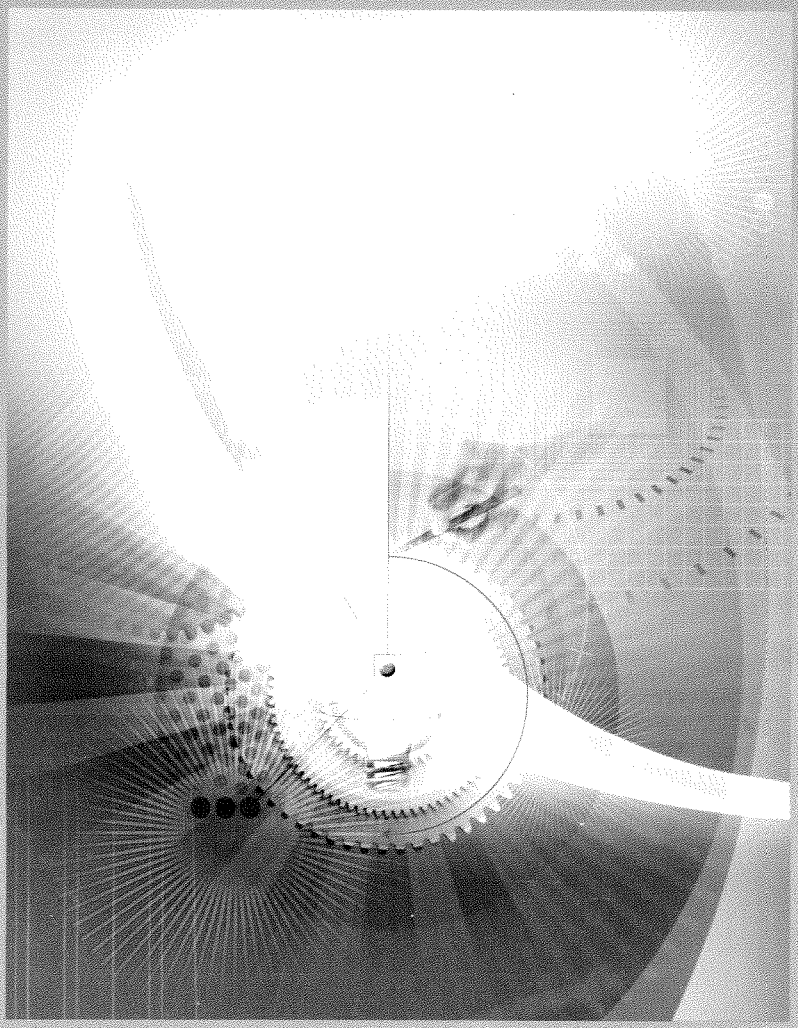
MERRILL EDUCATION/ASCD COLLEGE TEXTBOOK SERIES

Expanded 2nd Edition

UNDERSTANDING by DESIGN



GRANT WIGGINS AND JAY MCTIGHE

Chapter 7

Thinking like an Assessor

> We recognize understanding through a flexible performance. . . . Understanding shows its face when people can think and act flexibly around what they know. In contrast, when a learner cannot go beyond rote and routine thought and action, this signals lack of understanding. . . . To understand means to be able to perform flexibly.
>
> —David Perkins, "What Is Understanding?" in Martha Stone Wiske, Ed., *Teaching for Understanding,* 1998, p. 42

> The most important method of education . . . always has consisted of that in which the pupil was urged to actual performance.
>
> —Albert Einstein, *Ideas and Opinions,* 1954/1982, p. 60

Having clarified how to frame desired results in Stage 1, we now move to the second stage of backward design. Here we consider the assessment implications of our emerging design by asking (and reasking) the assessor's questions:

- What evidence can show that students have achieved the desired results (Stage 1)?
- What assessment tasks and other evidence will anchor our curricular units and thus guide our instruction?
- What should we look for, to determine the extent of student understanding?

Figure 7.1 lists the three stages of backward design and presents the considerations and design standards that apply. Stage 2 summarizes the elements to consider when planning for the collection of evidence from assessments.

Nowhere does the backward design process depart more from conventional practice than at this stage. Instead of moving from target to teaching, we ask, What would count as evidence of successful learning? Before we plan the activities, our question must first be, What assessment of the desired results logically follows Stage 1? And, specifically, what counts as evidence of the understanding sought?

Figure 7.1

The UbD Matrix: Focus on Stage 2

Key Design Questions	Chapters of the Book	Design Considerations	Filters (Design Criteria)	What the Final Design Accomplishes
Stage 1 • What are worthy and appropriate results? • What are the key desired learnings? • What should students come away understanding, knowing, and able to do? • What big ideas can frame all these objectives?	• Chapter 3—Gaining Clarity on Our Goals • Chapter 4—The Six Facets of Understanding • Chapter 5—Essential Questions: Doorways to Understanding • Chapter 6—Crafting Understandings	• National standards • State standards • Local standards • Regional topic opportunities • Teacher expertise and interest	• Focused on big ideas and core challenges	• Unit framed around enduring understandings and essential questions, in relation to clear goals and standards
Stage 2 • **What is evidence of the desired results?** • **In particular, what is appropriate evidence of the desired understanding?**	• **Chapter 7—Thinking like an Assessor** • **Chapter 8—Criteria and Validity**	• **Six facets of understanding** • **Continuum of assessment types**	• **Valid** • **Reliable** • **Sufficient**	• **Unit anchored in credible and useful evidence of the desired results**
Stage 3 • What learning activities and teaching promote understanding, knowledge, skill, student interest, and excellence?	• Chapter 9—Planning for Learning • Chapter 10—Teaching for Understanding	• Research-based repertoire of learning and teaching strategies • Appropriate and enabling knowledge and skill	Engaging and effective, using the elements of WHERETO: • *W*here is it going? • *H*ook the students • *E*xplore and equip • *R*ethink and revise • *E*xhibit and evaluate • *T*ailor to student needs, interests, and styles • *O*rganize for maximum engagement and effectiveness	• Coherent learning activities and teaching that will evoke and develop the desired understandings, knowledge, and skill; promote interest; and make excellent performance more likely

The mantra of this and the next chapter is to think like an assessor, not a teacher. Recall the logic of backward design, as shown in Figure 7.2. The text linking the first and second column shows what thinking like an assessor means.

As the logic of backward design reminds us, we are obligated to consider the assessment evidence implied by the outcomes sought, rather than thinking about assessment primarily as a means for generating grades. Given the goals, what performance evidence signifies that they have been met? Given the essential questions, what evidence would show that the learner had deeply considered them? Given the understandings, what would show that the learner "got it"? We urge teachers to consider a judicial analogy as they plan assessment. Think of students as juries think of the accused: innocent (of understanding, skill, and so on) until proven guilty by a preponderance of evidence that is more than circumstantial. In a world of standards-based accountability, such an approach is vital.

The following true stories illustrate the problem of failing to carefully consider the evidence needed.

- A kindergarten teacher has each student bring in a poster with 100 items for the hundredth day of school. But when asked to justify the assessment, the teacher refers to the state standard that references the "idea" of number and place value. But the learner had only to glue 100 items onto the poster. The students were not required to use or to explain rows, columns, or patterns. So we really only have evidence that the learner can count to 100, which is not the same as understanding "hundredness" as a concept linked to the base-10 system and the idea of place value, as the standard expects. In fact, because the poster was prepared at home, we do not have adequate evidence that the students did the counting on their own, without parental input.
- A 7th grade general science teacher captures the energy and imagination of his students by announcing that they will have to eat the results of their next science experiment. But what is engaging is not always what is most effective or appropriate, given the time available. In this instance, making peanut brittle offers little in the way of big ideas and enduring understanding for the week of experimentation allotted.
- A college history professor prepares a final exam consisting exclusively of 100 multiple-choice and short-answer questions for a syllabus in which "doing" history with primary sources is stressed as an important goal.

All of these assessments may have some merit when viewed through the lens of the individual lessons, but each needs to align better with curriculum goals. A more rigorous backward design—from the goals, generally (and key ideas to be understood, specifically), to the related assessments they imply—would have provided that link. These mistakes are common and not isolated. In fact, over the last decade we have observed that few educators have an adequate understanding of validity, and many harbor misunderstandings about assessment more generally, as reflected in both their comments and design work.

Figure 7.2

The Logic of Backward Design

Stage 1	Stage 2
If the desired result is for learners to . . .	***Then you need evidence of the student's ability to . . .***
Meet the standards . . . (G) Standard 6—Students will understand essential concepts about nutrition and diet. 6a—Students will use an understanding of nutrition to plan appropriate diets for themselves and others. 6c—Students will understand their own eating patterns and ways in which those patterns may be improved.	• Plan a diet for different kinds of people in different kinds of settings. • Reveal an understanding that the USDA guidelines are not absolute, but "guides"—and that there are other guides (as well as contextual variables). • Carefully note and analyze the habits of others as well as oneself, and make supported inferences about why people eat the way they do.
Understand that . . . (U) • A balanced diet contributes to physical and mental heath. • The USDA food pyramid presents relative guidelines for nutrition. • Dietary requirements vary for individuals based on age, activity level, weight, and overall health. • Healthful living requires an individual to act on available information about good nutrition even if it means breaking comfortable habits.	*That suggests the need for specific tasks or tests like . . .* (T) • Planning meals for diverse groups. • Reacting to excessively rigid or loose dietary plans made by others. • Making a good survey of what people actually eat and why.
Thoughtfully consider the questions . . . (Q) • What is healthful eating? • Are you a healthful eater? How would you know? • How could a healthy diet for one person be unhealthy for another? • Why are there so many health problems in the United States caused by poor eating despite all the available information?	
Know and be able to . . . (K) (S) • Use key terms—protein, fat, calorie, carbohydrate, cholesterol. • Identify types of foods in each food group and their nutritional values. • Be conversant with the USDA food pyramid guidelines. • Discuss variables influencing nutritional needs. • Identify specific health problems caused by poor nutrition.	*Quizzes:* On the food groups and the USDA food pyramid (OE) *Prompts:* Describe health problems that could arise as a result of poor nutrition and explain how these could be avoided; reflections on one's own eating habits and those of others.

More to the point of our focus on understanding, many teacher tests tend to focus on the accuracy of knowledge and skill rather than on evidence of *transferability*, based on big ideas in how to use knowledge and skill effectively. Our earlier discussion of the six facets and the need for transferability properly alerted designers to the importance of obtaining evidence of understanding through performance assessments. But the richness and complexity of all the desired results also demand variety in the evidence we collect.

Three basic questions

Thinking like an assessor boils down to a few basic questions. The first question is *What kinds of evidence do we need* to find hallmarks of our goals, including that of understanding? Before we design a particular test or task, it's important to consider the general types of performances that are implied. For example, regardless of content, understanding is often revealed through the exercises of comparing and contrasting or summarizing key ideas. After mapping a general approach to assessment, we then develop the assessment particulars.

The second question assumes that some particular task has been developed, about which we then ask, *What specific characteristics in student responses, products, or performances should we examine* to determine the extent to which the desired results were achieved? This is where criteria, rubrics, and exemplars come into play.

The third question has to do with a test for validity and reliability of the assessment: *Does the proposed evidence enable us to infer a student's knowledge, skill, or understanding?* In other words, does the evidence (Stage 2) align with our goals (Stage 1), and are the results sufficiently unambiguous? Few teachers are in the habit of testing their designs once the assessments have been fleshed out, but such self-testing is key to better results and to fairness.

In this chapter, we consider the first of the three aspects of thinking like an assessor: considering, in general terms, the kind of evidence needed to assess a variety of learning goals generally and understanding specifically. In the following chapter, we address the other two questions, related to criteria and the issues of validity and reliability.

An unnatural process

To think like an assessor prior to designing lessons does not come naturally or easily to many teachers. We are far more used to thinking like an activity designer or teacher once we have a target. That is, we easily and unconsciously jump to Stage 3—the design of lessons, activities, and assignments—without first asking ourselves what performances and products we need to teach toward.

Backward design demands that we overcome this natural instinct and comfortable habit. Otherwise our design is likely to be less coherent and focused

Figure 7.3
Two Approaches to Thinking About Assessment

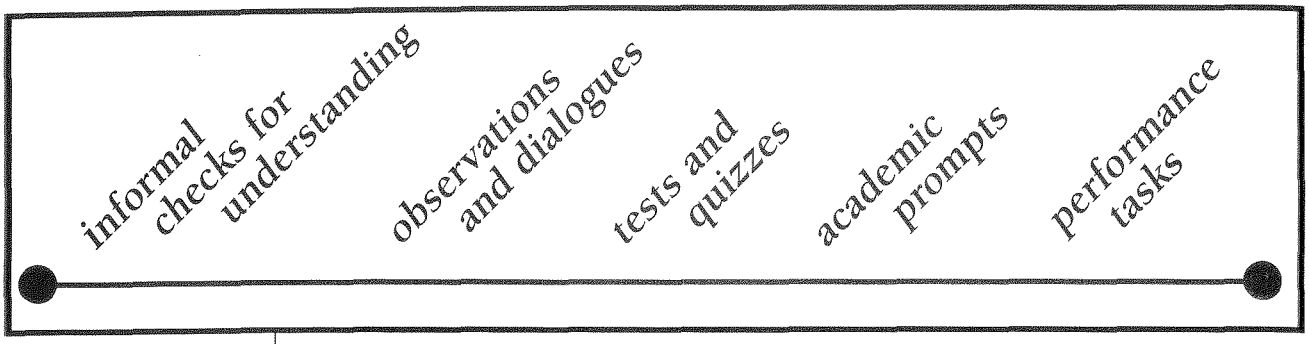
When thinking like an assessor, we ask—	When thinking like an activity designer (only), we ask—
• What would be sufficient and revealing evidence of understanding?	• What would be fun and interesting activities on this topic?
• Given the goals, what performance tasks must anchor the unit and focus the instructional work?	• What projects might students wish to do on this topic?
• What are the different types of evidence required by Stage 1 desired results?	• What tests should I give, based on the content I taught?
• Against what criteria will we appropriately consider work and assess levels of quality?	• How will I give students a grade (and justify it to their parents)?
• Did the assessments reveal and distinguish those who really understood from those who only seemed to? Am I clear on the reasons behind learner mistakes?	• How well did the activities work? • How did students do on the test?

on the desired results—and more the result of chance and the ability of students. In fact, a chief value of the UbD Template, and the backward design process more generally, is to provide tools and processes for short-circuiting this mental habit of overlooking the soundness of our assessments. Figure 7.3 summarizes how the two approaches—thinking like an assessor and thinking like an activity designer—differ.

The questions in the first column derive from the desired results and are likely to make the eventual activities and instructional strategies point toward the most appropriate assessments. The second column of questions, though sensible from the perspective of teaching and activity design, makes it far less likely that the assessments used will be appropriate. In effect, when we only think like an activity designer, we may well end up with something like the apples unit described in the Introduction. Although some students *may* develop important understandings and meet some standards as a result, it will be more by luck and happenstance than design. (See Chapter 8 for additional considerations regarding validity.)

Attention to the quality of local assessment could not be more important than it is now, when formal accountability demands assessments aligned with standards. Unless we use backward design frequently and carefully it is unlikely that the local assessment will provide the targeted feedback needed

Figure 7.4
A Continuum of Assessments



to inform teaching and enhance learning. Greater attention to self-assessment and peer review against design standards can greatly improve school-based assessments.

From snapshot to scrapbook

Effective assessment is more like a scrapbook of mementos and pictures than a single snapshot. Rather than using a single test, of one type, at the end of teaching, effective teacher-assessors gather lots of evidence along the way, using a variety of methods and formats. Thus, when planning to collect evidence of understanding, consider a range of assessment methods such as those shown in Figure 7.4.

This continuum of assessments includes checks of understanding (such as oral questions, observations, dialogues); traditional quizzes, tests, and open-ended prompts; and performance tasks and projects. They vary in terms of scope (from simple to complex), time frame (from short- to long-term), setting (from decontextualized to authentic contexts), and structure (from highly directive to unstructured). Because understanding develops as a result of ongoing inquiry and rethinking, the assessment of understanding should be thought of in terms of a collection of evidence over time instead of an "event"—a single moment-in-time test at the end of instruction—as so often happens in practice.

Given a focus on understanding, a unit or course will naturally be anchored by performance tasks or projects, because these provide evidence that students are able to use their knowledge in context. Our theory of understanding contends that contextualized application is the appropriate means of evoking and assessing *enduring* understandings. More traditional assessments (quizzes, tests, academic prompts, problem sets) round out the picture by assessing essential knowledge and skills that contribute to the culminating performances. The various types of evidence are summarized in Figure 7.5.

Figure 7.5
Types of Evidence

Performance Tasks

T

Complex challenges that mirror the issues and problems faced by adults. Ranging in length from short-term tasks to long-term, multistaged projects, they yield one or more tangible products and performances. They differ from academic prompts in the following ways:

- *Involve a real or simulated setting and the kind of constraints, background "noise," incentives, and opportunities an adult would find in a similar situation (i.e., they are authentic)*
- *Typically require the student to address an identified audience (real or simulated)*
- *Are based on a specific purpose that relates to the audience*
- *Allow students greater opportunity to personalize the task*
- *Are not secure: The task, evaluative criteria, and performance standards are known in advance and guide student work*

Academic Prompts

OE

Open-ended questions or problems that require the student to think critically, not just recall knowledge, and to prepare a specific academic response, product, or performance. Such questions or problems

- *Require constructed responses to specific prompts under school and exam conditions*
- *Are "open," with no single best answer or strategy expected for solving them*
- *Are often "ill structured," requiring the development of a strategy*
- *Involve analysis, synthesis, and evaluation*
- *Typically require an explanation or defense of the answer given and methods used*
- *Require judgment-based scoring based on criteria and performance standards*
- *May or may not be secure*
- *Involve questions typically only asked of students in school*

Quiz and Test Items

OE

Familiar assessment formats consisting of simple, content-focused items that

- *Assess for factual information, concepts, and discrete skill*
- *Use selected-response (e.g., multiple-choice, true-false, matching) or short-answer formats*
- *Are convergent, typically having a single, best answer*
- *May be easily scored using an answer key or machine*
- *Are typically secure (i.e., items are not known in advance)*

Informal Checks for Understanding

OE

Ongoing assessments used as part of the instructional process. Examples include teacher questioning, observations, examining student work, and think-alouds. These assessments provide feedback to the teacher and the student. They are not typically scored or graded.

Authentic performance—a necessity, not a frill

Understanding is revealed in performance. Understanding is revealed as transferability of core ideas, knowledge, and skill, on challenging tasks in a variety of contexts. Thus, assessment for understanding must be grounded in authentic performance-based tasks.

What do we mean by authentic tasks? An assessment task, problem, or project is authentic if it

• *Is realistically contextualized.* The task is set in a scenario that replicates or simulates the ways in which a person's knowledge and abilities are tested in real-world situations.

• *Requires judgment and innovation.* The student has to use knowledge and skills wisely and effectively to address challenges or solve problems that are relatively unstructured. Rather than a specific prompt or cue that tests a discrete piece of knowledge, realistic challenges require the learner to figure out the nature of the problem. What kind of knowledge and skill is being tapped here? How should I tackle it? Even when the goal may be quite clear, the student has to develop a plan and a procedure for solving the problem or addressing the issue.

• *Asks the student to "do" the subject.* Instead of reciting, restating, or replicating through demonstration what he was taught or already knows, the student has to carry out exploration and work in the discipline of science, history, or any other subject. The student's efforts resemble or simulate the kind of work done by people in the field.

• *Replicates key* challenging *situations in which adults are truly "tested" in the workplace, in civic life, and in personal life.* Real challenges involve specific situations with "messiness" and meaningful goals: important constraints, "noise," purposes, and audiences at work. In contrast, almost all school tests are without context (even when a writing prompt tries to suggest a sense of purpose and audience). In the real world—unlike schools—there is little if any secrecy about the goals or the criteria for success. Moreover, it is advantageous for the performer to ask questions of the "examiner" or boss, and ongoing feedback is typically available from colleagues. Students need to experience what it is like to perform tasks like those in the workplace and other real-life contexts, which tend to be complex and messy.

• *Assesses the student's ability to efficiently and effectively use a repertoire of knowledge and skill to negotiate a complex and multistage task.* Most conventional test items involve isolated bits of knowledge or elements of performance, similar to sideline drills in athletics, which differ from the integrated use of knowledge, skill, and feedback that a game requires. Although drills and tests are appropriate at times, performance is always more than the sum of the drills.

• *Allows appropriate opportunities to rehearse, practice, consult resources, and get feedback on and refine performances and products.* Although there is a role for the "secure" test that keeps questions secret and withholds resource materials from students, that type of testing must coexist with more transparent assessments of students if we are to focus their learning and improve their performance. As the apprenticeship model in the trades has proven, learning is maximized when cycles of *perform-feedback-revise-perform* guide the production of known high-quality products, judged against public performance standards. There is no room for "mystery testing" if we want students to demonstrate their understanding by using information, skills, and relevant resources to perform in context.

A call for greater authenticity in tests is not really new or inappropriate for a world of standards. Bloom and his colleagues signaled the importance of such assessments 40 years ago in the their description of *application* and in their account of synthesis: "a type of divergent thinking [in which] it is unlikely that the right solution to a problem can be set in advance" (Bloom, Madaus, & Hastings, 1981, p. 265).

An assessment approach grounded in authentic work calls for students (and teachers) to come to two important understandings: first, learning how adults in the larger world beyond the school *really* use or don't use the knowledge and skills that are taught in school; and second, how discrete lessons are meaningful, that is, how they lead to higher-quality performance or mastery of more important tasks. Just as the basketball player endures the drudgery of shooting endless foul shots and the flutist endures the monotony of playing scales—both with dreams of authentic achievement—so too must students experience that drills and quizzes have a pay-off in better performances on worthy endeavors.

Designing around problems not just exercises

Designers often find it helpful to consider the more general question implied in the basketball and flute examples to sharpen their assessments: Does the test amount to just simplified "drill" out of context? Or does the assessment require students to really "perform" wisely with knowledge and skill, in a problematic context of real issues, needs, constraints, and opportunities? To get evidence of true understanding requires that we elicit learner judgments made during genuine performance, not just seeing how they respond to easily followed cues that require mere recall and plugging in.

Put in different words, in authentic assessment we have to be sure that we have presented the learner with an *authentic problem*, to invoke an apt distinction made by Dewey almost a hundred years ago:

> *The most significant question which can be asked about any situation or experience proposed to induce [and reveal] learning is what quality of problem it involves . . . but it is indispensable to distinguish between genuine . . . or mock problems. The following questions may aid in making such a discrimination. . . . Does the question naturally suggest itself within some situation or personal experience? Or is it an aloof thing . . . ? Is it the sort of trying that would arouse observation and engage experimentation out side of school? [Or, is it] made a problem for the pupil only because he cannot get the required mark or be promoted or win the teacher's approval, unless he deals with it? (1916, p. 155)*

A variant of Dewey's distinction can be found in all the performance areas, whereby we distinguish exercises from the problems of performance. An exercise involves a straightforward execution of a "move" out of context. A problem is a demand within performance, requiring thought of the many choices and challenges that confront a performer in context. Lay-up drills in basketball

are exercises: Players form two lines, one for passers, the other for shooters, and they exchange free shots at the basket. Using that skill (shooting at or making a basket) in a game, however, requires the shooters to also work around the other team's defense.

A similar situation occurs in science. A typical science lab presents an exercise, not a problem: There is a right approach, a right answer, and thus no inherent puzzles or challenges to our understanding. By contrast, having to design and debug an effective, feasible, and cost-sensitive experiment to make sense of a puzzling phenomenon reflects true problem solving. All "doing" of a subject involves problem solving, so our assessments of understanding must be based on real problems, not just exercises requiring discrete facts and skills used in isolation.

Mathematics and history may well be the program areas in most need of thinking through this distinction. Almost every mathematics and history test in K–12 education is a set of exercises, not problems in the sense discussed: One need only respond on cue with the correct move. It doesn't matter whether the topic is adding fractions or understanding the civil rights era, the learner is invariably tested by unambiguous exercises having right answers. An authentic problem related to fractions or history must be like playing a basketball game—just shooting at the basket unhindered or just plugging in the obvious approach or facts isn't enough. The authentic problem solving requires deciding when to use which approach and which facts. Is this problem best solved by using fractions or decimals? Is the civil rights era best understood as a religious or secular movement?

To build math and history assessments out of only exercises (as we so often do) misses the essence of authentic performance in those fields. As we have said, real performance always involves transfer—that is, the flexible use of knowledge and skill in light of particular challenges. It requires puzzling out and making sense of what a situation demands, which is very different from merely responding to a highly structured exercise looking for the right response. Transferability is understanding revealed: The performers must figure out *which* knowledge and skill is needed on their own, without simplifying teacher prompts or cues, to solve the real problems of performance.

Figure 7.6 helps clarify the difference between a problem and an exercise. Note that exercises are necessary but not sufficient in developing competent performance; nor are exercises always reliable indicators of the ability to perform.

MISCONCEPTION ALERT!

Our goal in Stage 2 is appropriate evidence, not interesting projects or tasks. Although our aim should always be to make assessments interesting and thought-provoking (because we thereby evoke the best and most thorough work), that is not the main point in Stage 2. Many projects are fun and educational, but they may not provide enough evidence about the understandings sought in Stage 1—particularly if the work involves collaboration and freedom of choice in approach, content, and presentation. Many exercises are less engaging than complex performance tasks, but sometimes they yield more conclusive evidence about a specific understanding or skill. We must ensure that the project is designed backward from the evidence we need, not designed primarily with the learner's interests in mind. Beware of confusing interesting performance tasks or projects with valid evidence. This point is taken up in more detail in Chapter 8.

Figure 7.6
Problems Versus Exercises

	Problem	Exercise
The Framing of the Task	The problem statement is clear, but few if any cues or prompts are offered about how to best frame or solve the problem.	The task is either simple or made simple by specific cues or prompts as to the nature of the challenge or how to proceed in meeting it.
The Approach	Various approaches are possible. Figuring out what kind of problem this is and isn't is a key aspect of the challenge; that is, a strategy is needed. Some combination of logical method with trial and error will likely be required.	There is one best approach (though it might not be stated), and it is suggested by how the exercise is framed. The learner's ability to recognize and use the "right" tactic is a key goal of the exercise.
The Setting	Realistically "noisy" and complicated, typically involving different—sometimes competing—variables related to audience, purpose, criteria for judging work, and more.	Simplified to ensure that the only "variable" is the targeted skill or knowledge. (Similar to sideline drills in athletics or fingering exercises in music.)
The Solution	The goal is an appropriate solution, mindful of various requirements and perhaps competing variables and cost/benefit considerations. There may be a right answer, but it follows from sound reasoning and a supported argument or approach.	The goal is the right answer. The exercise is built to ensure that there is only one right answer, by design. Though it may be a puzzling challenge, there is a definite right answer that can be found via recall and plugging in of prior knowledge, with little or no modification.
Evidence of Success	The focus shifts from the answer to the justification of the approach and solution.	The accuracy of the answer and the choice of the "correct" approach.

Framing performance tasks using GRASPS

Authentic performance tasks are distinguished from other types of assessments by their particular features. Performance tasks typically present students with a problem: a real-world goal, set within a realistic context of challenges and possibilities. Students develop a tangible product or performance for an identified audience (sometimes real, sometimes simulated). And the evaluative criteria and performance standards are appropriate to the task—and known by the student in advance.

Because these elements characterize authentic assessments, we can use them during task design. We have created a design tool using the acronym GRASPS to assist in the creation of performance tasks. Each letter corresponds with a task element—Goal, Role, Audience, Situation, Performance, Standards.

Figure 7.7 presents each element with corresponding prompts to help designers construct performance tasks. Often, teachers transform existing assessments or engaging learning activities using GRASPS.

Here is an example of a performance task in science, constructed using GRASPS, for assessing understanding of multivariable experimental design:

- Goal and Role: As a scientist with a consumer research group, your task is to design an experiment to determine which of four brands of detergent will most effectively remove three different types of stains on cotton fabric.
- Audience: Your target audience is the testing department for *Consumer Research* magazine.
- Situation: You have a two-part challenge: (1) to develop an experimental design for isolating the key variables, and (2) to clearly communicate the procedure so that the staff of the testing department can conduct the experiment to determine which cleaner is most effective for each type of stain.
- Product: You need to develop a written experimental procedure (following the given format) outlining the steps in sequence. You may include an outline or graphic format to accompany the written description.
- Standards: Your experimental design needs to follow the criteria for good design accurately and completely; appropriately isolate the key variables; include a clear and accurate written description of the procedure (an outline or graphic to assist the testers is optional); and enable the testing department staff to determine which cleaner is most effective for each type of stain.

Not every performance assessment needs to be framed by GRASPS. However, we propose that at least one core performance task for assessing understanding in a major unit or course be developed in this fashion. Many teachers have observed that tasks framed this way provide students with clear performance targets as well as real-world meaningfulness not found in decontextualized test items or academic prompts.

Performance task vignettes

The following vignettes offer brief descriptions of performance tasks for possible use in assessing student understanding. Notice how they reflect the GRASPS elements.

- From the mountains to the seashore (history, geography; grades 6–8). A group of nine foreign students is visiting your school for one month as part of an international exchange program. (Don't worry, they speak English!) The principal has asked your class to plan and budget a four-day tour of Virginia to help the visitors understand the state's impact on the history and development of our nation. Plan your tour so that the visitors are shown sites that best capture the ways that Virginia has influenced our nation's development. Your task is to prepare a written tour itinerary, including an explanation of why each

Figure 7.7
GRASPS Task Design Prompts

Goal

- Your task is _____.
- The goal is to _____.
- The problem or challenge is _____.
- The obstacles to overcome are _____.

Role

- You are _____.
- You have been asked to _____.
- Your job is _____.

Audience

- Your clients are _____.
- The target audience is _____.
- You need to convince _____.

Situation

- The context you find yourself in is _____.
- The challenge involves dealing with _____.

Product, Performance, and Purpose

- You will create a _____
  in order to _____.
- You need to develop _____
  so that _____.

Standards and Criteria for Success

- Your performance needs to _____.
- Your work will be judged by _____.
- Your product must meet the following standards _____.

site was selected. Include a map tracing the route for the four-day tour and a budget for the trip.

• Garden design (mathematics, grades 6–8). You've been asked to plan a flower garden for a company with a logo that has side-by-side circular, rectangular, and triangular shapes. Your final product should be a labeled scale drawing and a list of how many plants of each type and color you need to execute the plan.

• Literary Hall of Fame (English, grades 10–12). The Council of Arts and Letters has announced the establishment of a Hall of Fame to honor the works of notable U.S. authors and artists. Since your class is finishing a course on U.S. literature, you have been asked to submit a nomination for an author to be admitted to the Hall of Fame. Complete the nomination form for an author whom you believe is worthy of induction. Your essay should include your analysis of the author's contribution to U.S. literature and your rationale for recommending the author for inclusion in the Hall of Fame.

• Mail-order friend (language arts, grades K–2). Imagine that you have an opportunity to order a friend by telephone from a mail-order catalog. Think about the qualities that you want in a friend. Before you order your friend over the telephone, practice asking for three characteristics that you want in a friend and give an example of each characteristic. Remember to speak clearly and loudly enough so that the sales person will know exactly what you're looking for. Your request will be taped and assessed against a rubric for clarity as well as how much thought you put into your request.

• Moving Van Go (mathematics and writing, grades 6–9). You are working for a moving company that plans to submit a bid for moving the contents of an office building to a new location. You are responsible for determining the minimum volume of furniture and equipment that must be moved. The exemplary product will take into account (a) the stackability of the items, (b) the interlocking nature of noncubical pieces, (c) the padding to protect the furniture, and (d) the number and size of the boxes needed to pack the small items. You will prepare a written report setting out the volume of items to be moved and a rationale for the findings, and a chart showing how the items will be placed to minimize the volume needed.

• Drywalling a home (mathematics, grades 8–10). When contractors give an estimate on home repairs, how can we know if the cost is reasonable? In this task, you will determine whether a drywalling contractor is giving accurate information, or trying to overcharge an uninformed customer. You will be given room dimensions and cost figures for materials and labor.

• The Cheyenne Indians—what really happened (history, college juniors and seniors). You will research a possible massacre during the Civil War about which no detailed narratives have been written. You will read Senate transcripts and various conflicting first-hand accounts, leading to your own narrative for inclusion in a history book. Your work will be reviewed by your peers and judged by professors serving as textbook editors.

• Fitness plan (physical education and health, secondary level). Playing the role of a trainer at a health club, you will develop a fitness program, consisting of aerobic, anaerobic, and flexibility exercises, for a new client. The fitness plan needs to take into account the client's lifestyle, age, activity level, and personal fitness goals. You will be given detailed descriptions of various clients.

Using the six facets as assessment blueprints

A basic requirement of assessing for understanding is that we need to know the learners' thought processes along with their "answers" or solutions. Their explanation of *why* they did what they did, their *support* for the approach or response, and their *reflection* on the result that we may gain fuller insight into their degree of understanding. Answers without reasons and support are typically insufficient to "convict" the learner of understanding. This is why we require both a dissertation and its defense for a doctorate. Assessment of understanding is enhanced when we make greater use of oral assessments, concept webs, portfolios, and constructed response items of all types to allow students to show their work and reveal their thinking. Selected response formats—multiple choice, matching pairs, true or false—in general provide insufficient (and sometimes misleading) evidence about understanding or its absence.

The six facets of understanding signal the types of performances we need as valid measures of understanding. They map out, in general terms, the kinds of performance evidence we need to successfully distinguish factual knowledge from an understanding of the facts. The value of the facets becomes clearer when we add them to our earlier backward design graphic, as shown in Figure 7.8.

The six facets provide a helpful scaffold for the second column by reminding us, in general, what understanding looks like. We can use the various abilities central to each facet to guide the design process in Stage 2. For example, Facet 1 involves the ability to explain, verify, or justify a position in one's own words. Starting with the stem, "A student who *really* understands . . ." and adding the key words from each facet produces suggestions for the kinds of assessment task we need, as illustrated in Figure 7.9.

This emerging list provides a useful start to a blueprint for assessing understanding. Regardless of our topic or the age of the students we teach, the verbs on this list suggest the kinds of assessments needed to determine the extent to which students understand. Then, in the third column in Figure 7.8, we can get more specific by asking, What kinds of tasks are suitable for the specific desired results of Stage 1 and the students we teach? Which facet (or facets) will most appropriately guide the design of a particular task, with specific performance, process, or product requirements?

Here are some starter ideas for performance tasks built around the six facets of understanding.

Facet 1: Explanation

Explanation asks students to tell the "big idea" in their own words, make connections, show their work, explain their reasoning, and induce a theory from data.

Figure 7.8

The Logic of Backward Design with the Six Facets

Stage 1	Stage 2	
If the desired result is for learners to . . .	***Then you need evidence of the student's ability to . . .***	***So the assessments need to require something like . . .***
understand that • A balanced diet contributes to physical and mental heath. • The USDA food pyramid presents relative guidelines for nutrition. • Dietary requirements vary for individuals based on age, activity level, weight, and overall health. • Healthful living requires an individual to act on available information about good nutrition even if it means breaking comfortable habits. *and thoughtfully consider the questions . . .* • What is healthful eating? • Are you a healthful eater? How would you know? • How could a healthy diet for one person be unhealthy for another? • Why are there so many health problems in the United States caused by poor eating despite all the available information?	*explain* • A balanced diet • The consequences of poor nutrition • Why we eat poorly, despite the information available *interpret* • Food nutrition labels • Data on the impact of fast foods on eating patterns *apply, by* • Planning healthy menus • Evaluating various plans and diets *see from the points of view of* • People of other cultures and regions in terms of their dietary beliefs and habits *empathize with* • A person living with significant dietary restrictions due to a medical condition *reflect on* • Personal eating habits • Whether foods that are good for you always taste bad	• Develop a brochure to help younger students understand what is meant by a balanced diet and the heath problems resulting from poor eating. • Discuss the popularity of fast foods and the challenges of eating a healthful diet in today's fast-paced world. • Plan a menu for a class party consisting of healthy, yet tasty, snacks. • Conduct and present research on the impact of diverse diets (i.e., Antarctica, Asia, the Middle East) on health and longevity. • Describe how your life would be affected (and how it might feel) to live with dietary restrictions due to a medical condition (such as diabetes). • Reflect: To what extent are you a healthy eater? How might you become a healthier eater?

Figure 7.9

Using the Six Facets to Build Assessments for Understanding

A student who *really* understands . . .

Facet 1. Can explain—*Demonstrates sophisticated explanatory power and insight. Is able to . . .*

a. Provide complex, insightful, and credible reasons—theories and principles, based on good evidence and argument—to explain or illuminate an event, fact, text, or idea; show meaningful connections; provide a systematic account, using helpful and vivid mental models.
   - Make fine, subtle distinctions; aptly qualify her opinions.
   - See and argue for what is central—the big ideas, pivotal moments, decisive evidence, key questions, and so on.
   - Make good predictions.

b. Avoid or overcome common misunderstandings and superficial or simplistic views—shown, for example, by avoiding overly simplistic, hackneyed, or imprecise theories or explanations.

c. Reveal a personalized, thoughtful, and coherent grasp of a subject—indicated, for example, by developing a reflective and systematic integration of what she knows. This integration would therefore be based in part upon significant and apt direct or simulated experience of specific ideas or feelings.

d. Substantiate or justify her views with sound argument and evidence.

Facet 2. Can interpret—*Offers powerful, meaningful interpretations, translations, narratives. Is able to . . .*

a. Effectively and sensitively interpret texts, data, and situations—shown, for example, by the ability to read between the lines and offer plausible accounts of the many possible purposes and meanings of any "text" (book, situation, human behavior, and so on).

b. Offer a meaningful and illuminating account of complex situations and people—shown, for example, by the ability to provide historical and biographical background to help make ideas more accessible and relevant.

Facet 3. Can apply—*Uses knowledge in context; has know-how. Is able to . . .*

a. Employ her knowledge effectively in diverse, authentic, and realistically messy contexts.

b. Extend or apply what she knows in a novel and effective way (invent in the sense of innovate, as Piaget discusses in *To Understand Is to Invent*[1]).

c. Effectively self-adjust as she performs.

Facet 4. Sees in perspective—*Is able to . . .*

a. Critique and justify a position, that is, see it as a point of view; to use skills and dispositions that embody disciplined skepticism and the testing of theories.

b. Place facts and theories in context; know the questions or problem to which the knowledge or theory is an answer or solution.

c. Infer the assumptions upon which an idea or theory is based.

d. Know the limits as well as the power of an idea.

e. See through argument or language that is biased, partisan, or ideological.

(continued on next page)

Figure 7.9 (continued)

f. See and explain the importance or worth of an idea.
g. Take a critical stance; wisely employ *both* criticism and belief (an ability summarized by Peter Elbow's maxim that we are likely to better understand when we methodically "believe when others doubt and doubt when others believe"[2]).

Facet 5. Demonstrates empathy—Is able to . . .

a. Project himself into, feel, and appreciate another's situation, affect, point of view.
b. Operate on the assumption that even an apparently odd or obscure comment, text, person, or set of ideas may contain insights that justify working to understand it.
c. See when incomplete or flawed views are plausible, even insightful, though perhaps somewhat incorrect or outdated.
d. See and explain how an idea or theory can be all too easily misunderstood by others.
e. Watch and listen sensitively and to perceive what others often do not.

Facet 6. Reveals self-knowledge—Is able to . . .

a. Recognize his own prejudices and style and how they color understanding; see and get beyond egocentrism, ethnocentrism, present-centeredness, nostalgia, either/or thinking.
b. Engage in effective metacognition; recognize intellectual style, strengths, and weaknesses.
c. Question his own convictions; like Socrates, sort out mere strong belief and habit from warranted knowledge, be intellectually honest, and admit ignorance.
d. Accurately self-assess and effectively self-regulate.
e. Accept feedback and criticism without defensiveness.
f. Regularly reflect on the meaning of one's learning and experiences.

[1]Jean Piaget. (1973). *To Understand Is to Invent: The Future of Education.* New York: Grossman's Publishing Co.
[2]Peter Elbow. (1973). *Writing Without Teachers.* New York: Oxford University Press.

• Mathematics—subtraction. Design a lesson plan, using manipulatives, to teach a new student to our class what "subtraction" is all about.

• Social studies—geography and economics. Create a graphic organizer to show connections between environment, natural resources and economy for two different regions.

• Science—electricity. Develop a trouble-shooting guide for an electric circuit system.

• Foreign language—language structure. Develop a guidebook in which you explain the difference between the various forms of past tense, and when they should and should not be used.

Facet 2: Interpretation

Interpretation requires the student to make sense of stories, art works, data, situations, or claims. Interpretation also involves translating ideas, feelings, or work done in one medium into another.

• History—U.S. history. Select 5–10 songs about the United States written since the Civil War. Use them to explore the questions: Are we the nation we set out to be? How have we seen ourselves as a nation? Which attitudes have changed and which have not?

• Literature—*The Catcher in the Rye* and *Frog and Toad Are Friends.* Answer the question, What's wrong with Holden? Study the words and actions of the main character, and the reaction of other characters to help you make sense of Holden Caulfield. Examine the question, Who is a true friend? Study the words and actions of the main characters, Frog and Toad. Look for patterns to help you answer the question.

• Visual and performing arts—any medium. Represent strong emotions (e.g., fear and hope) through a collage, dance, musical piece, or other medium. How does the medium affect the message?

• Science and mathematics—data patterns. Collect data over time on any complex phenomena (e.g., weather variables). Analyze and display the data in order to find patterns.

Facet 3: Application

Students who understand can use their knowledge and skill in new situations. Place emphasis on application in authentic contexts, with a real or simulated audience, purpose, setting, constraints, and background noise.

• Mathematics—area and perimeter. Design the shape of a fenced-in section of a yard, given a specified amount of fencing material, to maximize the play area for a new puppy.

• Social studies—map skills. Develop a scaled map of your school to help a new student find her way around.

• Health—nutrition. Develop a menu plan for healthful meals and snacks for a family of five for one week, staying within a defined budget.

• Science—environmental studies. Perform a chemical analysis of local stream water to monitor clean water compliance and present your findings to the regional EPA office.

Facet 4: Perspective

Perspective is demonstrated when the student can see things from different points of view, articulate the other side of the case, see the big picture, recognize underlying assumptions, and take a critical stance.

• History—compare and contrast. Review British, French, and Chinese textbook accounts of the U.S. Revolutionary War era. Identify the historical perspective of each, and defend or oppose their use as teaching resources at a simulated school board meeting.

• Arithmetic—different representations. Compare the pros and cons of different views of the same quantity represented in decimals, fractions, and percentages; and in different graphical and symbolic representations.

• English or language arts—literary analysis and writing. Assume you are the editor at a major publishing house. Review a submitted short story for possible plagiarism. (The teacher does not tell students that they are reviewing a story written by one of the authors they have studied this year.) Then write a tactful but firm letter back to the author on the likely source of this manuscript.

• Geometry. Compare the shortest distance between two points in three different spaces: physical corridors in their school building, on the earth's surface, and in Euclidean space.

• Music. Listen to three different recorded versions of the same song and critique each version, as if you are a producer working with your current star to choose an arrangement.

Facet 5: Empathy

Intellectual imagination is essential to understanding, and it manifests itself not only in the arts and literature, but more generally through the ability to appreciate people who think and act differently from us. The goal is not to have students accept the ways of others, but to help them better understand the diversity of thought and feeling in the world; that is, to develop their capacity to walk in someone else's shoes. In this way, students can avoid stereotyping and learn how yesterday's weird idea can be commonplace today.

• History. Using a *Meeting of Minds* format, role-play various characters with other students and discuss or debate an issue (e.g., settlers and Native Americans on Manifest Destiny, Truman deciding to drop the atomic bomb, the reasons for the collapse of the Soviet Union).

• English or language arts—writing. Imagine you are the newly selected poet laureate of the European Union and have been commissioned to write a sonnet about events in the Middle East. It will be published in the *Jerusalem Times* as well as the *Cairo Daily News*. Your goal is to promote empathy for the people suffering on both sides of this struggle.

• Science. Read and discuss premodern or discredited scientific writings to identify plausible or "logical" theories (given the information available at the time), such as Ptolemy's explanation for why the Earth must be at rest, and Lamarck's account of development.

• Literature—Shakespeare. Imagine you are Juliet from *Romeo and Juliet*, and consider your terrible, final act. Write your final diary entry to describe what are you thinking and feeling. *(Note: This prompt was used on a British national exam.)*

Facet 6: Self-Knowledge

It is important to require students to self-assess their past as well as their present work. It is only through self-assessment that we gain the most complete insight into how sophisticated and accurate students' views are of the tasks, criteria, and standards they are to master.

A simple strategy is to make the first and last written assignments for any course *the same question*, and require students to write a self-assessment postscript describing their sense of progress in understanding. Teachers who collected student work samples in portfolios use a related approach by asking students to review their portfolios and respond to reflective questions: How does your work show how you have improved? What task or assignment was the most challenging and why? Which selection are you most proud of and why? In what ways does your work illustrate your strengths and weaknesses as a learner?

Here are some other approaches to self-assessment and metacognition for any subject and level:

- Here I Come! At the end of the school year, write a letter to next year's teacher describing yourself as a learner. Describe your academic strengths, needs, interests, and learning styles. Set specific learning goals based on self-assessment of your performance during the year that is ending. (Ideally, these letters would be systematically collected and sent to the receiving teachers during the summer.)
- What have I learned? Add a postscript to any paper written for a course in which you must dispassionately self-assess the strengths, weaknesses, and gaps in your approach or response. Pose the question, Knowing what I now do, what would I do differently next time?
- How well do I think I did? Middle school, high school, and college students can produce a written or oral self-assessment against the criteria used to evaluate the work (rubrics). The accuracy of the self-assessment is a small part of the grade. *(Note: This practice is used on every major assignment at Alverno College in Milwaukee, Wisconsin.)*

First among equals

We generally need to include the first facet, *explanation,* as part of any task involving the other five facets. We need to know *why* the students performed the way they did, what they think it means, and what justifies their approach, not just that they did it. In performance-based assessment for understanding, in other words, the tasks and performances should require reflection, explicit self-assessment, and self-adjustment, with reasoning or rationale made as evident as possible.

Using essential questions for assessment

If we have done a good job in framing the unit around essential questions, then we have another helpful way to think through and to test the appropriateness of our assessment ideas. The performances should directly or indirectly require the students to address the essential questions.

Look back at our recurring unit on nutrition (Figure 7.10). Note how the Essential Questions provide a helpful framework upon which the right kinds of tasks can be built.

Figure 7.10
Essential Questions Leading to Performance Tasks

Essential Questions	Proposed Performance Tasks
• Why do people have such a difficult time eating right?	• Students collect and analyze survey data to find out where students eat most of their meals
• Must food that is really good for you taste bad and vice versa?	• Students investigate the nutritional value of various foods to compare taste with health benefits
• Why do experts often disagree about dietary guidelines? What agreement exists amidst the disagreement?	• Students compare and evaluate various approaches to good nutrition—USDA, Atkins, Mediterranean—culminating in poster display and oral report

You might start your work by simply assuming that the essential question will be like a blue-book exam question from college—begin your design work by thinking of the questions as final essay prompts. Then, see if you can take the prompt and devise a GRASPS situation in which the same question is being addressed in a more authentic manner.

If a GRASPS scenario seems contrived or you believe that a traditional writing prompt provides the most appropriate assessment, use the essential questions to focus learning and as a part of the final exam. Using the essential questions in this way provides a focus for both teachers and students and renders the assessment process far less mysterious and arbitrary than it needs to be.

Rounding out the evidence

The question we ask when thinking like an assessor is this: What's the evidence we need (given the desired results)? We should have no philosophical axe to grind in answering that question. We should use the best kinds of assessments, including, where appropriate, short-answer prompts and selected-response quizzes. Too often as teachers, we rely on only one or two types of assessment, then compound that error by concentrating on those aspects of the curriculum that are most easily tested and graded by multiple-choice or short-answer items. On the other hand, it is a common misconception that reform is about an exclusive reliance on authentic assessments. This is simply not the case. For evidence of many desired results, especially discrete knowledge and skill, objective quizzes, tests, and observations with checklists often suffice. We can

visually depict the relationship of various assessment types to curriculum priorities by considering the chart in Figure 7.11 (p. 170).

Frequently, too, we fail to consider the differences between tests and other forms of assessment that are particularly well suited for gathering evidence of understanding. In fact, in aiming for understanding, we usually err in assuming that formal and summative testing is needed for evidence gathering. The corollary is to assume that everything that is assessed must be graded.

On the contrary, as the phrases "check for understanding" and "feedback" imply, ongoing formative assessments are vital to reveal students' understanding and misunderstanding. A simple device for ongoing assessment of understanding is the "one-minute essay." At the end of each class, students are asked to answer two questions: (1) What is the big point you learned in class today? and (2) What is the main unanswered question you leave class with today? A quick scan of student responses provides the teacher with immediate feedback on the extent of student understanding (or lack thereof). Indeed, professors at Harvard University have called this technique one of the most effective innovations in their teaching (Light, 2001).

In our own teaching, we have required students to bring written questions to class each day. Class begins by having learners discuss their questions in groups of two or three, bringing their most important question to the entire class for consideration. Then, we look for patterns through a web of questions and possible answers. With a few minutes to go at the end of class, we ask one or two students to summarize the conversation and ask everyone to write notes. Perkins (1992) proposes many other strategies, and we suggest other such checks for understanding in Chapter 9.

The need for a variety of assessment evidence in Stage 2 is signaled in the Design Template by one box for key Performance Tasks and another box for all Other Evidence. A balance of types of assessment is good measurement and wise practice in teaching.

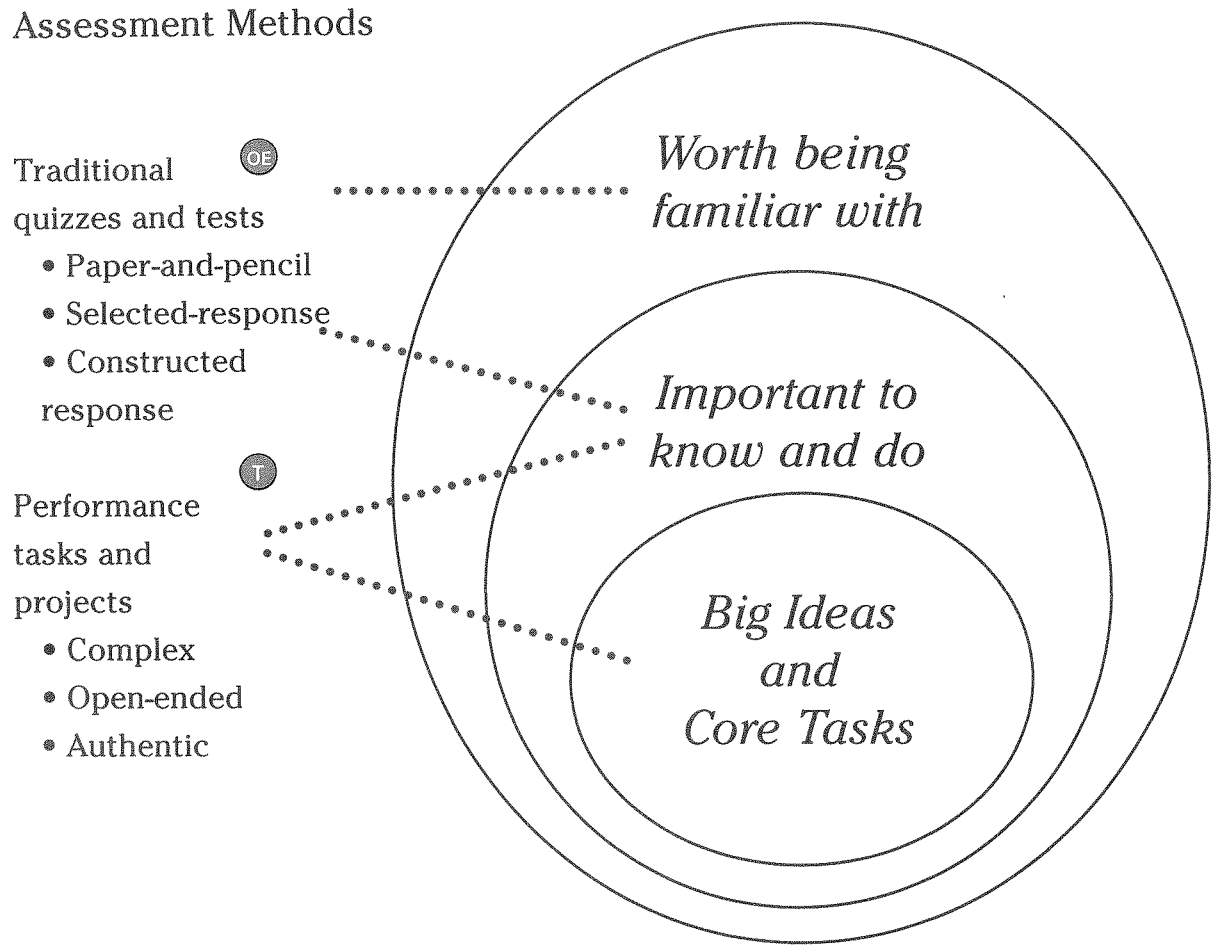
In this first look at assessment we have considered designing assessments by working backward from the desired results of Stage 1. We stressed that when understanding is the focus our evidence must be grounded in authentic performance tasks (supplemented as needed by "other evidence") that involve real problems, not mere exercises. The facets help us find the right kinds of tasks, and GRASPS helps us further refine each task to ensure its authenticity. And we reminded readers that there is always a need for variety of evidence.

MISCONCEPTION ALERT!

When we speak of evidence of understanding, we are referring to evidence gathered through a variety of formal and informal assessments during a unit of study or a course. We are not alluding only to end-of-teaching tests or culminating performance tasks. Rather, the collected evidence we seek may include observations and dialogues, traditional quizzes and tests, performance tasks and projects, as well as students' self-assessments gathered over time.

Figure 7.11
Curricular Priorities and Assessment Methods

In effective assessments, we see a match between the type or format of the assessment and the needed evidence of achieving the desired results. If the goal is for students to learn basic facts and skills, then paper-and-pencil tests and quizzes generally provide adequate and efficient measures. However, when the goal is deep understanding, we rely on more complex performances to determine whether our goal has been reached. The graphic below reveals the general relationship between assessment types and the evidence they provide for different curriculum targets.



Backward design in action with Bob James

Now I need to think about what would actually serve as evidence of the understandings I'm after. This will be a bit of a stretch for me. Typically in a 3–4 week unit like this one, I give one or two quizzes, have a project, which I grade, and conclude with a unit test (generally multiple choice or matching). Although this approach to assessment makes grading (and justifying the grades) fairly

easy, I have come to realize that these assessments don't always provide adequate evidence regarding the most important understandings of the unit. I tend to test what is easy to test instead of assessing what is most important, namely the understandings and attitudes students should take away, above and beyond nutritional facts. In fact, one thing that has always disturbed me is that the kids tend to focus on their grades rather than on their learning. Perhaps the way I've used assessments—more for grading purposes than to document learning—has contributed to their attitude.

Now I need to think about what would actually serve as evidence of the enduring understanding I'm after. After reviewing some examples of performance assessments and discussing ideas with my colleagues, I have decided on the following performance task:

> *Because we have been learning about nutrition, the camp director at the Outdoor Education Center has asked us to propose a nutritionally balanced menu for our three-day trip to the center later this year. Using the USDA food pyramid guidelines and the nutrition facts on food labels, we will design a plan for three days, including three main meals and three snacks (a.m., p.m., and campfire). Our goal is a tasty and nutritionally balanced menu.*

This task also links well with one of our unit projects—to analyze a hypothetical family's diet for a week and propose ways to improve their nutrition. With this task and project in mind, I can now use quizzes to check their prerequisite knowledge (of the food groups and the food pyramid recommendations) and a test for their understanding of how a nutritionally deficient diet contributes to health problems. This is the most complete assessment package I've ever designed for a unit, and I think that the task will motivate students as well as provide evidence of their understanding.

Looking ahead

We need now to consider the second and third questions that lie at the heart of thinking like an assessor: What should we look for when we assess? How can we be confident that our proposed assessments permit valid and reliable inferences back to Stage 1? In the next chapter we will turn to those two questions.

Chapter 8

Criteria and Validity

Assessment and feedback are crucial for helping people learn. Assessment that is consistent with principles of learning and understanding should:

- Mirror good instruction
- Happen continuously, but not intrusively, as part of instruction
- Provide information about the levels of understanding that students are reaching.

—John Bransford, Ann Brown, and Rodney R. Cocking, *How People Learn,* 2000, p. 244

The central problem . . . is that most widely used assessments of academic achievement are based on highly restrictive beliefs about learning and competence.

—Committee on the Foundations of Assessment, *Knowing What Students Know: The Science and Design of Educational Assessment,* 2001, p. 2

In Chapter 7 we focused on the kinds of assessments needed to provide appropriate evidence of our desired results. We noted that there is always a need for a variety of evidence and that assessment plans must be grounded in authentic performance tasks. We also found that the assessment of understanding requires performance assessment: We need to see how well the learner handles performance challenges in context, and what their thought processes were in doing so.

The need for criteria

Because the kinds of open-ended prompts and performance tasks needed to assess for understanding do not have a single, correct answer or solution process, evaluation of student work is based on judgment guided by criteria. Clear and appropriate criteria specify what we should look at to determine the degree of understanding and serve us in making a judgment-based process consistent and fair (Wiggins, 1998, pp. 91–99). How, then, do we come up with appropriate criteria and how do we make them clear to learners?

Appropriate criteria highlight the most revealing and important aspects of the work (given the goals), not just those parts of the work that are merely easy to see or score. For example, when reading a story we want to be engaged, to have our imagination sparked or interest fired. The best stories hook and hold our interest through an effective combination of plot and character. So a key criterion in judging stories is *engagement*. Another might be the author's *craftsmanship* in using effective literary devices and language choices. A third might relate to depth and credibility of the characters—or *character development*. The criteria of a story are not arbitrary. Every book should be engaging, well crafted, and built upon fully developed and credible characters.

Although these three criteria are related, they are also independent. A story might engage us despite cartoonish characters; the story might be engaging but filled with plot gaps or typos. Therefore, when identifying appropriate criteria, we must clarify a set of *independent variables in the performance* that affect our judgment of quality. The criteria would then specify the conditions that any performance must meet to be successful; they define, operationally, the task requirements.

Many teachers make the mistake of relying on criteria that are merely easy to see as opposed to central to the performance and its purpose. So it is common to see research papers that get high scores merely for having numerous footnotes (rather than well-supported research); understanding inferred because the speech was witty (instead of thorough); or exhibits judged as effective because they are colorful and creative (as opposed to supplying accurate information). Just as we need to derive assessments from the goals and understandings, we need to derive criteria from the goals.

From criteria to rubric

A rubric is a criterion-based scoring guide consisting of a fixed measurement scale (4 points, 6 points, or whatever is appropriate) and descriptions of the characteristics for each score point. Rubrics describe degrees of quality, proficiency, or understanding along a continuum. (If the assessment response needs only a yes/no or right/wrong determination, a checklist is used instead of a rubric.) Rubrics answer the questions:

- By what criteria should performance be judged and discriminated?
- Where should we look and what should we look for to judge performance success?
- How should the different levels of quality, proficiency, or understanding be described and distinguished from one another?

Two general types of rubrics—*holistic* and *analytic*—are widely used to judge student products and performances. A holistic rubric provides an overall impression of a student's work. Holistic rubrics yield a *single* score or rating for a product or performance.

Figure 8.1
Top-Level Descriptors from an NWREL Rubric for Writing

Development of Ideas: The paper is clear and focused. It holds the reader's attention. Relevant anecdotes and details enrich the central theme.

Organization: The organization enhances and showcases the central idea or theme. The order, structure, or presentation of information is compelling and moves the reader through the text.

Voice: The writer speaks directly to the reader in a way that is individual, compelling, and engaging. The writer crafts the writing with an awareness of and respect for the audience and the purpose for writing.

Word Choice: The words convey the intended message in a precise, interesting, and natural way. The words are powerful and engaging.

Sentence Fluency: The writing has an easy flow, rhythm, and cadence. Sentences are well built, with strong and varied structure that invites expressive oral reading.

Conventions: The writer shows a good grasp of standard writing conventions . . . and uses conventions effectively to enhance readability. Errors tend to be so few that just minor touch-ups would get this piece ready to publish.

Presentation: The form and presentation of the text enhance the ability of the reader to understand and connect with the message. It is pleasing to the eye.

Source: © NWREL, Portland, OR (2000). Reprinted with permission.
Note: Numerous helpful indicators exist for each level, on a five-point scale. In addition, more learner-friendly versions for younger students have been developed. See Arter & McTighe (2001) for this and other rubrics and a comprehensive look at design and implementation issues for rubrics.

An analytic rubric divides a product or performance into distinct traits or dimensions and judges each separately. Since an analytic rubric rates each of the identified traits independently, a separate score is provided for each. For example, a popular analytic rubric for writing examines six traits: (1) ideas, (2) organization, (3) voice, (4) word choice, (5) sentence fluency, and (6) conventions. A student's writing is rated according to the performance level on each trait. For example, a piece of writing might receive a *3* for *idea development* (trait 1), and a *4* for *use of conventions* (trait 6). The Northwest Regional Educational Laboratory has developed and used a widely implemented set of analytic rubrics involving six criteria (and an optional seventh) called 6 + 1. The traits scored, with the top descriptor for each criterion, are provided in Figure 8.1.

Although a holistic rubric is an appropriate scoring tool when an overall impression is required, we propose that assessors of understanding use analytic rubrics. Why? Because the quality of the feedback to the student is easily compromised in the name of efficiency when we boil down evaluation to a

single (holistic) score. For instance, two persuasive essays may be deemed unsatisfactory, but their defects are quite different. One paper is mechanically flawed but filled with wonderful arguments. Another paper is clearly written and grammatically correct, but contains superficial reasoning and an unsupported conclusion. Yet if we are obliged to assign a single score using a holistic rubric, we unwittingly mislead the learner, the parent, and others into thinking that the performances were the same. There are always independent criteria at work in performance, especially when understanding is a target, so we should try to strike a balance between appropriately varied criteria and feasibility.

Rubrics to assess understanding

To bring this general discussion about rubrics and criteria to understanding, recall that understanding is a matter of degree on a continuum. It is not a matter of simple right versus wrong but *more or less* naïve or sophisticated, *more or less* superficial or in-depth. Thus, a rubric for understanding must provide concrete answers to our key assessment questions: What does understanding look like? What differentiates a sophisticated understanding from a naïve understanding, in practice? What does a range of explanations look like, from the most naïve or simplistic to the most complex and sophisticated?

Let's look at two examples of rubrics that describe "understanding." A generic version of a rubric used in the advanced placement exam in U.S. history in the recent past asks readers to attend to the degree to which there is a supported thesis as opposed to a mere description of events:

- *Clear, well-developed thesis that deals in a sophisticated fashion with [key] components . . .*
- *Clear, developed thesis that deals with [key issues] . . .*
- *General thesis responding to all components superficially . . .*
- *Little or no analysis . . . (Educational Testing Service/College Board, 1992, p. 25).*

The rubric explicitly warns judges, first, to assess the degree of student understanding (sophisticated analysis versus mere retelling), and second, to not confuse either the number of factual errors or the quality of the writing with the student's understanding of the time period.

Here is a rubric from a Canadian provincial language arts exam that offers a caution to judges about distinguishing between insight versus the merits of any particular interpretation:

5 Proficient: An insightful understanding of the reading selection(s) is effectively established. The student's opinion, whether directly stated or implied, is perceptive and appropriately supported by specific details. Support is precise and thoughtfully selected.

4 Capable: A well-considered understanding. . . . Opinion is thoughtful. . . . Support is well defined and appropriate.

3 Adequate: A plausible understanding is established and sustained. The student's opinion is conventional but plausibly supported. Support is general but functional.

2 Limited: Some understanding is evidenced, but the understanding is not always defensible or sustained. Opinion may be superficial and support scant and/or vague.

1 Poor: An implausible conjecture. . . . The student's opinion, if present, is inappropriate or incomprehensible. Support is inappropriate or absent.

The evaluation of the answer should be in terms of the amount of evidence that the student has actually read something and thought about it, not a question of whether he/she has thought about it in the way an adult would, or in line with an adult's "correct" answer.

In both cases, the rubrics focus on describing degrees of understanding, the trait being scored. Other traits, such as mechanics, craftsmanship, and organization should be judged separately.

We recommend that assessors consider at least two different traits, regardless of whether the descriptors are formatted as one rubric in a grid or two separate rubrics. We suggest a rubric for "understanding" and a rubric for the qualities of the "performance" (including products and processes, where appropriate) in which that understanding was displayed.

Backward design from criteria and rubrics

> It helps when the students themselves identify the characteristics of an exemplary project so that they will have a clearer understanding of the parts of the whole. This means exposing students to many student-generated and professional writing samples, guiding students to identify exactly what makes each a strong (or weak) writing piece, identifying the necessary writing skills, and teaching those skills. Students now have a "map" for each unit, [which] seems to make them much more enthusiastic about the process. With clearly defined units, more purposeful lesson plans, and more enthusiastic students, UbD has made teaching a lot more fun!
>
> —6th grade language arts teacher

Backward design suggests another approach to help us with criteria and rubrics—albeit a counterintuitive one. It turns out that any explicit goal in Stage 1 implies the criteria needed in Stage 2, even *before* a particular task is designed. For example, consider what 6th grade students in Pennsylvania will need to include in their writing to show that they have met the state writing standard:

[Students will] write persuasive pieces with a clearly stated position or opinion and supporting detail, citing sources when needed.

Regardless of whether students compose a persuasive essay, a policy brief, or a letter to the editor, the following criteria (derived directly from the standard) should be employed when judging their writing:

- Clearly stated position or opinion
- Supporting details provided
- Appropriate sources cited (as needed)

The facets and criteria

Since we have argued that understanding is revealed via six facets, these prove useful in identifying criteria and constructing rubrics to assess the degree of understanding. Figure 8.2 provides a partial list of applicable criteria based on the six facets of understanding.

How, then, might we assess for increasing control over the facets of understanding, given these criteria? The rubric shown in Figure 8.3 provides a general framework for making helpful distinctions and sound judgments. The rubric reflects an appropriate continuum—from naïve understanding (at the bottom) to sophisticated understanding (at the top)—for each of the facets.

As the rubric makes clear, understanding may be thought of as a continuum—from misconception to insight or from self-conscious awkwardness to autonomic skill proficiency. Moreover, it reflects the reality that individuals can have diverse but valid understandings of the same ideas and experiences. In other words, one person's profile might look very different from another's even as we describe them both, in general, as "sophisticated" (in the same way we give holistic scores to writing performances consisting of different patterns of the analytic traits involved).

AN IMPLICATION FOR GIVING GRADES

The regular use of criterion-based rubrics and multiple checks for understanding has implications for grading, especially at the secondary and university level. Many upper-level teachers have two long-standing habits that are counterproductive: They often give grades to each piece of work without making clear the criteria and the appropriate weighting of each criterion, and they typically average those grades over the course of time to come up with a final grade. This latter practice especially makes little sense when assessing against understanding goals and rubrics over time: Averaging a learner's initial versus final level of comprehension of a complex idea will not provide an accurate representation of her understanding. See also Guskey, 2002; Wiggins, 1998; Marzano, 2000.

Figure 8.2
Facet-Related Criteria

Facet 1 Explanation	Facet 2 Interpretation	Facet 3 Application	Facet 4 Perspective	Facet 5 Empathy	Facet 6 Self-knowledge
• accurate	• meaningful	• effective	• credible	• sensitive	• self-aware
• coherent	• insightful	• efficient	• revealing	• open	• metacognitive
• justified	• significant	• fluent	• insightful	• receptive	• self-adjusting
• systematic	• illustrative	• adaptive	• plausible	• perceptive	• reflective
• predictive	• illuminating	• graceful	• unusual	• tactful	• wise

Figure 8.3
Six-Facet Rubric

Explained	Meaningful	Effective	In Perspective	Empathic	Reflective
Sophisticated and Comprehensive: an unusually thorough, elegant, or inventive account (model, theory, explanation); fully supported, verified, justified; deep and broad; goes well beyond the information given	*Insightful:* a powerful and illuminating interpretation or analysis of the importance, meaning, significance; tells a rich and insightful story; provides a revealing history or context	*Masterful:* Fluent, flexible, efficient, able to use knowledge and skill and adjust understandings well in diverse and difficult contexts—masterful ability to transfer	*Insightful and Coherent:* a thoughtful and circumspect viewpoint; effectively critiques, encompasses other plausible perspectives; takes a long and dispassionate critical view of the issues involved	*Mature:* disciplined; disposed and able to see and feel what others see and feel; unusually open to and willing to seek out the odd, alien, or different; able to make sense of texts, experiences, events that seem weird to others	*Wise:* deeply aware of the boundaries of own and others' understanding; able to recognize own prejudices and projections; has integrity—able and willing to act on understanding
Systematic: an atypical and revealing account, going beyond what is obvious or what was explicitly taught; makes subtle connections; well supported by argument and evidence; novel thinking displayed	*Revealing:* a thoughtful interpretation or analysis of the importance, meaning, significance; tells an insightful story; provides a helpful history or context	*Skilled:* competent in using knowledge and skill and adapting understandings in a variety of appropriate and demanding contexts	*Thorough:* a fully developed and coordinated critical view; makes own view more plausible by a fair consideration of the plausibility of other perspectives; makes apt criticisms, discriminations, and qualifications	*Sensitive:* disposed to see and feel what others see and feel; open to the unfamiliar or different; able to see the value and work that others do not see	*Circumspect:* aware of own ignorance and that of others; aware of own prejudices
In-Depth: an account that reflects some in-depth and personalized ideas; student is making the work his own, going beyond the given; there is supported theory, but insufficient or inadequate evidence and argument	*Perceptive:* a reasonable interpretation or analysis of the importance, meaning, or significance; tells a clear and instructive story; provides a revealing history or context	*Able:* limited but growing ability to be adaptive and innovative in the use of knowledge and skill	*Considered:* a reasonably critical and comprehensive look at major points of view in the context of her own; makes clear that there is plausibility to other points of view	*Aware:* knows and feels that others see and feel differently and is somewhat able to empathize with others	*Thoughtful:* generally aware of what he does and does not understand; aware of how prejudice and projection occur without awareness

Developed: an incomplete account, but with apt and insightful ideas; extends and deepens some of what was learned; some reading between the lines; account has limited support, argument, data, or sweeping generalizations; there is a theory with limited testing and evidence	*Interpreted:* a plausible interpretation or analysis of the importance, meaning, or significance; makes sense with a story; provides a telling history or context	*Apprentice:* relies on a limited repertoire of routines, able to perform well in a few familiar or simple contexts; limited use of judgment and responsiveness to feedback or situation	*Aware:* knows of different points of view and somewhat able to place own view in perspective, but weakness in considering worth of each perspective or critiquing each perspective, especially her own; uncritical about tacit assumptions	*Decentering:* has some capacity or self-discipline to walk in others shoes, but is still primarily limited to own reactions and attitudes, puzzled or put off by different feelings or attitudes	*Unreflective:* generally unaware of own specific ignorance; generally unaware of how prejudgments color understanding
Naïve: superficial account; more descriptive than analytical or creative; a fragmented or sketchy account of facts, ideas; glib generalizations; a black-and-white account; less theory than an unexamined hunch or borrowed idea	*Literal:* a simplistic or superficial reading; mechanical translation; a decoding with little or no interpretation; no sense of wider importance or significance; a restatement of what was taught or read	*Novice:* can perform only with coaching or relies on highly scripted, singular "plug-in" (algorithmic and mechanical) skills, procedures, or approaches	*Uncritical:* unaware of differing points of view, prone to overlook or ignore other perspectives; has difficulty imagining other ways of seeing things; prone to ad hominem criticisms	*Egocentric:* has little or no empathy, beyond intellectual awareness of others; see things through own ideas and feelings; ignores or is threatened or puzzled by different feelings, attitudes, views	*Innocent:* completely unaware of the bounds of own understanding and of the role of projections and prejudice in opinions and attempts to understand

The criteria, hence rubrics, are piling up! A practical strategy for addressing this complexity is to frame multiple rubrics in light of the fewest key differing aspects of understanding, knowledge, and skill. Here is an example of a set of five criteria in mathematics (edited to just the top score for each of the five rubrics), which can be used to assess the key dimensions of most complex mathematical performance:

MISCONCEPTION ALERT!

Where do the most appropriate criteria and indicators come from? How do rubrics move from general to specific descriptors? The answers involve yet another element of backward design: For the descriptors to be appropriate, detailed, and helpful, they must emerge from reviews of many concrete samples of work. The descriptors reflect the distinguishing characteristics of the pile of work at that level. Thus, a rubric is never complete until it has been used to evaluate student work *and* an analysis of different levels of work is used to sharpen the descriptors.

- Mathematical Insight: Shows a sophisticated understanding of the subject matter involved. The concepts, evidence, arguments, qualifications made, questions posed, and methods used are expertly insightful, going well beyond the grasp of the topic typically found at this level of experience. Grasps the essence of the problem and applies the most powerful tools for solving it. The work shows that the student is able to make subtle distinctions and to relate the particular problem to more significant, complex, or comprehensive mathematical principles, formulas, or models.
- Reasoning: Shows a methodical, logical, and thorough plan for solving the problem. The approach and answers are explicitly detailed and reasonable throughout (whether the knowledge used is sophisticated or accurate). The student justifies all claims with thorough argument: Counterarguments, questionable data, and implicit premises are fully explicated.
- Effectiveness of Solution: The solution to the problem is effective and often inventive. All essential details of the problem, and audience, purpose, and other contextual matters, are fully addressed in a graceful and effective way. The solution may be creative in many possible ways: an unorthodox approach, unusually clever juggling of conflicting variables, the bringing in of unobvious mathematics, or imaginative evidence.
- Accuracy of Work: The work is accurate throughout. All calculations are correct, provided to the proper degree of precision and measurement error, and properly labeled.
- Quality of Presentation: The student's performance is persuasive and unusually well presented. The essence of the research and the problems to be solved are summed up in a highly engaging and efficient manner, mindful of the audience and the purpose of the presentation. Craftsmanship in the final product is obvious. Effective use is made of supporting material (e.g., visuals, models, overheads, and videos) and of team members (where appropriate). The audience shows enthusiasm and confidence that the presenter understands what she is talking about and understands the listeners' interests.

If the thought of using so many rubric traits seems overwhelming, start small. Go back to the two basic criteria—quality of the understandings and the quality of the performance. Add a third for process when appropriate, and other rubric traits as time and interest permit. Later, when you have identified multiple traits, use only parts of the set, as appropriate to each assignment. (In the chapter on Macro Design issues, we will argue that sets of such rubrics should be established at the Program level.)

Designing and refining rubrics based on student work

Important criteria for evaluating student understanding and proficiency are initially derived from the desired results of Stage 1. Yet as the Misconception Alert makes clear, the process of building and revising a rubric also relies on an analysis of student performance. The following is a summary of the six-step process that Arter and McTighe (2001, pp. 37–44) propose for analyzing student performance:

Step 1: Gather samples of student performance that illustrate the desired understanding or proficiency. *Choose as large and diverse a set of samples as possible.*

Step 2: Sort student work into different "stacks" and write down the reasons. *For example, place the samples of student work into three piles: strong, middle and weak. As the student work is sorted, write down reasons for placing pieces in the various stacks. If a piece is placed in the "sophisticated" pile, describe its distinguishing features. What cues you that the work reflects sophisticated understanding? What are you saying to yourself as you place a piece of work into a pile? What might you say to a student as you return this work? The qualities or attributes that you identify reveal the important criteria indicators. Keep sorting work until you are not adding anything new to your list of attributes.*

Step 3: Cluster the reasons into traits or important dimensions of performance. *The sorting process used thus far in this exercise is "holistic." Participants in this process end up with a list of comments for high, medium and low performance; any single student product gets only one overall score. Usually, during the listing of comments someone will say something to the effect that, "I had trouble placing this paper into one stack or another because it was strong on one trait but weak on another." This brings up the need for analytical trait scoring systems; i.e., evaluating each student's product or performance on more than one dimension.*

Step 4: Write a definition of each trait. *These definitions should be "value neutral"—they describe what the trait is about, not what good performance looks like. (Descriptions of good performance on the trait are accorded to the "highest" rubric rating.)*

Step 5: Select samples of student performance that illustrate each score point on each trait. *Find samples of student work that illustrate strong, weak and mid range performance on each trait. These examples are sometimes called "anchors" since they provide concrete examples of the levels in a rubric. The anchors can be used to help students come to understand what "good" looks like. (Note: It's important to have more than a single example. If you show students only a single example of what a good performance looks like, they are likely to imitate or copy it.)*

Step 6: Continuously refine. *Criteria and rubrics evolve with use. As you try them out, invariably you will find some parts of the rubric that work fine and some that don't. Add and modify descriptions so that they communicate more precisely, and choose better anchors that illustrate what you mean.*

The challenge of validity

The third question in thinking like an assessor asks us to be careful that we evoke the most appropriate evidence, namely evidence of the desired results of Stage 1. We are not trying to create *merely* interesting and realistic tasks in Stage 2 but to obtain the most appropriate evidence of the desired results framed in Stage 1. This is the challenge of validity.

Validity refers to the meaning we can and cannot properly make of specific evidence, including traditional test-related evidence. We see a student commit a kind act on the playground. What should we infer about that student's propensity to "be kind"? That's the challenge of validity: At what events or data should we look to obtain the most telling evidence of more general abilities?

Consider the challenge currently in any conventional classroom. Mrs. Metrikos, a 6th grade teacher at Carson Middle School, makes up a 20-problem test on fractions. Jose gets 11 right. The teacher infers that Jose's control of the *entire realm of fractions* is very shaky. Valid conclusion? Not necessarily. First, we need to look at the test items and determine if they are representative of all types of problems with fractions. Given that Jose is a recent immigrant, maybe his English is weak but his math strong; does the test factor out the English to let us see only his math ability? Is the test so laden with word problems that the test is really a test of English comprehension? What about the relative difficulty of the problems? Each question counted the same as the others. But what if some are much harder than others?

In scoring the test, Mrs. Metrikos focused solely on the correctness of the answers, ignoring the process each student used to set up and solve each problem. Is correctness indicative of understanding? Not necessarily. The best test papers may simply reflect recall of the formulas involved, without any understanding of why they work. Further, what should we infer when Jose runs up after the papers are handed back to explain his understanding of fractions and why his mistakes were "just" carelessness. Should that affect his grade or our understanding of his understanding? Perhaps as Mrs. Metrikos looks over

the results that evening, she sees not only that Jose seemed to have trouble with the English in the word problems, but that Jose has trouble with fractions in which the denominators differ, but had no difficulty in explaining the rule and why you need a common denominator. To say that Jose "doesn't understand" fractions based on the wrong answers is thus an invalid conclusion.

A focus on understanding makes the issue of validity challenging in any assessment. Suppose Jenny got 19 of the 20 problems right, but the one she got wrong asked for an explanation as to why common denominators are needed. Suppose Sara gets all the history facts right on the multiple-choice test part of her history exam, but completely fails the document-based question that calls for analysis of key events during the same time frame? What if Ian does a superb poster on the water cycle, but fails the quiz? These are the challenges that face us all. We have to be sure that the performances we demand are appropriate to the particular understandings sought. Could a student perform well on the test without understanding? Could a student with understanding nonetheless forget or jumble together key facts? Yes and yes—it happens all the time. We want to avoid doubtful inferences when assessing any student work, but especially so when assessing for understanding.

As we noted earlier, understanding is a matter of degree. As the fraction example suggests, we typically pay too much attention to *correctness* (in part because scoring for correctness makes assessment so much easier and seemingly "objective"—machines can do it) and too little attention to the *degree* of understanding (in which someone has to make a valid judgment). So understanding easily falls through the cracks of typical testing and grading.

The issue is made harder still by a common confusion in performance assessment design. Many teacher-designers confuse interesting and engaging learning activities with appropriate evidence from performance. Just because the performance is complex and the task interesting, it doesn't follow that the evidence we gain from student project work is appropriate for the desired results.

We can sum up the challenge in the story about a 5th grade teacher in Virginia. She proposed assessing her students' mastery of standards related to the Civil War by having them construct a diorama. She was developing a unit on the Civil War in a workshop where the goal was twofold: Find creative ways to address the state standards, and honor UbD ideas. She was trying to assess her students' understanding of the causes and effects of the Civil War through the use of an engaging performance task.

She asked if she could use a tried and true project (one that the "kids love") since it involved performance and yielded an assessable product. We said that, in the abstract, there was no reason not to, as long as the project would generate the right kind of evidence. She wasn't sure what we meant, so we asked her to describe the project. Well, she said, the kids must build a diorama of one great battle in the Civil War for a simulated Civil War museum. There have to be maps, explanatory plaques, and relevant artifacts. So we asked for the particulars of the state standard:

AN ESSENTIAL QUESTION ABOUT INSIGHT REMAINS

This discussion about validity does not directly address or settle a long-standing controversy among philosophers and psychologists: Whether the act of understanding primarily involves a mental picture separate from the performance. To frame it as a cognitive research essential question, the debate involves asking: Is performance ability necessarily *preceded* by a mental model? Or is understanding more like successful jazz improvisation—something that is *inherently* a performance ability and sensitivity in which prior deliberate thought plays no critical or determining role? Although we don't take sides here, readers interested in the issue might want to read Gilbert Ryle's *The Concept of Mind* (1949), Perkins's chapter in *Teaching for Understanding* (Wiske, 1998), and *The Nature of Insight* (Sternberg & Davidson, 1995).

Civil War and Reconstruction: 1860s to 1877

USI.9 The student will demonstrate knowledge of the causes, major events, and effects of the Civil War by

a. describing the cultural, economic, and constitutional issues that divided the nation;
b. explaining how the issues of states' rights and slavery increased sectional tensions;
c. identifying on a map the states that seceded from the Union and those that remained in the Union;
d. describing the roles of Abraham Lincoln, Jefferson Davis, Ulysses S. Grant, Robert E. Lee, Thomas "Stonewall" Jackson, and Frederick Douglass in events leading to and during the war;
e. using maps to explain critical developments in the war, including major battles;
f. describing the effects of war from the perspectives of Union and Confederate soldiers (including black soldiers), women, and slaves.

We responded by asking her to self-assess the proposed assessment task design against two questions. How likely is it that:

- A student could do well on this performance task, but really not demonstrate the understandings you are after?
- A student could perform poorly on this task, but still have significant understanding of the ideas and show them in other ways?

If the answer to either question is "yes," then the assessment will probably *not* provide valid evidence.

"Oh, of course!" she quickly said. "How could I have been so foolish? It really only gets at a small slice of the standards, and bypasses entirely the issue of cause and effect. How did I miss that?"

Her mistake is a common one—confusing interesting projects or authentic *activities* with valid assessments. In this case, she had taken one small link between her project and the standard (the major military turning points) and tried to draw a conclusion from the evidence that was not warranted. The good news? When asked to self-assess against the two validity questions, she saw the problem immediately. The bad news? Most people don't self-assess their proposed assessments against any design standards, and they often end up with invalid inferences. The aim of Stage 2 is not engaging work; the aim is good evidence for judging achievement against stated goals.

The anecdote also reminds us of the importance of deriving the general criteria from the goals. Given that the content standard focused on *causes and effects* of the Civil War, if the teacher had considered appropriate criteria related to the standard *prior* to designing the specific diorama task, she may have averted the validity problem. In terms of assessing for causal reasoning, any student performance would need to (1) identify multiple causes, (2) identify multiple effects, (3) be historically accurate, and (4) include a clear explanation. Thinking this way also suggests other, more appropriate task possibilities, such as a cause-effect poster showing multiple causes and multiple effects of the war.

The analysis illustrates nicely the paradox of designing local assessments: Left to our own instincts, seeing validity issues is very difficult. With a little disciplined self-assessment against the right standards (not to mention some quick peer review), however, we can solve most of the problems that we encounter.

Backward design to the rescue

Recall the horizontal version of the Template (Figure 7.2, p. 149) and see how it asks us to look at the logical links between Stage 1 and Stage 2. Notice in Figure 8.4 how backward design, using two of the six facets, helps us to better "think like an assessor."

To become more attentive to issues of validity, designers are encouraged to regularly apply the self-test in Figure 8.5 to their current (or past) assessments, which expands on this line of questioning and can be used for any assessment design idea, past or future, to improve validity.

Your answers will likely be less than certain, of course. There are no rules or recipes in validity. Sometimes we just have to make a thoughtful judgment, mindful of our fallibility. But don't underestimate the power of self-assessment in design. It can solve many of your problems and make you more confident and courageous as an assessor—so that you assess what really matters, not merely what is easy to see and score.

MISCONCEPTION ALERT!

Validity is about inference, not the test itself. Validity concerns the meaning of evidence: what we ask students to do, and how we assess the resulting work. In other words, validity is about our understanding of the results, not the test itself. We have to be a bit more careful in our talk. Although everyone casually uses the words "valid" and "invalid" as adjectival modifiers of "test," strictly speaking this is inaccurate. Validity is about the *inferences* we try to make from particular test results. And sharpening the power of those inferences is key to becoming a better assessor.

Figure 8.4
Using Backward Design to Think like an Assessor

Stage 1	Stage 2	
If the desired result is for learners to . . .	***Then you need evidence of the student's ability to . . .***	***So the assessments need to include some things like . . .***
Understand that . . . (U) • Statistical analysis and graphic display often reveal patterns in data. • Pattern recognition enables prediction. • Inferences from data patterns can be plausible but invalid (as well as implausible but valid). • Correlation does not ensure causality. (Q) **And thoughtfully consider the questions . . .** • What's the trend? • What will happen next? • In what ways can data and statistics "lie" as well as reveal?	APPLY: What applications would enable us to infer student understanding of what they have learned? What kinds of performances and products, if done well, would provide valid ways of distinguishing between understanding and mere recall? EXPLAIN: What must students be able to explain, justify, support, or answer about their work for us to infer genuine understanding? How can we test their ideas and applications to find out if they really understand what they have said and done?	(T) (OE) • Using past performances in the men's and women's marathon, predict the men's and women's marathon times for 2020. • Chart various scenarios for a savings program (e.g., for college, retirement). Give financial advice. Explain the implausibility of compound interest. • Analyze the past 15 years of AIDS cases to determine the trend. (Note: The data start out looking linear but become exponential.) • Write an article or a letter to the editor about why the marathon analysis is plausible but incorrect. • Develop a brochure to would-be investors on why early saving with small amounts is better than later with large amounts. • Create a graphic display with accompanying written explanation to illustrate the exponential nature of AIDS cases.

Figure 8.5
Self-Test of Assessment Ideas

Stage 1 — **Desired Results:**

Stage 2 — **Proposed Assessment:**

	very likely	somewhat likely	very unlikely
How likely is it that a student could do* well *on the assessment by			
1. Making clever guesses based on limited understanding?	❑	❑	❑
2. Parroting back or plugging in what was learned, with accurate recall but limited or no understanding?	❑	❑	❑
3. Making a good-faith effort, with lots of hard work and enthusiasm, but with limited understanding?	❑	❑	❑
4. Producing lovely products and performances, but with limited understanding?	❑	❑	❑
5. Applying natural ability to be articulate and intelligent, with limited understanding of the content in question?	❑	❑	❑
How likely is it that a student could do* poorly *on the assessment by			
6. Failing to meet the performance goals despite having a deep understanding of the big ideas? (For example, the task is not relevant to the goals.)	❑	❑	❑
7. Failing to meet the scoring and grading criteria used, despite having a deep understanding of the Big Ideas? (For example, some of the criteria are arbitrary, placing undue or inappropriate emphasis on things that have little to do with the desired results or true excellence at such a task.)	❑	❑	❑

Goal: Make all your answers "very unlikely"

Validity affects rubric design, too. Validity issues arise in rubrics, not just tasks. We have to make sure that we employ the right criteria for judging understanding (or any other target), not just what is easy to count or score. In assessing for understanding we must especially beware of confusing mere correctness or skill in performance (i.e., writing, PowerPoint, graphic representations) with degree of understanding. A common problem in assessment is that many scorers presume greater understanding in the student who knows all the facts or communicates with elegance versus the student who makes mistakes or communicates poorly. But what if the findings of the papers with mistakes are truly insightful and the paper that is well written and based on facts is superficial? Getting clear on what we can and cannot conclude from the evidence—that's always the issue in validity, and it applies to how we score, not just what we score.

In practice, variants of the two questions asked earlier also help us self-assess the validity of criteria and rubrics. Given the criteria you are proposing and the rubrics being drafted from them, consider

- Could the proposed criteria be met but the performer still not demonstrate deep understanding?
- Could the proposed criteria not be met but the performer nonetheless still show understanding?

If your answer to either question is yes, then the proposed criteria and rubric are not yet ready to provide valid inferences.

Reliability: Our confidence in the pattern

A discussion on the appropriateness of the assessment evidence is vital but not sufficient. We need not only a valid inference but a trustworthy one. We need to be confident that a result reflects a pattern. Maybe Jose's 9 errors out of 20 would only end up being 9 out of 50 if he were given another test the next day. The proposed test might be appropriate, but a single result on it unreliable or anomalous. This is the problem of reliability and why we argued in Chapter 7 for having a scrapbook of evidence as opposed to a single snapshot.

Consider your favorite winning sports team to see the reliability problem. Their performance in games is surely an appropriate measure of their achievement. Game results yield valid inferences about achievement in the sport, by definition. But any one game result might not be representative. Consider any night on which the team was upset by a historically weak team. That score is out of the ordinary—unreliable—once we have many results in hand, because the team did quite well over the entire season. Reliable assessments reveal a credible pattern, a clear trend.

Please note that whether various judges agree with one another is a different problem, usually termed "inter-rater reliability." In that case, we want the judgments of multiple judges to form a consistent pattern. But those multiple judges might still only be scoring a single event. In that case, the judges

could be reliable, that is, they could all give the same score, but the performance that day may not be "reliable" or typical of the student's pattern of performance.

A second aphorism we like to use in framing the challenge of assessment (in addition to "innocent until proven guilty") is a famous line by Binet, the creator of the IQ test and the founder of modern measurement techniques: "It doesn't matter what tests you use as long as they are *varied* and *many*." That's why in Understanding by Design we ask designers to use a mix of different types of evidence over time.

General guidelines

We can sum up the concerns in Chapters 7 and 8 by offering the following questions and guidelines to consider when constructing a balanced set of local assessments of understanding:

1. The needed evidence is inherently less direct and more complicated than that obtained from objective tests to assess knowledge and skill. We need to look at more than just the percentage of correct answers. Why? Sometimes getting the right answer occurs as a result of rote recall, good test-taking skills, or lucky guessing. In assessing for understanding, we need to ferret out the reasons behind the answers and what meaning the learner makes of the results.

2. Assessment of understanding requires evidence of "application" in performance or products, but that complicates judging results. What do we do when parts of a complex performance are shaky, but we discern clear insight in the content? Or the result is fine, yet we sense that little insight was required to complete the project? How do we design performances that enable us to make precise judgments about the different parts of performance?

3. Since understanding involves the six facets, do some facets take precedence over others? *Which* performances matter most, in *what* situations? What can we infer, for instance, when the "application" and "explanation" of strategy is strong but the "interpretation" of the situation is weak? Or the particular "application" was ineffective, but verbal analysis and self-assessment makes clear that the learner has a solid understanding of the content and process?

4. Try to have parallel versions of the same content across different assessment formats. In other words, counteract the "messiness" of a complex task with a simple quiz in the same content. Or use constructed response questions on the same content to make sure that correct answers cannot hide lack of understanding. Whenever possible, have parallel assessments in diverse formats improve the quality of the evidence of desired results.

5. Try to anticipate key misunderstandings and develop quick preassessments and postassessments to find out if those misunderstandings were overcome—regardless of what other assessment tasks you are using. For example, the following quick assessment task reveals whether students understand the process of isolating variables as part of a science investigation:

Roland wants to decide which of two spot removers is best. First, he tried Spot Remover A on a T-shirt that had fruit stains and chocolate stains. Next, he tried Spot Remover B on jeans that had grass stains and rust stains. Then he compared the results. Is there a problem with Roland's plan that will make it hard for him to know which spot remover is best? Explain.

6. Given that a single application or product may or may not link to larger goals, regularly ask students to "show their work," give reasons for answers, and show connections to larger principles or ideas in the answers.

7. Given that an articulate explanation may be more a function of verbal ability and verbal knowledge with no real understanding, ask the student to "transfer" that explanation to a new or different problem, situation, or issue.

8. Tap into various facets to broaden the evidence: When demanding a hands-on application (Facet 3), also require interpretation (Facet 2), and self-assessment (Facet 6) to make sure that the final product is not overvalued. Require a blend of perspective and empathy whenever possible.

A caveat before closing

Although we've concentrated on more formal and *summative* assessments of understanding in this chapter, daily teacher checks are the vehicles through which we monitor whether students understand. The iterative nature of understanding, the likelihood of confusions or misconceptions, and the need for interactive evidence make it imperative, in fact, that teachers know how to use ongoing assessments to inform their teaching and needed adjustments. Since Stage 2 is about summative assessment, we postpone a further consideration of informal checks for understanding and feedback until Stage 3.

We have postponed for many chapters the work we all typically like to do most: the design of the learning plan. Stage 3 now beckons, where we determine more fully what the learning plan needs to accomplish, given not only the desired understandings and assessment evidence, but who our learners are and what is in their best interest.