Four themes guide the discussion in this chapter of how advances in the cognitive sciences and new approaches to measurement have created opportunities, not yet fully realized, for assessments to be used in ways that better serve the goals of learning.

- One type of assessment does not fit all. The purpose and context of an assessment set priorities and constraints on the design. The power of classroom assessment resides in its close connections to instruction and teachers’ knowledge of their students’ instructional histories. Large-scale, standardized assessments can communicate across time and place, but by so constraining the content and timeliness of the message that they often have limited utility in the classroom. These kinds of trade-offs are an inescapable aspect of assessment design.
- It is in the context of classroom assessment that the most significant benefit can be gained from advances in cognitive theory. Learning is enhanced by assessment that provides feedback to students about particular qualities of their work and what they can do to improve their understanding. To provide this kind of information, teachers must have a foundation of knowledge about how students learn the subject matter.
- Large-scale assessments are further removed from instruction, but can still benefit learning if well designed and properly used. They can signal worthy goals and display to the public what competency in a domain looks like, along with typical learning pathways. They can also play an important role in communicating and fostering public dialogue about educational goals. However, fully capitalizing on a merger of cognitive and measurement principles will require relaxing some of the constraints that drive current large-scale assessment practices.
- Multiple measures are needed to serve the assessment needs of an educational system. Currently, however, conflicts between classroom and large-scale assessments in terms of both goals and feedback cause confusion for educators, students, and parents. We describe a vision of coordinated systems of assessment in which multiple assessments work together, along with curriculum and instruction, to support a shared set of learning goals. In this vision, a greater portion of the investment in assessment is shifted toward the classroom, where it can be used most directly to assist learning.
Although assessments are currently used for many purposes in the educational system, a premise of this report is that their effectiveness and utility must ultimately be judged by the extent to which they promote student learning. The aim of assessment should be “to educate and improve student performance, not merely to audit it” (Wiggins, 1998, p.7). To this end, people should gain important and useful information from every assessment situation. In education, as in other professions, good decision making depends on access to relevant, accurate, and timely information. Furthermore, the information gained should be put to good use by informing decisions about curriculum and instruction and ultimately improving student learning (Falk, 2000; National Council of Teachers of Mathematics, 1995).

Assessments do not function in isolation; an assessment’s effectiveness in improving learning depends on its relationships to curriculum and instruction. Ideally, instruction is faithful and effective in relation to curriculum, and assessment reflects curriculum in such a way that it reinforces the best practices in instruction. In actuality, however, the relationships among assessment, curriculum, and instruction are not always ideal. Often assessment taps only a subset of curriculum and without regard to instruction, and can narrow and distort instruction in unintended ways (Klein, Hamilton, McCaffrey, and Stecher, 2000; Koretz and Barron, 1998; Linn, 2000; National Research Council [NRC], 1999b). In this chapter we expand on the idea, introduced in Chapter 2, that synergy can best be achieved if the three parts of the system are bound by or grow out of a shared knowledge base about cognition and learning in the domain.
PURPOSES AND CONTEXTS OF USE

Educational assessment occurs in two major contexts. The first is the classroom. Here assessment is used by teachers and students mainly to assist learning, but also to gauge students’ summative achievement over the longer term. Second is large-scale assessment, used by policy makers and educational leaders to evaluate programs and/or obtain information about whether individual students have met learning goals.

The sharp contrast that typically exists between classroom and large-scale assessment practices arises because assessment designers have not been able to fulfill the purposes of different assessment users with the same data and analyses. To guide instruction and monitor its effects, teachers need information that is intimately connected with the work their students are doing, and they interpret this evidence in light of everything else they know about their students and the conditions of instruction. Part of the power of classroom assessment resides in these connections. Yet precisely because they are individualized and highly contextualized, neither the rationale nor the results of typical classroom assessments are easily communicated beyond the classroom. Large-scale, standardized tests do communicate efficiently across time and place, but by so constraining the content and timeliness of the message that they often have little utility in the classroom. This contrast illustrates the more general point that one size of assessment does not fit all. The purpose of an assessment determines priorities, and the context of use imposes constraints on the design, thereby affecting the kinds of information a particular assessment can provide about student achievement.

Inevitability of Trade-Offs in Design

To say that an assessment is a good assessment or that a task is a good task is like saying that a medical test is a good test; each can provide useful information only under certain circumstances. An MRI of a knee, for example, has unquestioned value for diagnosing cartilage damage, but is not helpful for diagnosing the overall quality of a person’s health. It is natural for people to understand medical tests in this way, but not educational tests. The same argument applies nonetheless, but in ways that are less familiar and perhaps more subtle.

In their classic text *Psychological Tests and Personnel Decisions*, Cronbach and Gleser (1965) devote an entire chapter to the trade-off between *fidelity* and *bandwidth* when testing for employment selection. A high-fidelity, narrow-bandwidth test provides accurate information about a small number of focused questions, whereas a low-fidelity, broad-bandwidth test provides noisier information for a larger number of less-focused questions. For a
fixed level of resources—the same amount of money, testing time, or tasks—the designer can choose where an assessment will fall along this spectrum. Following are two examples related to the fidelity-bandwidth (or depth versus breadth) trade-offs that inevitably arise in the design of educational assessments. They illustrate the point that the more purposes one attempts to serve with a single assessment, the less well that assessment can serve any given purpose.

**Trade-Offs in Assessment Design: Examples**

**Accountability Versus Instructional Guidance for Individual Students**

The first example expands on the contrast between classroom and large-scale assessments described above. A starting point is the desire for statewide accountability tests to be more helpful to teachers or the question of why assessment designers cannot incorporate in the tests items that are closely tied to the instructional activities in which students are engaged (i.e., assessment tasks such as those effective teachers use in their classrooms). To understand why this has not been done, one must look at the distinct purposes served by standardized achievement tests and classroom quizzes: who the users are, what they already know, and what they want to learn.

In this example, the chief state school officer wants to know whether students have been studying the topics identified in the state standards. (Actually, by assessing these topics, the officer wants to increase the likelihood that students will be studying them.) But there are many curriculum standards, and she or he certainly cannot ascertain whether each has been studied by every student. A broad sample from each student is better for his or her purposes—not enough information to determine the depth or the nature of any student’s knowledge across the statewide curriculum, but enough to see trends across schools and districts about broad patterns of performance. This information can be used to plan funding and policy decisions for the coming year.

The classroom teacher wants to know how well an individual student, or class of students, is learning the things they have been studying and what they ought to be working on next. What is important is the match among what the teacher already knows about the things students have been working on, what the teacher needs to learn about their current understanding, and how that knowledge will help shape what the students should do now to learn further.

For the chief state school officer, the ultimate question is whether larger aggregates of students (such as schools, districts, or states) have had “the opportunity to learn.” The state assessment is constructed to gather information to support essentially the same inference about all students, so the
information can most easily be combined to meet the chief officer’s purpose. For the teacher, the starting point is knowing what each student as an individual has had the opportunity to learn. The classroom quiz is designed to reveal patterns of individual knowledge (compared with the state grade-level standards) within the small content domain in which students have been working so the teacher can make tailored decisions about next steps for individual students or the class. For the teacher, combining information across classes that are studying and testing different content is not important or possible. Ironically, the questions that are of most use to the state officer are of the least use to the teacher.

National Assessment of Educational Progress (NAEP): Estimates for Groups Versus Individual Students

The current public debate over whether to provide student-level reports from NAEP highlights a trade-off that goes to the very heart of the assessment and has shaped its sometimes frustratingly complex design from its inception (see Forsyth, Hambleton, Linn, Mislevy, and Yen, 1996 for a history of NAEP design trade-offs). NAEP was designed to survey the knowledge of students across the nation with respect to a broad range of content and skills, and to report the relationships between that knowledge and a large number of educational and demographic background variables. The design selected by the founders of NAEP (including Ralph Tyler and John Tukey) to achieve this purpose was multiple-matrix sampling. Not all students in the country are sampled. A strategically selected sample can support the targeted inferences about groups of students with virtually the same precision as the very familiar approach of testing every student, but for a fraction of the cost. Moreover, not all students are administered all items. NAEP can use hundreds of tasks of many kinds to gather information about competencies in student populations without requiring any student to spend more than a class period performing those tasks; it does so by assembling the items into many overlapping short forms and giving each sampled student a single form.

Schools can obtain useful feedback on the quality of their curriculum, but NAEP’s benefits are traded off against several limitations. Measurement at the level of individual students is poor, and individuals can not be ranked, compared, or diagnosed. Further analyses of the data are problematic. But a design that served any of these purposes well (for instance, by testing every student, by testing each student intensively, or by administering every student parallel sets of items to achieve better comparability) would degrade the estimates and increase the costs of the inferences NAEP was created to address.
Reflections on the Multiple Purposes for Assessment

As noted, the more purposes a single assessment aims to serve, the more each purpose will be compromised. Serving multiple purposes is not necessarily wrong, of course, and in truth few assessments can be said to serve a single purpose only. But it is incumbent on assessment designers and users to recognize the compromises and trade-offs such use entails. We return to notions of constraints and trade-offs later in this chapter.

Multiple assessments are thus needed to provide the various types of information required at different levels of the educational system. This does not mean, however, that the assessments need to be disconnected or working at cross-purposes. If multiple assessments grow out of a shared knowledge base about cognition and learning in the domain, they can provide valuable multiple perspectives on student achievement while supporting a core set of learning goals. Stakeholders should not be unduly concerned if differing assessments yield different information about student achievement; in fact, in many circumstances this is exactly what should be expected. However, if multiple assessments are to support learning effectively and provide clear and meaningful results for various audiences, it is important that the purposes served by each assessment and the aspects of achievement sampled by any given assessment be made explicit to users.

Later in the chapter we address how multiple assessments, including those used across both classroom and large-scale contexts, could work together to form more complete assessment systems. First, however, we discuss classroom and large-scale assessments in turn and how each can best be used to serve the goals of learning.

CLASSROOM ASSESSMENT

The first thing that comes to mind for many people when they think of “classroom assessment” is a midterm or end-of-course exam, used by the teacher for summative grading purposes. But such practices represent only a fraction of the kinds of assessment that occur on an ongoing basis in an effective classroom. The focus in this section is on assessments used by teachers to support instruction and learning, also referred to as formative assessment. Such assessment offers considerable potential for improving student learning when informed by research and theory on how students develop subject matter competence.

As instruction is occurring, teachers need information to evaluate whether their teaching strategies are working. They also need information about the current understanding of individual students and groups of students so they can identify the most appropriate next steps for instruction. Moreover, students need feedback to monitor their own success in learning and to know
how to improve. Teachers make observations of student understanding and performance in a variety of ways: from classroom dialogue, questioning, seatwork and homework assignments, formal tests, less formal quizzes, projects, portfolios, and so on.

Black and Wiliam (1998) provide an extensive review of more than 250 books and articles presenting research evidence on the effects of classroom assessment. They conclude that ongoing assessment by teachers, combined with appropriate feedback to students, can have powerful and positive effects on achievement. They also report, however, that the characteristics of high-quality formative assessment are not well understood by teachers and

BOX 6-1 Transforming Classroom Assessment Practices

A project at King’s College London (Black and Wiliam, 2000) illustrates some of the issues encountered when an effort is made to incorporate principles of cognition and reasoning from evidence into classroom practice. The project involved working closely with 24 science and mathematics teachers to develop their formative assessment practices in everyday classroom work. During the course of the project, several aspects of the teaching and learning process were radically changed.

One such aspect was the teachers’ practices in asking questions in the classroom. In particular, the focus was on the notion of wait time (the length of the silence a teacher would allow after asking a question before speaking again if nobody responded), with emphasis on how short this time usually is. The teachers altered their practice to give students extended time to think about any question posed, often asking them to discuss their ideas in pairs before calling for responses. The practice of students putting up their hands to volunteer answers was forbidden; anyone could be asked to respond. The teachers did not label answers as right or wrong, but instead asked a student to explain his or her reasons for the answer offered. Others were then asked to say whether they agreed and why. Thus questions opened up discussion that helped expose and explore students’ assumptions and reasoning. At the same time, wrong answers became useful input, and the students realized that the teacher was interested in knowing what they thought, not in evaluating whether they were right or wrong. As a consequence, teachers asked fewer questions, spending more time on each.
In addition, teachers realized that their lesson planning had to include careful thought about the selection of informative questions. They discovered that they had to consider very carefully the aspects of student thinking that any given question might serve to explore. This discovery led them to work further on developing criteria for the quality of their questions. Thus the teachers confronted the importance of the cognitive foundations for designing assessment situations that can evoke important aspects of student thinking and learning. (See Bonniol [1991] and Perrenoud [1998]) for further discussion of the importance of high-quality teacher questions for illuminating student thinking.)

In response to research evidence that simply giving grades on written work can be counterproductive for learning (Butler, 1988), teachers began instead to concentrate on providing comments without grades—feedback designed to guide students’ further learning. Students also took part in self-assessment and peer-assessment activities, which required that they understand the goals for learning and the criteria for quality that applied to their work. These kinds of activities called for patient training and support from teachers, but fostered students’ abilities to focus on targets for learning and to identify learning goals for which they lacked confidence and needed help (metacognitive skills described in Chapter 3). In these ways, assessment situations became opportunities for learning, rather than activities divorced from learning.
There is a rich literature on how classroom assessment can be designed and used to improve instruction and learning (e.g., Falk, 2000; Niyogi, 1995; Shepard, 2000; Stiggins, 1997; Wiggins, 1998). This literature presents powerful ideas and practical advice to assist teachers across the K-16 spectrum in improving their classroom assessment practices. We do not attempt to summarize all of the insights and implications for practice presented in this literature. Rather, our emphasis is on what could be gained by thinking about classroom assessment in light of the principles of cognition and reasoning from evidence emphasized throughout this report.

Formative Assessment, Curriculum, and Instruction

At the 2000 annual meeting of the American Educational Research Association, Shepard (2000) began her presidential address by quoting Graue’s (1993, p. 291) observation, that “assessment and instruction are often conceived as curiously separate in both time and purpose.” Shepard asked:

How might the culture of classrooms be shifted so that students no longer feign competence or work to perform well on the test as an end separate from real learning? Could we create a learning culture where students and teachers would have a shared expectation that finding out what makes sense and what doesn’t is a joint and worthwhile project, essential to taking the next steps in learning? …How should what we do in classrooms be changed so that students and teachers look to assessment as a source of insight and help instead of its being the occasion for meting out reward and punishments. To accomplish this kind of transformation, we have to make assessment more useful, more helpful in learning, and at the same time change the social meaning of evaluation. (pp. 12-15)

Shepard proceeded to discuss ways in which classroom assessment practices need to change: the content and character of assessments need to be significantly improved to reflect contemporary understanding of learning; the gathering and use of assessment information and insights must become a part of the ongoing learning process; and assessment must become a central concern in methods courses in teacher preparation programs. Shepard’s messages were reflective of a growing belief among many educational assessment experts that if assessment, curriculum, and instruction were more integrally connected, student learning would improve (e.g., Gipps, 1999; Pellegrino, Baxter, and Glaser, 1999; Snow and Mandinach, 1991; Stiggins, 1997).

Sadler (1989) provides a conceptual framework that places classroom assessment in the context of curriculum and instruction. According to this framework, three elements are required for formative assessment to promote learning:
• A clear view of the learning goals.
• Information about the present state of the learner.
• Action to close the gap.

These three elements relate directly to assessment, curriculum, and instruction. The **learning goals** are derived from the curriculum. The **present state of the learner** is derived from assessment, so that the gap between it and the learning goals can be appraised. **Action** is then taken through instruction to close the gap. An important point is that assessment information by itself simply reveals student competence at a point in time; the process is considered formative assessment only when teachers use the information to make decisions about how to adapt instruction to meet students’ needs.

Furthermore, there are ongoing, dynamic relationships among formative assessment, curriculum, and instruction. That is, there are important bidirectional interactions among the three elements, such that each informs the other. For instance, formulating assessment procedures for classroom use can spur a teacher to think more specifically about learning goals, thus leading to modification of curriculum and instruction. These modifications can, in turn, lead to refined assessment procedures, and so on.

The mere existence of classroom assessment along the lines discussed here will not ensure effective learning. The clarity and appropriateness of the curriculum goals, the validity of the assessments in relationship to these goals, the interpretation of the assessment evidence, and the relevance and quality of the instruction that ensues are all critical determinants of the outcome. Starting with a model of cognition and learning in the domain can enhance each of these determinants.

**Importance of a Model of Cognition and Learning**

For most teachers, the ultimate goals for learning are established by the curriculum, which is usually mandated externally (e.g., by state curriculum standards). However, teachers and others responsible for designing curriculum, instruction, and assessment must fashion intermediate goals that can serve as an effective route to achieving the ultimate goals, and to do so they must have an understanding of how people represent knowledge and develop competence in the domain.

National and state standards documents set forth learning goals, but often not at a level of detail that is useful for operationalizing those goals in instruction and assessment (American Federation of Teachers, 1999; Finn, Petrilli, and Vanourek, 1998). By dividing goal descriptions into sets appropriate for different age and grade ranges, current curriculum standards provide broad guidance about the nature of the progression to be expected in various subject domains. Whereas this kind of epistemological and concep-
tual analysis of the subject domain is an essential basis for guiding assessment, deeper cognitive analysis of how people learn the subject matter is also needed. Formative assessment should be based on cognitive theories about how people learn particular subject matter to ensure that instruction centers on what is most important for the next stage of learning, given a learner’s current state of understanding. As described in Chapter 3, cognitive research has produced a rich set of descriptions of how people develop problem-solving and reasoning competencies in various content areas, particularly for the domains of mathematics and science. These models of learning provide a fertile ground for designing formative assessments.

It follows that teachers need training to develop their understanding of cognition and learning in the domains they teach. Preservice and professional development are needed to uncover teachers’ existing understandings of how students learn (Strauss, 1998), and to help them formulate models of learning so they can identify students’ naive or initial sense-making strategies and build on those strategies to move students toward more sophisticated understandings. The aim is to increase teachers’ diagnostic expertise so they can make informed decisions about next steps for student learning. This has been a primary goal of cognitively based approaches to instruction and assessment that have been shown to have a positive impact on student learning, including the Cognitively Guided Instruction program (Carpenter, Fennema, and Franke, 1996) and others (Cobb et al., 1991; Griffin and Case, 1997), some of which are described below. As these examples point out, however, such approaches rest on a bedrock of informed professional practice.

**Cognitively Based Approaches to Classroom Assessment: Examples**

**Cognitively Guided Instruction and Assessment**

Carpenter, Fennema, and colleagues have demonstrated that teachers who are informed regarding children’s thinking about arithmetic will be in a better position to craft more effective mathematics instruction (Carpenter et al., 1996; Carpenter, Fennema, Peterson, and Carey, 1988). Their approach, called Cognitively Guided Instruction (CGI), borrows much from cognitive science, yet recasts that work at a higher level of abstraction, a midlevel model designed explicitly to be easily understood and used by teachers. As noted earlier, such a model permits teachers to “read and react” to ongoing events in real time as they unfold during the course of instruction. In a sense, the researchers suggest that teachers use this midlevel model to support a process of continuous formative assessment so that instruction can be modified frequently as needed.
The cornerstone of CGI is a coarse-grained model of student thinking that borrows from work done in cognitive science to characterize the semantic structure of word problems, along with typical strategies children use for their solution. For instance, teachers are informed that problems apparently involving different operations, such as $3 + 7 = 10$ and $10 - 7 = 3$, are regarded by children as similar because both involve the action of combining sets. The model that summarizes children’s thinking about arithmetic word problems involving addition or subtraction is summarized by a three-dimensional matrix, in which the rows define major classes of semantic relations, such as combining, separating, or comparing sets; the columns refer to the unknown set (e.g., $7 + 3 = ?$ vs. $7 + ? = 10$); and the depth is a compilation of typical strategies children employ to solve problems such as these. Cognitive-developmental studies (Baroody, 1984; Carpenter and Moser, 1984; Siegler and Jenkins, 1989) suggest that children’s trajectories in this space are highly consistent. For example, direct modeling strategies are acquired before counting strategies; similarly, counting on from the first addend (e.g., $2 + 4 = ?$, 2, 3(1), 4(2), 5(3), 6(4)) is acquired before counting on from the larger addend (e.g., 4, 5(1), 6(2)).

Because development of these strategies tends to be robust, teachers can quickly locate student thinking within the problem space defined by CGI. Moreover, the model helps teachers locate likely antecedent understandings and helps them anticipate appropriate next steps. Given a student’s solution to a problem, a classroom teacher can modify instruction in a number of ways: (1) by posing a developmentally more difficult or easier problem; (2) by altering the size of the numbers in the set; or (3) by comparing and contrasting students’ solution strategies, so that students can come to appreciate the utility and elegance of a strategy they might not yet be able to generate on their own. For example, a student directly modeling a joining of sets with counters (e.g., $2 + 3$ solved by combining 2 chips with 3 chips and then counting all the chips) might profit by observing how a classmate uses a counting strategy (such as 2, 3(1), etc.) to solve the same problem. In a program such as CGI, formative assessment is woven seamlessly into the fabric of instruction (Carpenter et al., 1996).

**Intelligent Tutors**

As described in previous chapters, intelligent tutoring systems are powerful examples of the use of cognitively based classroom assessment tools blended with instruction. Studies indicate that when students work alone with these computer-based tutors, the relationship between formative assessment and the model of student thinking derived from research is comparatively direct. Students make mistakes, and the system offers effective
A large-scale experiment evaluated the benefits of intelligent tutoring in an urban high school (Koedinger, Anderson, Hadley, and Mark, 1997). Researchers compared achievement levels of ninth-grade students who received the PUMP curriculum, which is supported by an intelligent tutor, the PUMP Algebra Tutor (PAT) (experimental group), with those of students who received more traditional algebra instruction (control group).* The results, presented below, demonstrate strong learning benefits from using the curriculum that included the intelligent tutoring program.

The researchers did not collect baseline data to ensure similar starting achievement levels across experimental and control groups. However, they report that the groups were similar in terms of demographics. In addition, they looked at students’ mathematics grades in the previous school year to check for differences in students’ prior knowledge that would put the experimental group at an advantage. In fact, the average prior grades for the experimental group were lower than those for the control group.

*The researchers note that their research strategy is first to establish the success of the whole package and then to examine the effects of the curriculum and intelligent tutoring components independently; this work is still to be finished.

remediation. As a result, students on average learn more with the system than with other, traditional instruction (see Box 6-2).

On the other hand, some research suggests that the relationship between formative assessment and cognitive theory can be more complex. In a study of Anderson’s geometry tutor with high school students and their teachers, Schofield and colleagues found that teachers provided more articulate and better-tuned feedback than did the intelligent tutor (Schofield, Eurich-Fulcer, and Britt, 1994). Nevertheless, students preferred tutor-based to traditional instruction, not for the reasons one might expect, but because the tutor helped teachers tune their assistance to problems signaled by a student’s interaction with the tutor. Thus, student interactions with the tutor
(and sometimes their problems with it) served to elicit and inform more knowledgeable teacher assistance, an outcome that students apparently appreciated. Moreover, the assistance provided by teachers to students was less public. Hence, formative assessment and subsequent modification of instruction—both highly valued by these high school students—were mediated by a triadic relationship among teacher, student, and intelligent tutor. Interestingly, these interactions were not the ones originally intended by the designers of the tutor. Not surprisingly, rather than involving direct correspondence between model-based assessments and student learning, these relationships are more complex in actual practice. And the Schofield et al. study suggests that some portion of the effect may be due to stimulating positive teacher practices.

*SOURCE: Adapted from Koedinger, Anderson, Hadley, and Mark (1997). Used with permission of the American Association for the Advancement of Science.*
Reflections on the Teacher’s Role

Intelligent tutors and instructional programs such as Facets (described in Chapter 5) and CGI share an emphasis on providing clearer benchmarks of student thinking so that teachers can understand precursors and successors to the performances they are observing in real time. Thus these programs provide a “space” of student development in which teachers can work, a space that emphasizes ongoing formative assessment as an integral part of teaching practice. Yet these approaches remain under specified in important senses. Having good formative benchmarks in mind directs attention to important components and landmarks of thinking, yet teachers’ flexible and sensitive repertoires of assistance are still essential to achieving these goals. In general, these programs leave to teachers the task of generating and testing these repertoires. Thus, as noted earlier, the effectiveness of formative assessment rests on a bedrock of informed professional practice. Models of learning flesh out components and systems of reasoning, but they derive their purpose and character from the practices within which they are embedded. Similarly, descriptions of typical practices make little sense in the absence of careful consideration of the forms of knowledge representation and reasoning they entail (Cobb, 1998).

Complex cognitively based measurement models can be embedded in intelligent tutoring systems and diagnostic assessment programs and put to good use without the teacher’s having to participate in their construction. Many of the examples of assessments described in this report, such as Facets, intelligent tutoring systems, and BEAR (see Chapter 4), use statistical models and analysis techniques to handle some of the operational challenges. Providing teachers with carefully designed tools for classroom assessment can increase the utility of the information obtained. A goal for the future is to develop tools that make high-quality assessment more feasible for teachers. The topic of technology’s impact on the implementation of classroom assessment is one to which we return in Chapter 7.

The Quality of Feedback

As described in Chapter 3, learning is a process of continuously modifying knowledge and skills. Sometimes new inputs call for additions and extensions to existing knowledge structures; at other times they call for radical reconstruction. In all cases, feedback is essential to guide, test, challenge, or redirect the learner’s thinking.

Simply giving students frequent feedback in the classroom may or may not be helpful. For example, highly atomized drill-and-practice software can provide frequent feedback, but in so doing can foster rote learning and context dependency in students. A further concern is whether such software is being used appropriately given a student’s level of skill development. For
instance, a drill-and-practice program may be appropriate for developing fluency and automatizing a skill, but is usually not as appropriate during the early phase of skill acquisition (Goldman, Mertz, and Pellegrino, 1989). It is also noteworthy that in an environment where the teacher dominates all transactions, the frequent evocation and use of feedback can make that dominance all the more oppressive (Broadfoot, 1986).

There is ample evidence, however, that formative assessment can enhance learning when designed to provide students with feedback about particular qualities of their work and guidance on what they can do to improve. This conclusion is supported by several reviews of the research literature, including those by Natriello (1987), Crooks (1988), Fuchs and Fuchs (1986), Hattie (1987, 1990), and Black and Wiliam (1998). Many studies that have examined gains between pre- and post-tests, comparing programs in which formative assessment was the focus of the innovation and matched control groups were used, have shown effect sizes in the range of 0.4 to 0.7\(^1\) (Black and Wiliam, 1998).

When different types of feedback have been compared in experimental studies, certain types have proven to be more beneficial to learning than others. Many studies in this area have shown that learning is enhanced by feedback that focuses on the mastery of learning goals (e.g., Butler, 1988; Hattie, 1987, 1990; Kluger and DeNisi, 1996). This research suggests that other types of feedback, such as when a teacher focuses on giving grades, on granting or withholding special rewards, or on fostering self-esteem (trying to make the student feel better, irrespective of the quality of his or her work), may be ineffective or even harmful.

The culture of focusing on grades and rewards and of seeing classroom learning as a competition appears to be deeply entrenched and difficult to change. This situation is more apparent in the United States than in some other countries (Hattie, Biggs, and Purdie, 1996). The competitive culture of many classrooms and schools can be an obstacle to learning, especially when linked to beliefs in the fixed nature of ability (Vispoel and Austin, 1995; Wolf, Bixby, Glen, and Gardner, 1991). Such beliefs on the part of educators can lead both to the labeling—overtly or covertly—of students as “bright” or “dull” and to the confirmation and enhancement of such labels through tracking practices.

International comparative studies—notably case studies and video studies conducted for the Third International Mathematics and Science Study

\(^1\)To give a sense of the magnitude of such effect sizes, an effect size of 0.4 would mean that the average student who received the treatment would achieve at the same level as a student in the top 35 percent of those who did not receive the treatment. An effect size of 0.7, if realized in the Third International Mathematics and Science Study, would raise the United States from the middle of the 41 countries participating to one of the top 5.
that compare mathematics classrooms in Germany, Japan, and the United States—highlight the effects of these cultural beliefs. The studies underscore the difference between the culture of belief in Japan that the whole class can and should succeed through collaborative effort and the culture of belief endemic to many western countries, particularly the United States, that emphasizes the value of competition and differentiation (Cohen and Stevenson, 1995; Holloway, 1988).

The issues involved in students’ views of themselves as learners may be understood at a more profound level by regarding the classroom as a community of practice in which the relationships formed and roles adopted between teacher and students and among students help to form and interact with each member’s sense of personal identity (Cobb et al., 1991; Greeno and The Middle-School Mathematics Through Applications Group, 1997). Feedback can either promote or undermine the student’s sense of identity as a potentially effective learner. For example, a student might generate a conjecture that was later falsified. One possible form of feedback would emphasize that the conjecture was wrong. A teacher might, instead, emphasize the disciplinary value of formulating conjectures and the fruitful mathematics that often follows from generating evidence about a claim, even (and sometimes especially) a false one.

A voluminous research literature addresses characteristics of learners that relate to issues of feedback. Important topics of study have included students’ attributions for success and failure (e.g., Weiner, 1986), intrinsic versus extrinsic motivation (e.g., Deci and Ryan, 1985), and self-efficacy (e.g., Bandura and Schunk, 1981). We have not attempted to synthesize this large body of literature (for reviews see Graham and Weiner, 1996; Stipek, 1996). The important point to be made here is that teachers should be aware that different types of feedback have motivational implications that affect how students respond. Black and William (1998) sum up the evidence on feedback as follows:

. . . the way in which formative information is conveyed to a student, and the context of classroom culture and beliefs about ability and effort within which feedback is interpreted by the individual recipient, can affect these personal features for good or ill. The hopeful message is that innovations which have paid careful attention to these features have produced significant gains when compared with the existing norms of classroom practice. (p. 25)

The Role of the Learner

Students have a crucial role to play in making classroom assessment effective. It is their responsibility to use the assessment information to guide their progress toward learning goals. Consider the following assessment ex-
ample, which illustrates the benefits of having students engage actively in peer and self-assessment.

Researchers White and Frederiksen (2000) worked with teachers to develop the ThinkerTools Inquiry Project, a computer-enhanced middle school science curriculum that enables students to learn about the processes of scientific inquiry and modeling as they construct a theory of force and motion. The class functions as a research community, and students propose competing theories. They then test their theories by working in groups to design and carry out experiments using both computer models and real-world materials. Finally, students come together to compare their findings and to try to reach consensus about the physical laws and causal models that best account for their results. This process is repeated as the students tackle new research questions that foster the evolution of their theories of force and motion.

The ThinkerTools program focuses on facilitating the development of metacognitive skills as students learn the inquiry processes needed to create and revise their theories. The approach incorporates a reflective process in which students evaluate their own and each other’s research using a set of criteria that characterize good inquiry, such as reasoning carefully and collaborating well. Studies in urban classrooms revealed that when this reflective process is included, the approach is highly effective in enabling all students to improve their performance on various inquiry and physics measures and helps reduce the performance gap between low- and high-achieving students (see Box 6-3).

As demonstrated by the ThinkerTools example, peer and self-assessment are useful techniques for having learners share and grasp the criteria of quality work—a crucial step if formative assessment is to be effective. Just as teachers should adopt models of cognition and learning to guide instruction, they should also convey a model of learning (perhaps a simplified version) to their students so the students can monitor their own learning. This can be done through techniques such as the development of scoring rubrics or criteria for evaluating student work. As emphasized in Chapter 3, metacognitive awareness and control of one’s learning are crucial aspects of developing competence.

Students should be taught to ask questions about their own work and revise their learning as a result of reflection—in effect, to conduct their own formative assessment. When students who are motivated to improve have opportunities to assess their own and others’ learning, they become more capable of managing their own educational progress, and there is a transfer of power from teacher to learner. On the other hand, when formative feed-

Website: <garnet.berkeley.edu:7019/mchap.html>. [September 5, 2000].
White and Frederiksen (2000) carried out a controlled study comparing ThinkerTools classes in which students engaged in the reflective-assessment process with matched control classes in which they did not. Each teacher’s classes were evenly divided between the two treatments. In the reflective-assessment classes, the students continually engaged in monitoring and evaluating their own and each other’s research. In the control classes, the students were not given an explicit framework for reflecting on their research; instead, they engaged in alternative activities in which they commented on what they did and did not like about the curriculum. In all other respects, the classes participated in the same ThinkerTools inquiry-based science curriculum. There were no significant differences in students’ initial average standardized test scores (the Comprehensive Test of Basic Skills [CTBS] was used as a measure of prior achievement) between the classes assigned (randomly) to the different treatments.

One of the outcome measures was a written inquiry assessment that was given both before and after the ThinkerTools Inquiry Curriculum was administered. Presented below are the gain scores on this assessment for both low- and high-achieving students and for students in the reflective-assessment and control classes. Note first that students in the reflective-assessment classes gained more on this inquiry
assessment. Note also that this was particularly true for the low-achieving students. This is evidence that the metacognitive reflective-assessment process is beneficial, particularly for academically disadvantaged students.

This finding was further explored by examining the gain scores for each component of the inquiry test. As shown in the figure below, one can see that the effect of reflective assessment is greatest for the more difficult aspects of the test: making up results, analyzing those results, and relating them back to the original hypotheses. In fact, the largest difference in the gain scores is that for a measure termed “coherence,” which reflects the extent to which the experiments the students designed addressed their hypotheses, their made-up results related to their experiments, their conclusions followed from their results, and their conclusions were related back to their original hypotheses. The researchers note that this kind of overall coherence is a particularly important indication of sophistication in inquiry.

![Graph showing mean gain scores for hypothesis, experiment, results, conclusions, and coherence for assessment and control groups.](SOURCE: White and Frederiksen (2000, p. 347). Used with permission of the American Association for the Advancement of Science.)
back is “owned” entirely by the teacher, the power of the learner in the classroom is diminished, and the development of active and independent learning is inhibited (Deci and Ryan, 1994; Fernandes and Fontana, 1996; Grolnick and Ryan, 1987).

**Fairness**

Because the assessor, in this context typically the classroom teacher, has interactive contact with the learner, many of the construct-irrelevant barriers associated with external standardized assessments (e.g., language barriers, unfamiliar contexts) can potentially be detected and overcome in the context of classroom assessment. However, issues of fairness can still arise in classroom assessment. Sensitive attention by the teacher is paramount to avoid potential sources of bias. In particular, differences between the cultural backgrounds of the teacher and the students can lead to severe difficulties. For example, the kinds of questions a middle-class teacher asks may be quite unlike, in form and function, questions students from a different socio-economic or cultural group would experience at home, placing those students at a disadvantage (Heath, 1981, 1983).

Apart from the danger of a teacher’s personal bias, possibly unconscious, against any particular individual or group, there is also the danger of a teacher’s subscribing to the belief that learning ability or intelligence is fixed. Teachers holding such a belief may make self-confirming assumptions that certain children will never be able to learn, and may misinterpret or ignore assessment evidence to the contrary. However, as emphasized in the above discussion, there is great potential for formative assessment to assist and improve learning, and some studies, such as the ThinkerTools study described in Box 6-3, have shown that students initially classified as less able show the largest learning gains. There is some indication from other studies that the finding of greater gains for less able students may be generalizable, and this is certainly an area to be further explored. For now, these initial findings suggest that effective formative assessment practices may help overcome disadvantages endured at earlier stages in education.

Another possible source of bias may arise when students do not understand or accept learning goals. In such a case, responses that should provide the basis for formative assessment may not be meaningful or forthcoming.

---

3The literature reviews on mastery learning by Block and Burns (1976), Guskey and Gates (1986), and Kulik, Kulik, and Bangert-Drowns (1990) confirm evidence of extra learning gains for the less able, gains that have been associated with the feedback enhancement in such regimes. However, Livingston and Gentile (1996) have cast doubt on this attribution. Fuchs and Fuchs (1986) report that studies with children with learning handicaps showed mean gain effect sizes of 0.73, compared with a mean of 0.63 for nonhandicapped children.
This potential consequence argues for helping learners understand and share learning goals.

**LARGE-SCALE ASSESSMENT**

We have described ways in which classroom assessment can be used to improve instruction and learning. We now turn to a discussion of assessments that are used in large-scale contexts, primarily for policy purposes. They include state, national, and international assessments. At the policy level, large-scale assessments are often used to evaluate programs and/or to set expectations for individual student learning (e.g., for establishing the minimum requirements individual students must meet to move on to the next grade or graduate from high school). At the district level, such assessments may be used for those same purposes, as well as for matching students to appropriate instructional programs. At the classroom level, large-scale assessments tend to be less relevant but still provide information a teacher can use to evaluate his or her own instruction and to identify or confirm areas of instructional need for individual students. Though further removed from day-to-day instruction than classroom assessments, large-scale assessments have the potential to support instruction and learning if well designed and appropriately used. For parents, large-scale assessments can provide information about their own child’s achievement and some information about the effectiveness of the instruction their child is receiving.

**Implications of Advances in Cognition and Measurement**

Substantially more valid and useful information could be gained from large-scale assessments if the principles set forth in Chapter 5 were applied during the design process. However, fully capitalizing on the new foundations described in this report will require more substantial changes in the way large-scale assessment is approached, as well as relaxation of some of the constraints that currently drive large-scale assessment practices.

As described in Chapter 5, large-scale summative assessments should focus on the most critical and central aspects of learning in a domain as identified by curriculum standards and informed by cognitive research and theory. Large-scale assessments typically will reflect aspects of the model of learning at a less detailed level than classroom assessments, which can go into more depth because they focus on a smaller slice of curriculum and instruction. For instance, one might need to know for summative purposes whether a student has mastered the more complex aspects of multicolumn subtraction, including borrowing from and across zero, rather than exactly which subtraction bugs lead to mistakes. At the same time, while policy makers and parents may not need all the diagnostic detail that would be
useful to a teacher and student during the course of instruction, large-scale summative assessments should be based on a model of learning that is compatible with and derived from the same set of knowledge and beliefs about learning as classroom assessment.

Research on cognition and learning suggests a broad range of competencies that should be assessed when measuring student achievement, many of which are essentially untapped by current assessments. Examples are knowledge organization, problem representation, strategy use, metacognition, and kinds of participation in activity (e.g., formulating questions, constructing and evaluating arguments, contributing to group problem solving). Furthermore, large-scale assessments should provide information about the nature of student understanding, rather than simply ranking students according to general proficiency estimates.

A major problem is that only limited improvements in large-scale assessments are possible under current constraints and typical standardized testing scenarios. Returning to issues of constraints and trade-offs discussed earlier in this chapter, large-scale assessments are designed to serve certain purposes under constraints that often include providing reliable and comparable scores for individuals as well as groups; sampling a broad set of curriculum standards within a limited testing time per student; and offering cost-efficiency in terms of development, scoring, and administration. To meet these kinds of demands, designers typically create assessments that are given at a specified time, with all students taking the same (or parallel) tests under strictly standardized conditions (often referred to as “on-demand” assessment). Tasks are generally of the kind that can be presented in paper-and-pencil format, that students can respond to quickly, and that can be scored reliably and efficiently. In general, competencies that lend themselves to being assessed in these ways are tapped, while aspects of learning that cannot be observed under such constrained conditions are not addressed. To design new kinds of situations for capturing the complexity of cognition and learning will require examining the assumptions and values that currently drive assessment design choices and breaking out of the current paradigm to explore alternative approaches to large-scale assessment.

Alternative Approaches

To derive real benefits from the merger of cognitive and measurement theory in large-scale assessment requires finding ways to cover a broad range of competencies and to capture rich information about the nature of student understanding. This is true even if the information produced is at a coarse-grained as opposed to a highly detailed level. To address these challenges it is useful to think about the constraints and trade-offs associated
with issues of sampling—sampling of the content domain and of the student population.

The tasks on any particular assessment are supposed to be a representative sample of the knowledge and skills encompassed by the larger content domain. If the domain to be sampled is very broad, which is usually the case with large-scale assessments designed to cover a large period of instruction, representing the domain may require a large number and variety of assessment tasks. Most large-scale test developers opt for having many tasks that can be responded to quickly and that sample broadly. This approach limits the sorts of competencies that can be assessed, and such measures tend to cover only superficially the kinds of knowledge and skills students are supposed to be learning. Thus there is a need for testing situations that enable the collection of more extensive evidence of student performance.

If the primary purpose of the assessment is program evaluation, the constraint of having to produce reliable individual student scores can be relaxed, and population sampling can be useful. Instead of having all students take the same test (also referred to as “census testing”), a population sampling approach can be used whereby different students take different portions of a much larger assessment, and the results are combined to obtain an aggregate picture of student achievement.

If individual student scores are needed, broader sampling of the domain can be achieved by extracting evidence of student performance from classroom work produced during the course of instruction (often referred to as “curriculum-embedded” assessment). Student work or scores on classroom assessments can be used to supplement the information collected from an on-demand assessment to obtain a more comprehensive sampling of student performance. Although rarely used today for large-scale assessment purposes, curriculum-embedded tasks can serve policy and other external purposes of assessment if the tasks are centrally determined to some degree, with some flexibility built in for schools, teachers, and students to decide which tasks to use and when to have students respond to them.

Curriculum-embedded assessment approaches afford additional benefits. In on-demand testing situations, students are administered tasks that are targeted to their grade levels but not otherwise connected to their personal educational experiences. It is this relatively low degree of contextualization that renders these data good for some inferences, but not as good for others (Mislevy, 2000). If the purpose of assessment is to draw inferences about whether students can solve problems using knowledge and experiences they have learned in class, an on-demand testing situation in which every student receives a test with no consideration of his or her personal instruction history can be unfair. In this case, to provide valuable evidence of learning, the assessment must tap what the student has had the opportunity to learn (NRC, 1999b). In contrast to on-demand assessment, embedded
assessment approaches use techniques that link assessment tasks to concepts and materials of instruction. Curriculum-embedded assessment offers an alternative to on-demand testing for cases in which there is a need for correspondence among the curriculum, assessment, and actual instruction (see the related discussion of conditional versus unconditional inferences at the end of Chapter 5).

The following examples illustrate some cases in which these kinds of alternative approaches are being used successfully to evaluate individuals and programs in large-scale contexts. Except for DIAGNOSER, these examples are not strictly cognitively based and do not necessarily illustrate the features of design presented in Chapter 5. Instead they were selected to illustrate some alternative ways of approaching large-scale assessment and the trade-offs entailed. The first two examples show how population sampling has been used for program evaluation at the national and state levels to enable coverage of a broader range of learning goals than would be possible if each student were to take the same form of a test. The third and fourth examples involve approaches to measuring individual attainment that draw evidence of student performance from the course of instruction.

**Alternative Approaches to Large-Scale Assessment: Examples**

**National Assessment of Educational Progress**

As described earlier in this chapter, NAEP is a national survey intended to provide policymakers and the public with information about the academic achievement of students across the nation. It serves as one source of information for policymakers, school administrators, and the public for evaluating the quality of their curriculum and instructional programs. NAEP is a unique case of program evaluation in that it is not tied to any specific curriculum. It is based on a set of assessment frameworks that describe the knowledge and skills to be assessed in each subject area. The performances assessed are intended to represent the leading edge of what all students should be learning. Thus the frameworks are broader than any particular curriculum (NRC, 1999a). The challenge for NAEP is to assess the breadth of learning goals that are valued across the nation. The program approaches this challenge through the complex matrix sampling design described earlier.

NAEP’s design is beginning to be influenced by the call for more cognitively informed assessments of educational programs. Recent evaluations of NAEP (National Academy of Education, 1997; NRC, 1999a) emphasize that the current survey does not adequately capitalize on advances in our understanding of how people learn particular subject matter. These study
committees have strongly recommended that NAEP incorporate a broader conceptualization of school achievement to include aspects of learning that are not well specified in the existing NAEP frameworks or well measured by the current survey methods. The National Academy of Education panel recommended that particular attention be given to such aspects of student cognition as problem representation, the use of strategies and self-regulatory skills, and the formulation of explanations and interpretations, contending that consideration of these aspects of student achievement is necessary for NAEP to provide a complete and accurate assessment of achievement in a subject area. The subsequent review of NAEP by the NRC reiterated those recommendations and added that large-scale survey instruments alone cannot reflect the scope of these more comprehensive goals for schooling. The NRC proposed that, in addition to the current assessment blocks, which are limited to 50-minute sessions and paper-and-pencil responses, NAEP should include carefully designed, targeted assessments administered to smaller samples of students that could provide in-depth descriptive information about more complex activities that occur over longer periods of time. For instance, smaller data collections could involve observations of students solving problems in groups or performing extended science projects, as well as analysis of writing portfolios compiled by students over a year of instruction.

Thus NAEP illustrates how relaxing the constraint of having to provide individual student scores opens up possibilities for population sampling and coverage of a much broader domain of cognitive performances. The next example is another illustration of what can be gained by such a sampling approach.

**Maryland State Performance Assessment Program**

The Maryland State Performance Assessment Program (MSPAP) is designed to evaluate how well schools are teaching the basic and complex skills outlined in state standards called Maryland Learner Outcomes. Maryland is one of the few states in the country that has decided to optimize the use of assessment for program evaluation, forgoing individual student scores.4 A population sampling design is used, as opposed to the census testing design used by most states.

MSPAP consists of criterion-referenced performance tests in reading, mathematics, writing, language usage, science, and social studies for students in grades 3, 5, and 8. The assessment is designed to measure a broad range of competencies. Tasks require students to respond to questions or directions that lead to a solution for a problem, a recommendation or decision, or an explanation or rationale for their responses. Some tasks assess one content

4Website: <www.mdk12.org>. [June 29, 2000].
area; others assess multiple content areas. The tasks may encompass group or individual activities; hands-on, observation, or reading activities; and activities that require extended written responses, limited written responses, lists, charts, graphs, diagrams, webs, and/or drawings. A few MSPAP items are released each year to educators and the public to provide a picture of what the assessment looks like and how it is scored.5

To cover this broad range of learning outcomes, Maryland uses a sampling approach whereby each student takes only one-third of the entire assessment. This means an individual student’s results do not give a complete picture of how that child is performing (although parents can obtain a copy of their child’s results from the local school system). What is gained is a program evaluation instrument that covers a much more comprehensive range of learning goals than that addressed by a traditional standardized test.

**AP Studio Art**

The above two examples do not provide individual student scores. The AP Studio Art portfolio assessment is an example of an assessment that is designed to certify individual student attainment over a broad range of competencies and to be closely linked to the actual instruction students have experienced (College Board, 1994). Student work products are extracted during the course of instruction, collected, and then evaluated for summative evaluation of student attainment.

AP Studio Art is just one of many Advanced Placement (AP) programs designed to give highly motivated high school students the opportunity to take college-level courses in areas such as biology, history, calculus, and English while still in high school. AP programs provide course descriptions and teaching materials, but do not require that specific textbooks, teaching techniques, or curricula be followed. Each program culminates in an exam intended to certify whether individual students have mastered material equivalent to that of an introductory college course. AP Studio Art is unique in that at the end of the year, instead of taking a written summative exam, students present a portfolio of materials selected from the work they have produced during the AP course for evaluation by a group of artists and teachers. Preparation of the portfolio requires forethought; work submitted for the various sections must meet the publicly shared criteria set forth by the AP program.

The materials presented for evaluation may have been produced in art classes or on the student’s own time and may cover a period of time longer than a single school year. Instructional goals and the criteria by which students’ performance will be evaluated are made clear and explicit. Portfolio

---

5Website: <www.mdk12.org/mspp/mspap/look/prt_mspap.html>. [June 29, 2000].
requirements are carefully spelled out in a poster distributed to students and teachers; scoring rubrics are also widely distributed. Formative assessment is a critical part of the program as well. Students engage in evaluation of their own work and that of their peers, then use that feedback to inform next steps in building their portfolios. Thus while the AP Studio Art program is not directly based on cognitive research, it does reflect general cognitive principles, such as setting clear learning goals and providing students with opportunities for formative feedback, including evaluation of their own work.

Portfolios are scored quickly but fairly by trained raters. It is possible to assign reliable holistic scores to portfolios in a short amount of time. Numerous readings go into the scoring of each portfolio, enhancing the fairness of the assessment process (Mislevy, 1996). In this way, technically sound judgments are made, based on information collected through the learning process, that fulfill certification purposes. Thus by using a curriculum-embedded approach, the AP Studio Art program is able to collect rich and varied samples of student work that are tied to students’ instructional experiences over the course of the year, but can also be evaluated in a standardized way for the purposes of summative assessment.

It should be noted that some states attempting to implement large-scale portfolio assessment programs have encountered difficulties (Koretz and Barron, 1998). Therefore, while this is a good example of an alternative approach to on-demand testing, it should be recognized that there are many implementation challenges to be addressed.

**Facets DIAGNOSER**

We return to Minstrell and Hunt’s facets-based DIAGNOSER (Minstrell, 2000), described in some detail in Chapter 5, to illustrate another way of thinking about assessment of individuals’ summative achievement. The DIAGNOSER, developed for use at the classroom level to assist learning, does not fit the mold of traditional large-scale assessment. Various modules (each of which takes 15 to 20 minutes) cover small amounts of material fairly intensively. However, the DIAGNOSER could be used to certify individual attainment by noting the most advanced module a student had completed at a successful level of understanding in the course of instruction. For instance, the resulting assessment record would distinguish between students who had completed only Newtonian mechanics and those who had completed modules on the more advanced topics of waves or direct-circuit electricity. Because the assessment is part of instruction, there would be less concern about instructional time lost to testing.

Minstrell (2000) also speculates about how a facets approach could be applied to the development of external assessments designed to inform decisions at the program and policy levels. Expectations for learning, currently
conveyed by state and national curriculum standards, would be enhanced by facets-type research on learning. Current standards based on *what we want our students to know and be able to do* could be improved by incorporating findings from research on *what students know and are able to do along the way to competence*. By using a matrix sampling design, facet clusters could be covered extensively, providing summary information for decision makers about specific areas of difficulty for learners—information that would be useful for curriculum revision.

**Use of Large-Scale Assessment to Signal Worthy Goals**

Large-scale assessments can serve the purposes of learning by signaling worthwhile goals for educators and students to pursue. The challenge is to use the assessment program to signal goals at a level that is clear enough to provide some direction, but not so prescriptive that it results in a narrowing of instruction. Educators and researchers have debated the potential benefits of “teaching to a test.” Proponents of performance-based assessment have suggested that assessment can have a positive impact on learning if authentic tasks are used that replicate important performances in the discipline. The idea is that high-quality tasks can clarify and set standards of academic excellence, in which case teaching to the test becomes a good thing (Wiggins, 1989). Others (Miller and Seraphine, 1993) have argued that teaching to a test will always result in narrowing of the curriculum, given that any test can only sample the much broader domain of learning goals.

These views can perhaps be reconciled if the assessment is based on a well-developed model of learning that is shared with educators and learners. To make appropriate instructional decisions, teachers should teach to the model of learning—as conveyed, for example, by progress maps and rubrics for judging the quality of student work—rather than focusing on the particular items on a test. Test users must understand that any particular set of assessment tasks represents only a sample of the domain and that tasks will change from year to year. Given this understanding, assessment items and sample student responses can provide valuable exemplars to help teachers and students understand the underlying learning goals. Whereas teaching directly to the items on a test is not desirable, teaching to the set of beliefs about learning that underlie an assessment—which should be the same set of beliefs that underlies the curriculum—can provide positive direction for instruction.

High-quality summative assessment tasks are ones for which students can prepare only through active learning, as opposed to rote drill and practice or memorization of solutions. The United Kingdom’s Secondary School Certification Exam in physics (described in more detail later in this chapter) produces a wide variety of evidence that can be used to evaluate students’ summative achieve-
ment. The exam includes some transfer tasks that have been observed to be highly motivating for students (Morland, 1994). For instance, there is a task that assesses whether students can read articles dealing with applications of physics that lie outside the confines of the syllabus. Students know they will be presented with an article they have not seen before on a topic not specified in the syllabus, but that it will be at a level they should be able to understand on the basis of the core work of the syllabus. This task assesses students’ competency in applying their understanding in a new context in the process of learning new material. The only way for students to prepare for this activity is to read a large variety of articles and work systematically to understand them.

Another goal of the U.K. physics curriculum is to develop students’ capacity to carry out experimental investigations on novel problems. Students are presented with a scientific problem that is not included in the routine curriculum materials and must design an experiment, select and appropriately use equipment and procedures to implement the design, collect and analyze data, and interpret the data. Again, the only way students can prepare for this task is by engaging in a variety of such investigations and learning how to take responsibility for their design, implementation, and interpretation. In the United Kingdom, these portions of the physics exam are administered by the student’s own teacher, with national, standardized procedures in place for ensuring and checking fairness and rigor. When this examination was first introduced in the early 1970s, it was uncommon in classrooms to have students read on topics outside the syllabus and design and conduct their own investigations. The physics exam has supported the message, also conveyed by the curriculum, that these activities are essential, and as a result students taking the exam have had the opportunity to engage in such activities in the course of their study (Tebbutt, 1981).

Feedback and Expectations for Learning

In Chapters 4 and 5, we illustrated some of the kinds of information that could be obtained by reporting large-scale assessment results in relation to developmental progress maps or other types of learning models. Assessment results should describe student performance in terms of different states and levels of competence in the domain. Typical learning pathways should be displayed and made as recognizable as possible to educators, students, and the public.

Large-scale assessments of individual achievement could be improved by focusing on the potential for providing feedback that not only measures but also enhances future learning. Assessments can be designed to say both that this person is unqualified to move on and that this person’s difficulty lies
in these particular areas, and that is what has to be improved, the other components being at the desired level.

Likewise, assessments designed to evaluate programs should provide the kinds of information decision makers can use to improve those programs. People tend to think of school administrators and policy makers as removed from concerns about the details of instruction. Thus large-scale assessment information aimed at those users tends to be general and comparative, rather than descriptive of the nature of learning that is taking place in their schools. Practices in some school districts, however, are challenging these assumptions (Resnick and Harwell, 1998).

Telling an administrator that mathematics is a problem is too vague. Knowing how a school is performing in mathematics relative to past years, how it is performing relative to other schools, and what proportions of students fall in various broadly defined achievement categories also provides little guidance for program improvement. Saying that students do not understand probability is more useful, particularly to a curriculum planner. And knowing that students tend to confuse conditional and compound probability can be even more useful for the modification of curriculum and instruction. Of course, the sort of feedback needed to improve instruction depends on the program administrator’s level of control.

Not only do large-scale assessments provide means for reporting on student achievement, but they also convey powerful messages about the kinds of learning valued by society. Large-scale assessments should be used by policy makers and educators to operationalize and communicate among themselves, and to the public, the kinds of thinking and learning society wishes to encourage in students. In this way, assessments can foster valuable dialogue about learning and its assessment within and beyond the education system. Models of learning should be shared and communicated in accessible ways to show what competency in a domain looks like. For example, Developmental Assessment based on progress maps is being used in the Commonwealth of Victoria to assess literacy. An evaluation of the program revealed that users were “overwhelmingly positive about the value and potential of Developmental Assessment as a means for developing shared understandings and a common language for literacy development” (Meiers and Culican, 2000, p. 44).

Example: The New Standards Project

The New Standards Project, as originally conceived (New Standards™, 1997a, 1997b, 1997c), illustrates ways to approach many of the issues of large-scale assessment discussed above. The program was designed to provide clear goals for learning and assessments that are closely tied to those
goals. A combination of on-demand and embedded assessment was to be used to tap a broad range of learning outcomes, and priority was given to communicating the performance standards to various user communities. Development of the program was a collaboration between the Learning Research and Development Center of the University of Pittsburgh and the National Center on Education and the Economy, in partnership with states and urban school districts. Together they developed challenging standards for student performance at grades 4, 8, and 10, along with large-scale assessments designed to measure attainment of those standards.6

The New Standards Project includes three interrelated components: performance standards, a portfolio assessment system,7 and an on-demand exam. The performance standards describe what students should know and the ways they should demonstrate the knowledge and skills they have acquired. The performance standards include samples of student work that illustrate high-quality performances, accompanied by commentary that shows how the work sample reflects the performance standards. They go beyond most content standards by describing how good is good enough, thus providing clear targets to pursue.

The Reference Exam is a summative assessment of the national standards in the areas of English Language Arts and Mathematics at grades 4, 8, and 10. The developers state explicitly that the Reference Exam is intended to address those aspects of the performance standards that can be assessed in a limited time frame under standardized conditions. The portfolio assessment system was designed to complement the Reference Exam by providing evidence of achievement of those performance standards that depend on extended work and the accumulation of evidence over time.

The developers recognized the importance of making the standards clear and presenting them in differing formats for different audiences. One version of the standards is targeted to teachers. It includes relatively detailed language about the subject matter of the standards and terms educators use to describe differences in the quality of work produced by students. The standards are also included in the portfolio material provided for student use. In these materials, the standards are set forth in the form of guidelines to help students select work for inclusion in their portfolios. In addition, there were plans to produce a less technical version for parents and the community in general.

6Aspects of the program have since changed, and the Reference Exam is now administered by Harcourt Educational Measurement.

7The portfolio component was field tested but has not been administered on a large scale.


ASSESSMENT SYSTEMS

In the preceding discussion we have addressed issues of practice related to classroom and large-scale assessment separately. We now return to the matter of how such assessments can work together conceptually and operationally.

As argued throughout this chapter, one form of assessment does not serve all purposes. Given that reality, it is inevitable that multiple assessments (or assessments consisting of multiple components) are required to serve the varying educational assessment needs of different audiences. A multitude of different assessments are already being conducted in schools. It is not surprising that users are often frustrated when such assessments have conflicting achievement goals and results. Sometimes such discrepancies can be meaningful and useful, such as when assessments are explicitly aimed at measuring different school outcomes. More often, however, conflicting assessment goals and feedback cause much confusion for educators, students, and parents. In this section we describe a vision for coordinated systems of multiple assessments that work together, along with curriculum and instruction, to promote learning. Before describing specific properties of such systems, we consider issues of balance and allocation of resources across classroom and large-scale assessment.

Balance Between Classroom and Large-Scale Assessment

The current educational assessment environment in the United States clearly reflects the considerable value and credibility accorded external, large-scale assessments of individuals and programs relative to classroom assessments designed to assist learning. The resources invested in producing and using large-scale testing in terms of money, instructional time, research, and development far outweigh the investment in the design and use of effective classroom assessments. It is the committee’s position that to better serve the goals of learning, the research, development, and training investment must be shifted toward the classroom, where teaching and learning occurs.

Not only does large-scale assessment dominate over classroom assessment, but there is also ample evidence of accountability measures negatively impacting classroom instruction and assessment. For instance, as discussed earlier, teachers feel pressure to teach to the test, which results in a narrowing of instruction. They also model their own classroom tests after less-than-ideal standardized tests (Gifford and O’Connor, 1992; Linn, 2000; Shepard, 2000). These kinds of problems suggest that beyond striking a better balance between classroom and large-scale assessment, what is needed are coordinated assessment systems that collectively support a common set of learning goals, rather than working at cross-purposes.
Ideally in a balanced assessment environment, a single assessment does not function in isolation, but rather within a nested assessment system involving states, local school districts, schools, and classrooms. Assessment systems should be designed to optimize the credibility and utility of the resulting information for both educational decision making and general monitoring. To this end, an assessment system should exhibit three properties: comprehensiveness, coherence, and continuity. These three characteristics describe an assessment system that is aligned along three dimensions: vertically, across levels of the education system; horizontally, across assessment, curriculum, and instruction; and temporally, across the course of a student’s studies. These notions of alignment are consistent with those set forth by the National Institute for Science Education (Webb, 1997) and the National Council of Teachers of Mathematics (1995).

Features of a Balanced Assessment System

Comprehensiveness

By comprehensiveness, we mean that a range of measurement approaches should be used to provide a variety of evidence to support educational decision making. Educational decisions often require more information than a single measure can provide. As emphasized in the NRC report *High Stakes: Testing for Tracking, Promotion, and Graduation*, multiple measures take on particular importance when important, life-altering decisions (such as high school graduation) are being made about individuals. No single test score can be considered a definitive measure of a student’s competence. Multiple measures enhance the validity and fairness of the inferences drawn by giving students various ways and opportunities to demonstrate their competence. The measures could also address the quality of instruction, providing evidence that improvements in tested achievement represent real gains in learning (NRC, 1999c).

One form of comprehensive assessment system is illustrated in Table 6-1, which shows the components of a U.K. examination for certification of top secondary school students who have studied physics as one of three chosen subjects for 2 years between ages 16 and 18. The results of such examinations are the main criterion for entrance to university courses. Components A, B, C, and D are all taken within a few days, but E and F involve activities that extend over several weeks preceding the formal examination.

This system combines external testing on paper (components A, B, and C) with external performance tasks done using equipment (D) and teachers’ assessment of work done during the course of instruction (E and F). While
TABLE 6-1 Six Components of an A-Level Physics Examination

<table>
<thead>
<tr>
<th>Component</th>
<th>Title</th>
<th>No. of Questions or Tasks</th>
<th>Time</th>
<th>Weight in Marks</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Coded Answer</td>
<td>40</td>
<td>75 min.</td>
<td>20%</td>
<td>Multiple choice questions, all to be attempted.</td>
</tr>
<tr>
<td>B</td>
<td>Short Answer</td>
<td>7 or 8</td>
<td>90 min.</td>
<td>20%</td>
<td>Short with structured subcomponents, fixed space for answer, all to be attempted.</td>
</tr>
<tr>
<td>C</td>
<td>Comprehension</td>
<td>3</td>
<td>150 min.</td>
<td>24%</td>
<td>a) Answer questions on a new passage.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>b) Analyze and draw conclusions from a set of presented data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>c) Explain phenomena described in short paragraphs: select 3 from 5.</td>
</tr>
<tr>
<td>D</td>
<td>Practical Problems</td>
<td>8</td>
<td>90 min.</td>
<td>16%</td>
<td>Short problems with equipment set up in a laboratory, all to be attempted.</td>
</tr>
<tr>
<td>E</td>
<td>Investigation</td>
<td>1</td>
<td>About 2 weeks</td>
<td>10%</td>
<td>In normal school laboratory time, investigate a problem of the student's own choice.</td>
</tr>
<tr>
<td>F</td>
<td>Project Essay</td>
<td>1</td>
<td>About 2 weeks</td>
<td>10%</td>
<td>In normal school time, research and write about a topic chosen by the student.</td>
</tr>
</tbody>
</table>

SOURCE: Adapted from Morland (1994).

This particular physics examination is now subject to change,\(^8\) combining the results of external tests with classroom assessments of particular aspects of achievement for which a short formal test is not appropriate is an established feature of achievement testing systems in the United Kingdom and

---

\(^8\)Because the whole structure of the 16-18 examinations is being changed, this examination and the curriculum on which it is based, which have been in place for 30 years, will no longer be in use after 2001. They will be replaced by a new curriculum and examination, based on the same principles.
several other countries. This feature is also part of the examination system for the International Baccalaureate degree program. In such systems, work is needed to develop procedures for ensuring the comparability of standards across all teachers and schools.

Overall, the purpose is to reflect the variety of the aims of a course, including the range of knowledge and simple understanding explored in A, the practical skills explored in D, and the broader capacities for individual investigation explored in E and F. Validity and comprehensiveness are enhanced, albeit through an expensive and complex assessment process.

There are other possible ways to design comprehensive assessment systems. Portfolios are intended to record “authentic” assessments over a period of time and a range of classroom contexts. A system may assess and give certification in stages, so that the final outcome is an accumulation of results achieved and credited separately over, say, 1 or 2 years of a learning course; results of this type may be built up by combining on-demand externally controlled assessments with work samples drawn from coursework. Such a system may include assessments administered at fixed times or at times of the candidate’s choice using banks of tasks from which tests can be selected to match the candidate’s particular opportunities to learn. Thus designers must always look to the possibility of using the broader approaches discussed here, combining types of tasks and the timing of assessments and of certifications in the optimum way.

Further, in a comprehensive assessment system, the information derived should be technically sound and timely for given decisions. One must be able to trust the accuracy of the information and be assured that the inferences drawn from the results can be substantiated by evidence of various types. The technical quality of assessment is a concern primarily for external, large-scale testing; but if classroom assessment information is to feed into the larger assessment system, the reliability, validity, and fairness of these assessments must be addressed as well. Researchers are just beginning to explore issues of technical quality in the realm of classroom assessment (e.g., Wilson and Sloane, 2000).

Coherence

For the system to support learning, it must also have a quality the committee refers to as coherence. One dimension of coherence is that the conceptual base or models of student learning underlying the various external and classroom assessments within a system should be compatible. While a large-scale assessment might be based on a model of learning that is coarser than that underlying the assessments used in classrooms, the conceptual base for the large-scale assessment should be a broader version of one that makes sense at the finer-grained level (Mislevy, 1996). In this way, the exter-
nal assessment results will be consistent with the more detailed understanding of learning underlying classroom instruction and assessment. As one moves up and down the levels of the system, from the classroom through the school, district, and state, assessments along this vertical dimension should align. As long as the underlying models of learning are consistent, the assessments will complement each other rather than present conflicting goals for learning.

To keep learning at the center of the educational enterprise, assessment information must be strongly linked to curriculum and instruction. Thus another aspect of coherence, emphasized earlier, is that alignment is needed among curriculum, instruction, and assessment so that all three parts of the education system are working toward a common set of learning goals. Ideally, assessment will not simply be aligned with instruction, but integrated seamlessly into instruction so that teachers and students are receiving frequent but unobtrusive feedback about their progress. If assessment, curriculum, and instruction are aligned with common models of learning, it follows that they will be aligned with each other. This can be thought of as alignment along the horizontal dimension of the system.

To achieve both the vertical and horizontal dimensions of coherence or alignment, models of learning are needed that are shared by educators at different levels of the system, from teachers to policy makers. This need might be met through a process that involves gathering together the necessary expertise, not unlike the approach used to develop state and national curriculum standards that define the content to be learned. But current definitions of content must be significantly enhanced based on research from the cognitive sciences. Needed are user-friendly descriptions of how students learn the content, identifying important targets for instruction and assessment (see, e.g., American Association for the Advancement of Science, 2001). Research centers could be charged with convening the appropriate experts to produce a synthesis of the best available scientific understanding of how students learn in particular domains of the curriculum. These models of learning would then guide assessment design at all levels, as well as curriculum and instruction, effecting alignment in the system. Some might argue that what we have described are the goals of current curriculum standards. But while the existing standards emphasize what students should learn, they do not describe how students learn in ways that are maximally useful for guiding instruction and assessment.

**Continuity**

In addition to comprehensiveness and coherence, an ideal assessment system would be designed to be continuous. That is, assessments should measure student progress over time, akin more to a videotape record than to
the snapshots provided by the current system of on-demand tests. To provide such pictures of progress, multiple sets of observations over time must be linked conceptually so that change can be observed and interpreted. Models of student progression in learning should underlie the assessment system, and tests should be designed to provide information that maps back to the progression. With such a system, we would move from “one-shot” testing situations and cross-sectional approaches for defining student performance toward an approach that focused on the processes of learning and an individual’s progress through that process (Wilson and Sloane, 2000). Thus, continuity calls for alignment along the third dimension of time.

**Approximations of a Balanced System**

No existing assessment systems meet all three criteria of comprehensiveness, coherence, and continuity, but many of the examples described in this report represent steps toward these goals. For instance, the Developmental Assessment program shows how progress maps can be used to achieve coherence between formative and summative assessments, as well as among curriculum, instruction, and assessment. Progress maps also enable the measurement of growth (continuity). The Australian Council for Educational Research has produced an excellent set of resource materials for teachers to support their use of a wide range of assessment strategies—from written tests to portfolios to projects at the classroom level—that can all be designed to link back to the progress maps (comprehensiveness) (see, e.g., Forster and Masters, 1996a, 1996b; Masters and Forster, 1996). The BEAR assessment shares many similar features; however, the underlying models of learning are not as strongly tied to cognitive research as they could be. On the other hand, intelligent tutoring systems have a strong cognitive research base and offer opportunities for integrating formative and summative assessments, as well as measuring growth, yet their use for large-scale assessment purposes has not yet been explored. Thus, examples in this report offer a rich set of opportunities for further development toward the goal of designing assessment systems that are maximally useful for both informing and improving learning.

**CONCLUSIONS**

Guiding the committee’s work were the premises that (1) something important should be learned from every assessment situation, and (2) the information gained should ultimately help improve learning. The power of classroom assessment resides in its close connections to instruction and teachers’ knowledge of their students’ instructional histories. Large-scale, standardized assessments can communicate across time and place, but by so
constraining the content and timeliness of the message that they often have limited utility in the classroom. Thus the contrast between classroom and large-scale assessments arises from the different purposes they serve and contexts in which they are used. Certain trade-offs are an inescapable aspect of assessment design.

Students will learn more if instruction and assessment are integrally related. In the classroom, providing students with information about particular qualities of their work and about what they can do to improve is crucial for maximizing learning. It is in the context of classroom assessment that theories of cognition and learning can be particularly helpful by providing a picture of intermediary states of student understanding on the pathway from novice to competent performer in a subject domain.

Findings from cognitive research cannot always be translated directly or easily into classroom practice. Most effective are programs that interpret the findings from cognitive research in ways that are useful for teachers. Teachers need theoretical training, as well as practical training and assessment tools, to be able to implement formative assessment effectively in their classrooms.

Large-scale assessments are further removed from instruction, but can still benefit learning if well designed and properly used. Substantially more valid and useful inferences could be drawn from such assessments if the principles set forth in this report were applied during the design process.

Large-scale assessments not only serve as a means for reporting on student achievement, but also reflect aspects of academic competence societies consider worthy of recognition and reward. Thus large-scale assessments can provide worthwhile targets for educators and students to pursue. Whereas teaching directly to the items on a test is not desirable, teaching to the theory of cognition and learning that underlies an assessment can provide positive direction for instruction.

To derive real benefits from the merger of cognitive and measurement theory in large-scale assessment, it will be necessary to devise ways of covering a broad range of competencies and capturing rich information about the nature of student understanding. Indeed, to fully capitalize on the new foundations described in this report will require substantial changes in the way large-scale assessment is approached and relaxation of some of the constraints that currently drive large-scale assessment practices. Alternatives to on-demand, census testing are available. If individual student scores are needed, broader sampling of the domain can be achieved by extracting evidence of student performance from classroom work produced during the course of instruction. If the primary purpose of the assessment is program evaluation, the constraint of having to produce reliable individual student scores can be relaxed, and population sampling can be useful.
For classroom or large-scale assessment to be effective, students must understand and share the goals for learning. Students learn more when they understand (and even participate in developing) the criteria by which their work will be evaluated, and when they engage in peer and self-assessment during which they apply those criteria. These practices develop students’ metacognitive abilities, which, as emphasized above, are necessary for effective learning.

The current educational assessment environment in the United States assigns much greater value and credibility to external, large-scale assessments of individuals and programs than to classroom assessment designed to assist learning. The investment of money, instructional time, research, and development for large-scale testing far outweighs that for effective classroom assessment. More of the research, development, and training investment must be shifted toward the classroom, where teaching and learning occur.

A vision for the future is that assessments at all levels—from classroom to state—will work together in a system that is comprehensive, coherent, and continuous. In such a system, assessments would provide a variety of evidence to support educational decision making. Assessment at all levels would be linked back to the same underlying model of student learning and would provide indications of student growth over time.