6

Tensions, Trends, Tools, and Technologies:

Time for an Educational Sea Change

Valerie J. Shute
Educational Testing Service

> Education is not preparation for life; education is life itself.
>
> John Dewey *(1859–1952)*

Introduction

This chapter outlines three educational approaches: (a) *traditional*: the currently dominant approach, a largely lecture-oriented authoritarian style that makes heavy use of "assessments *of* learning," which are useful for accountability purposes but only marginally useful for guiding day-to-day instruction; (b) *progressive*: a highly student-centered approach that relies on "assessments *for* learning," which can be very useful in guiding day-to-day instruction; and (c) *unified*: a new, integrated approach that uses the best of both kinds of assessments—"for" and "of" learning—and that leverages computer technology, educational measurement, and cognitive science to address factors that undermined earlier attempts to implement the Progressive approach. This chapter examines some of the research, trends, and factors that should be considered, understood, and, in some cases, leveraged in order to move toward the unified approach. Further, this chapter presents examples of how Educational Testing

Service (ETS) projects are moving toward the new approach to harness assessment in the service of learning.

Déjà Vu

Think back on your high school years. Whether Elvis, the Beatles, Led Zeppelin, Madonna, Run-DMC, Pearl Jam, or Britney Spears dominated the charts, odds are that you spent your day going from one 50-minute class to another, with a different subject each period. In class, you probably spent most of your time sitting at your desk, listening to lectures from a teacher who was the repository of knowledge to be learned. Your job was to learn the facts and other knowledge that your teacher knew, and you were periodically tested on just how well you absorbed the information and could retrieve the relevant facts. Direct cooperation with other students was a relatively rare event (except perhaps in team sports). This traditional scenario captures the norm for U.S. schools that have underserved too many students for too long (e.g., Barton, 2005).

Now imagine the following: public schools that apply progressive methods—such as individualizing instruction, motivating students by considering their interests, and developing cooperative group projects—to achieve the goal of producing knowledgeable and skilled lifelong learners. The teachers are happy, hard working, and valued by the community. In addition, they hold leadership roles in the school, and work individually and collectively to figure out the best ways to reach and teach their students. These same teachers create new textbooks and conduct research to see whether their methods worked. School days are structured to allow teachers time to meet and discuss their findings with colleagues.

Is this an ideal vision of schools of the future? Yes and no. According to Ravitch (2000), the image above describes several model public schools in the U.S. in the 1920s and 1930s, inspired by John Dewey's vision of education (e.g., the Lincoln School at Teachers College in New York, and the Winnetka, Illinois, public schools). These schools were engaging places for children to learn and were attractive places for teachers to teach; they avoided the monotonous, unfruitful routines of traditional schools.

What happened to these exciting experiments of educational reform, and more importantly, what lessons can we learn from them? First, according to Kliebard (1987), they failed because the techniques and founding ideas were misapplied by so-called experts who believed that mass education could be accomplished cheaply, employing low-paid and poorly trained teachers who would either follow their manuals or stand aside while students pursued their interests. Second, they failed because the reforms rejected traditional subject-matter curricula and substituted vocational training for the 90% of the student population who, at the time, were not expected to seek or hold professional careers (see Bobbitt, 1912, "The Elimination of Waste in Education"). Finally, this period also saw mass IQ testing (e.g., Lemann, 1999) gaining a firm foothold in education, with systematic use of Terman's National Intelligence Test in senior and junior high schools. The testing was aimed specifically at efficiently assigning students into high, middle, or low educational tracks according to their supposedly innate mental abilities (Terman, 1916).

In general, there was a fundamental shift to practical education going on in the country during the early 1900s, countering "wasted time" in schools and abandoning the classics as useless and inefficient for the masses. Bobbitt, along with some other early educational researchers and administrators such as Ellwood and Ayers (Kliebard, 1987, pp. 103–104), inserted into the national educational discourse the metaphor of the school as a "factory." This metaphor has persisted to this day; yet if schools were actual factories, they would have been shut down years ago.

The basic idea I present in this chapter is that serious problems exist in education today, but viable solutions are possible. The particular solution described herein is based on the claims that (1) individual differences among students have powerful effects on learning, (2) these effects can be quantified and predicted, and (3) technology can capitalize on these effects to the benefit of teachers and students (as well as others, such as administrators and parents).

This chapter is organized as follows. First, I describe two distinct educational approaches—traditional and progressive—that have been battling it out in our country for almost a century, although both have valuable contributions to make to education. Second, I summarize factors that are influencing the current state of educational flux, fueling the need for an educational *sea change*.[1] Third, research is presented that seems promising for addressing the particular problem areas that are delineated. I also present specific models and methods that we can use right now to create diagnostic, formative assessments[2] that are woven directly into the fabric of the curriculum, linked to targeted instruction as well as standards, and are likely to make a real difference in the landscape (or seascape) of education. Finally, I sketch out a prototype system currently under development at Educational Testing Service that employs many of the methods and tools cited in the chapter.

The Chasm Between Traditional and Progressive Approaches

The model of school-as-a-factory is inappropriate, particularly in today's rapidly changing and information-rich world. So what is a better model (or models) that we can use to focus educational reform? Very simply: *There are two competing views of education—traditional and progressive—from which we can draw the best features to combine into a new, unified model.* On the one hand, traditional education invokes a more "outside-in" approach whereby teachers provide knowledge to awaiting students. On the other hand, progressive education is more "inside-out," defining the role of the student as an active, creative, and reflective participant in the learning process.

John Dewey believed that the more authoritarian approach of traditional education was too concerned with delivering knowledge, and not enough with understanding students' actual learning or experiences, the cornerstone of progressive education (see Flanagan, 1994). However, he was also highly critical of completely "free, student-driven" education because students often do not know how to structure their own learning experiences for maximum benefit. Fast forward 70 to 80 years and we see the paradigm conflict continuing today.[3]

Traditional Approach

Many educators, administrators, and policy makers support relatively structured, didactic, traditional education. This approach came to the fore with the recession and tax revolt of the 1970s, followed by the publication of the report, *A Nation at Risk* (National Commission of Excellence in Education, 1983), leading to an increased emphasis on basics, national learning standards, and improving results on standardized tests.

Lending credible support for this position, consider the findings from a project called "Follow Through" (e.g., Proper & St. Pierre, 1980; Stebbins, St. Pierre, Proper, Anderson, & Cerva, 1977). This was an enormous, federally funded research project launched in 1967 in response to President Johnson's request to "follow through" on project Head Start. Summaries of the study (e.g., Adams, 1996; Stone & Clements, 1998) describe nine educational Models[4] that were compared in 51 school districts over a 4 to 6 year period. Each of the nine models was yoked to a comparison school. Of the nine, all but two (i.e., Direct Instruction and Behavior Analysis Models; see complete listing in footnote) were, to various degrees, learner centered. Contrary to expectations, the two exceptions significantly outperformed the other Models. Furthermore, Stone and Clements (1998) noted that five of the seven learner-centered models produced worse results than the traditional school programs (i.e., the control groups) to which each Follow Through approach was compared. By far, the most successful of the nine Models was Direct Instruction[5] (Engelmann, Becker, Carnine, & Gersten, 1988), which showed positive scores on all three types of outcome measures—basic skills, cognitive skills, and affective variables (Adams, 1996).

At least three other major re-analyses of the data were independently conducted (see Mac Iver & Kemper, 2002), yet none of these analyses show significant disagreement with respect to achievement data. Results of the national evaluation and all subsequent analyses converge on the finding that the highest achievement scores were attained by students in the Direct Instruction model.

Progressive Approach

There are equally ardent supporters of progressive education,[6] which generally refers to classroom methods that focus on individualized instruction, encourage collaboration among students, provide hands-on learning activities, and stress informality in the classroom (e.g., Brown & Campione, 1990; Darling-Hammond, 1997; Darling-Hammond, Griffin, & Wise, 1992; Pea, 1994; Scardamalia & Bereiter, 1994). Researchers report that intrinsic motivation is enhanced when learning is student-centered, i.e., when students are provided with opportunities to exert control, to determine their fate, or at least have a perception that they are doing so (e.g., Lepper & Chabay, 1985; Ng, Guthrie, Van Meter, McCann, & Alao, 1998). For example, Deci and colleagues (e.g., Deci & Ryan, 1985; Deci, Vallerand, Pelletier, & Ryan, 1991) found that when students have control over their own learning, they achieve more positive learning outcomes, greater interest, more trust, higher self-esteem, and greater persistence. Additional research has reported the increased benefits to students in relation to self-determination (Papert, 1980) and feelings of control (Keller & Kopp, 1987).

Examples from research employing interactive instructional materials report positive outcomes relating student control to improved learning (e.g., Carrier & Williams, 1988). And motivational theory research (Keller, 1979) has similarly demonstrated that when students are given some control over aspects of their learning, they are more likely to have positive feelings toward the task combined with intrinsic motivation. Finally, Laurillard (1984, 1991) reported findings that learning enjoyment increased when students were given control.

Toward a New Approach: A Look at the Interactions

This dichotomy between the two opposing educational philosophies (i.e., traditional and progressive) may also be seen in the implementation of computerized learning environments. Among other variables, such systems can differ in the amount of learner control (one of the main features of the progressive approach) supported during the learning process. The research literature is about evenly mixed in relation to the effectiveness of these two approaches—traditional and progressive (specifically, in this case, less and more learner control)—and the arguments are similar to those described earlier with regard to classroom settings. That is, one approach argues that it is more efficacious to develop straightforward learning environments that do not permit "garden path" digressions (e.g., Koedinger, Anderson, Hadley, & Mark, 1997; Sleeman, Kelly, Martinak, Ward, & Moore, 1989). In contrast, the other approach argues that student learning is enhanced by environments containing assorted tools that allow the learner freedom to explore and learn, unfettered (e.g., Bunt, Conati, Huggett, & Muldner, 2001; de Jong, van Joolingen, Scott, deHoog, Lapied, & Valent, 1994; Shute, Glaser, & Raghavan, 1989).

The disparity between positions becomes more complex because the issue is not just about which approach—traditional or progressive—is the better learning environment; i.e., it is unrealistic to suppose that a statistical "main effect" for "approach" would provide an adequate picture. Instead, a better question may be the following: Which is the better approach for what type(s) of students? In other words, we should examine the data for evidence of classic aptitude-treatment interactions (Cronbach & Snow, 1977), for which the main effect would be an inadequate summary. This may be further extended to include other variables as well, such as outcome and demographic variables. To arrive at recommendations for instructional design, one also needs to consider the *goal* of the instructional environment (Shute, Gawlick, & Gluck, 1998), such as ensuring mastery or efficient topic coverage.

Extreme positions are rarely helpful, and the concept of a single best method of instruction for everyone is overly simplified. On the one hand, traditional education, with its focus on content rather than the learning process, tends to lack a basic understanding of students. On the other hand, progressive education, as Dewey himself noted, can be too reactionary. That is, freedom for the sake of freedom is a weak philosophy of education according to Dewey (1938). Instead, he asserted, experience arises from the interplay of two principles—continuity and interaction. One's current experience is a function of the interaction between one's past experiences and the present situation. Dewey believed, like many educators who followed, that no single experience has preordained value. A rewarding experience for one person could well be a detrimental experience for another.

In short, as with fashion (e.g., Nehru jackets), cars (e.g., the Edsel), and toys (e.g., pet rocks), educational reforms tend to come and go, causing a flurry for some duration, but rarely influencing teaching practices in any lasting or significant way. According to Cuban (2004), and supporting the look-to-the-interactions perspective, there will never be a clear victory for either traditional or progressive education because students differ in their motivations, interests, and backgrounds, and learn at different speeds in different subjects. The bottom line is simply that there is no single best way for teachers to teach, or for children to learn, that optimally fits all situations. Features from both traditional and progressive ways of teaching and learning need to be incorporated into a school's approach.

Bridging the Chasm With Research

The idea of improving teaching through the application of science has been around since the earliest days of organized teacher training. Dewey believed that the scientific study of child development would improve classroom instruction by suggesting ways in which teaching might be fitted to the learner (Dewey, 1916). It was not until the 1960s, however, that government-funded research began expanding toward present-day levels. And it was during this time (1960s and 1970s) that aptitude-treatment interaction (ATI) research flourished. But despite the fact that hundreds of studies were conducted, the jury remained out, and ATI's popularity declined after the 1970s. It is likely that the reason for this decline is that the classroom data were confounded by many extraneous variables (e.g., personality of the teacher, instructional materials, classroom dynamics), making ATIs hard to find and difficult to interpret. During the 1990s, with the emergence of computers and the ability to control extraneous variables, interest renewed (see Shute & Towle, 2003, for more on this topic).

Anderson, Reder, and Simon (1996) provide compelling arguments in support of more research before the adoption of any educational techniques. They point out that new so-called theories of education are introduced into schools every day, solely on the basis of their philosophical or common-sense plausibility, but lacking in empirical support. Substantially more emphasis should be provided for responsible experimentation that explicitly tests such new ideas. In their article, they argue for the equivalent of an "FEA," analogous to the FDA, requiring well-designed clinical trials for every educational "drug" introduced into the market place. Six years later, this idea has materialized in the form of the What Works Clearinghouse, established in 2002 by the U.S. Department of Education's Institute of Education Sciences to provide educators, policy makers, researchers, and the public with a central and trusted source of scientific evidence of what works in education.

From the standpoint of science, experimental studies are far more convincing than descriptive and correlational ones, yet school personnel often ignore the more rigorous studies and adopt innovations suggested by the descriptive ones. For example, during the 1960s and 1970s, correlational studies suggested that enhancing self-esteem was related to improved achievement. This led to substantial changes in teacher training and schooling. Experimental findings to the contrary were ignored. For example, Scheirer and Kraut (1979) showed that self-esteem and achievement are correlated mainly because achievement enhances self-esteem, not because self-esteem enhances achievement.

Educational Needs and the Factors Fostering Flux

Current circumstances make it important and urgent to move to a new way of thinking about and conducting education. Technological advances, growth in research on cognition and learning, and other factors make successful outcomes much more likely. Success depends on what we do. We are in an excellent position to create a sea change, responsive to some of the urgent needs in education.

The basic premise of this chapter is that the seascape of education is unquestionably ready for an extreme makeover, and our goal should be to guide its transformation in the best, most effective direction possible, based on results from research. One salient source of educational discord is the No Child Left Behind (NCLB) Act of 2001, with its requirements for increased assessment and school accountability. Hundreds (if not thousands) of articles appear in the press each day, describing phenomena like a national "grassroots rebellion" against the No Child Left Behind Act, as reported by organizations such as NCLBGrassroots.org.

NCLB Dissatisfaction In general, dissatisfaction with the NCLB Act is neither a rejection of accountability nor a lack of commitment to narrow the achievement gap. Rather, the shared sentiment among many educators in the field seems to be that the pressure to teach to the test undermines quality education and deepens the adversarial relationship between parents and teachers. More specific complaints raised against NCLB include the following: (a) it is an unprecedented federal intrusion into education, historically an area reserved for states; (b) its one-size-fits-all approach ignores the realities of good teaching and learning; (c) the law devotes too much valuable class time to test preparation; and (d) it is too narrow in its substantive focus, concentrating on reading and mathematics to the exclusion of such basic skills as communication and creative problem solving (see, e.g., Civil Society Institute, 2005; Kahl, 2003).

New 21st-Century Skills Another factor contributing to the need for a sea change has to do with the aforementioned factory metaphor and its incongruence with our current information age. Students are not acquiring sufficient knowledge and skills to prepare them for careers in mathematics, science, and technology with the traditional approach to schooling, as evidenced by the Program for International Student Assessment (PISA) results (e.g., Lemke et al., 2004), described in more detail later in this chapter. Moreover, students today need new skills (e.g., information communication and technology [ICT] skills: how to define, access, manage, integrate, evaluate, and communicate information) to deal successfully with the deluge of data in the 21st century. The term "lifelong learner" describes this phenomenon and suggests (if not demands) that we change the way we structure learning and the way people access and acquire information and transform it into knowledge. Toward this end, we must figure out what skills we value, and support those for a society producing knowledge workers, not simply service workers. At the same time, we need to be cautious about moving from one extreme to the other, and to be informed by ongoing research-based tests of educational effectiveness, by which procedures, models, and curriculum are rigorously compared. "We need to look to science to give us answers. We need to engage our best researchers in research on how children learn … and how instruction can be improved" (Paige, 2001).

Major Educational Trends of Today

Over the past 10 years or so, some major educational developments have emerged and gained dominance, as indicated by their increased popularity at educational and psychological research conferences. These trends are characterized by "new" models of teaching and learning, but on closer inspection, many appear very similar to ideas originally envisioned by Dewey. The most salient of these trends relates to curricula characterized by tightly integrated formative assessments that are diagnostic,[7] criterion-referenced, linked to targeted instruction (or instructional prescriptions), and that fit the particular needs of the learner at just the right time (see Appendix for definitions of these terms).

Assessments Tied to Instruction Bass and Glaser (2004) describe the principles of what they refer to as "informative assessments," to draw attention to the instructional goal of improving student learning. They see the design of such assessments as having a substantial influence on the quality of information provided to teachers and students to support instructional decision making and more meaningful learning. This is, however, conditioned on presenting assessment results in an easy, intelligible, and actionable format—to both teachers and students. Shepard (2000) presents a constructive and comprehensive conceptual framework in which to house many of these new ideas and models. She describes how classroom assessment practices might be structured and implemented to be more effective in enhancing teaching and learning. She outlines the principles of a "social-constructivist" conceptual framework, bringing together cognitive, constructivist,[8] and sociocultural theories as a reformed and nicely blended view of education.

Another good example of this blended approach can be seen in Web-based cognitive tutors called "assistments," the merging together of assessment with instructional assistance into one system (see Razzaq, Feng, Nuzzo-Jones, et al., 2005, for more on the topic).

Black and Wiliam (1998a, 1998b) very clearly establish the importance of formative assessments to both teaching and learning. They conducted a large research review of the relationship(s) between assessment and learning, and their landmark papers have had a major influence on both research and the teaching profession. In addition, they originated the widely used distinction between (a) assessment *for* learning and (b) assessment *of* learning.

Finally, and in line with the "best of both worlds" position of this chapter, Stiggins (2002) argues that both assessment of learning and assessment for learning are essential. Unfortunately, while assessment *of* learning is currently well-entrenched in our educational system (such as through NCLB), assessment *for* learning is not. We need to strike a better, more scientifically informed, balance. For example, if formative assessments (representing assessments *for* learning) were employed throughout the school year, then at the end of the year or marking period, the need for formal summative tests (a common type of assessment *of* learning) would be greatly reduced. To accomplish this goal would require that the student data—collected, analyzed, and recorded by the formative assessments—be valid, reliable, and of a manageable, actionable grain size.

Student-Centered Practices Another trend appears to be renewed interest in student-centered approaches to teaching (e.g., Pellegrino, 2004), where teacher and student roles are basically redefined. The teacher becomes a facilitator of learning instead of the sole dispenser of knowledge, and students take more responsibility for their own learning. The main idea behind this approach is that learning is most meaningful when topics are relevant to the students' lives, needs, and interests and when the students themselves are actively engaged in creating, understanding, and connecting to knowledge (McCombs & Whistler, 1997). Students will have a higher motivation to learn when they feel they have a real stake in their own learning. In keeping with the idea of bridging the chasm between the traditional and progressive approaches, implementing student-centered practices will require the provision of more freedom than is currently in place, but in a structured way. For example, students can use assessment information to regulate and guide their learning. Sharing assessment information with students is a way to empower them (e.g., Brna, Self, Bull, & Pain, 1999; Zapata-Rivera & Greer, 2004), thus transitioning to a new role for students—from passive assessment recipients to active participants. Furthermore, self-assessments can provide another source of evidence, contributing to a more complete picture of what the student really knows (e.g., Mitrovic & Martin, 2002; Zapata-Rivera, 2003).

Cognitive Modeling The final big trend being applied to educational research is cognitive modeling, which refers to a set of ideas and procedures that come from cognitive psychology and computer science. Cognitive modeling is generally defined as the representation of what is inside the learner's head: thinking, knowing, and learning. Cognitive models can help predict or control complex human behavior, including skill learning, problem solving, and other types of cognitive activities (e.g., Aleven & Koedinger, 2002; Anderson & Lebiere, 1998; Heffernan & Koedinger, 2002). Computer tutors that have been built using cognitive models have been very successful in improving student learning, especially in mathematics (e.g., Koedinger & Anderson, 1998). One major advantage we have today compared with even a decade ago, is technology to engender and support many of the reform ideas, some of which were presented nearly a century ago. It has often been said that Dewey was ahead of his time. Perhaps now his time has come, particularly given the confluence of (a) the growing dissatisfaction with NCLB as a vehicle for educational reform (see Hart & Winston, 2005; Phi Delta Kappa, 2005); (b) the presence of the What Works Clearinghouse to evaluate new ideas and interventions; and (c) the collection of available technologies to support innovative ideas that were previously not easy, or even possible, to accomplish in the classrooms and culture of the past.

Issues and Solutions

For the remainder of the chapter, I will define specific educational issues and present concrete solutions, highlighting evidence-centered design (ECD) as a viable tool to design assessment to support learning.[9] This will be followed by a description of the theoretical foundation and implementation of a prototype system we are currently developing at ETS. The system is designed to help struggling middle school students learn mathematics—specifically Algebra I content. The prototype exemplifies the idea of merging assessment and instruction to support learning.

The combination of fields needed to accomplish these objectives includes assessment design, cognitive psychology, educational measurement/psychometrics, artificial intelligence, instructional system design, educational psychology, and others as well. The bottom line, however, is that it's *all about learning*, using informative assessments, tied to good instruction, integrated within the curriculum, and linked to state and/or national standards, in order to maximally support both teachers and learners.

The Problems We Face

In 2004–2005, the United States invested $536 billion in K–12 education, and another $373 billion for higher education (U.S. Department of Education, 2005). But although the U.S. is a world leader in education investment, nations that spend far less regularly achieve much higher levels of student performance (Lemke et al., 2004). The rest of this chapter will focus on assessments within the area of mathematics, but the arguments and findings are applicable to other areas as well, such as reading, science, and cross-cutting skills such as problem solving and reflection.

International Comparison of Mathematics Assessments America's 15-year-olds performed below the international average in mathematics literacy and problem solving, according to the latest results from PISA. The test, given in the spring of 2003, assesses the ability of 15-year-old students from various countries (including 30 of the most developed) to apply learning to problems with a real-world context (see Lemke et al., 2004). Students in the following countries outperformed the United States in mathematics literacy in 2003: Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Hong Kong-China, Iceland, Ireland, Japan, South Korea, Liechtenstein, Luxembourg, Macao-China, Netherlands, New Zealand, Norway, the Slovak Republic, Sweden, and Switzerland. These same 23 countries, plus Hungary and Poland, outperformed the United States in mathematics problem solving. U.S. 15-year-olds scored measurably better than their counterparts[10] in only 3 of 30 nations on the new international test of problem solving in math. Moreover, the U.S. has the poorest outcomes per dollar spent on education. In short, U.S. students are performing poorly on mathematics tasks that involve transfer of learning and problem-solving skills. We need to bolster our students' problem-solving skills to compete effectively internationally in the near future.

Widening Achievement Gaps Shifting attention from the international to the home front, there are also some disturbing differences in mathematics achievement among U.S. student subpopulations. Despite substantial educational reform efforts directed at poor and minority students across the last two decades, current data show large and growing achievement gaps between ethnic minorities and white students (e.g., Haycock, 2001; Lee, 2002). For example, in 1990, there was a 33-point gap between the scores of black and white students on the National Assessment of Educational Progress (NAEP) mathematics test at the eighth-grade level.[11] By 2000, the gap had grown to 39 points. Latino students were 28 points behind white students in 1990 and 33 points behind a decade later. In California in 2004, fourth- and eighth-grade black and Latino students were found to perform, on average, 3 years behind comparable groups of white students in mathematics. According to Mora (2001a, 2001b), it is reasonable to conclude that for students in California, the achievement gap is most likely due to factors such as language proficiency and its impact on literacy, which relates to accessibility issues, addressed next. And linking PISA findings and the achievement gap, Bracey (2004) analyzed 2003 PISA data, excluding Asian, black, and Hispanic students from the sample. When ranking only white U.S. students in relation to students from the other 30 countries, the U.S. ranked as follows: reading: 2, math: 7, and science: 4.

Accessibility The third main problem we face concerns the need in K–12 education for better curricula, including embedded diagnostic assessments, that are more "universally designed"—that is, more accessible, effective, and valid for students with greater diversity in terms of disability and English language capability. A committee of the National Research Council recently examined accommodation policies for the NAEP and other large-scale assessments. They reported that, "Overall, existing research does not provide definitive evidence about which procedures will, in general, produce the most valid estimates of performance for students with disabilities and English language learners" (National Research Council, 2004, p. 6). In addition to the call for universally designed assessments, there are accessibility issues associated with instructional materials. For example, most classroom materials (books, chalkboard, quizzes, etc.) tend to be written in English, and are highly visual in nature. Obviously, this presents obstacles for individuals who are not fluent in English and/or have visual disabilities. *If content is not accessible, it cannot be learned.*

Proaction So what can we do about these troubling findings? Obviously, many variables contribute to the poor showing by U.S. students relating to students in other countries; and within the U.S., by ethnic minority students. One thing we can do is focus on developing and evaluating, in controlled research studies, valid and reliable tools—technological and methodological—that can expedite the development and implementation of informative assessments that help teachers to teach, students to learn, and learning outcomes to improve. A key component of informative assessments is valid diagnosis; and a key component of valid diagnosis is good evidence, i.e., performance data that form the basis for inferences about proficiencies. Fortunately, technological, educational, and psychological measurement approaches have advanced, and we can now more accurately diagnose student proficiencies. Information, collected and analyzed from the student, can inform both the teacher (for decisions about what to do next, with the student or classroom) and the student (who can use the information to understand what he or she did wrong or right). In addition, proficiencies themselves may be validated through the examination of data. That is, careful inspection of the data provides valuable insights into whether the proficiencies are effective and useful, as defined, or whether they should be modified.

Research, methods, and models will now be described that can be used to design and implement informative assessments. This is followed by a description of a prototype system that is being developed at ETS as a possible solution to some of the major problems facing American education today.

Specific Solutions

This section begins with a brief review of relevant learning research; i.e., timely feedback, tailored content, and multiple representations. Together, these three areas form the research basis for the prototype solution described later in this chapter.

Timely Feedback Timely feedback in the context of problem solving is generally viewed as important to enhancing student learning (e.g., Corbett & Anderson, 1989; Epstein, Lazarus, Calvano, et al., 2002). In addition to exerting positive effects on achievement, feedback has also been found to be a significant factor in motivating learning (e.g., Narciss & Huth, 2004). However, the story is not quite so simple. According to Cohen (1985), feedback is "one of the more instructionally powerful and least understood features in instructional design" (p. 33). Because of the many differences in types of feedback, results relating to its timing and effects on learning outcomes can conflict. Mathan & Koedinger (2002) review some conflicting results on the timing of feedback, and conclude that the effectiveness of feedback depends on the nature of the task and the capability of the learner. This suggests the need to further explore optimal ways to tailor the type and timing of feedback to learning tasks and to students' individual needs and characteristics (e.g., Schimmel, 1988; Smith & Ragan, 1999).

Tailored Content Adjusting learning environments and content to suit student needs can substantially improve learning (e.g., Corno & Snow; 1986; Shute, 1993). Computer-based adaptive learning systems are beginning to accommodate differences in learner interests, aptitudes, and background (e.g., Bajraktarevic, Hall, & Fullick, 2003; Conlan & Wade, 2004; De Bra, Aerts, Berden, et al., 2003; Papanikolaou, Grigoriadou, Magoulas, & Kornilakis, 2002; Weber & Brusilovsky, 2001). These systems effectively can act as personal tutors, build models of learners, and intervene with relevant information when needed. Technology has advanced to the point that we can more easily implement adaptive instructional techniques on the Internet (e.g., differential sequencing of content, depending on learners' needs). See Brusilovsky (2003) and Brusilovsky and Vassileva (2003) for more on this topic.

Multiple Representations of Content Finally, presenting alternative representations of the same concept (in tasks, examples, and so forth) can not only augment comprehension but also accommodate various disabilities, preferences, or learning styles. Research supports the importance of multiple-strategy use and representations in mathematics in terms of skill acquisition, understanding, and transfer (e.g., Katz, Lipps, & Trafton, 2002; Koedinger & Tabachneck, 1994; Tabachneck, Koedinger, & Nathan, 1994). The requirement for integrating different types of response formats, and hence representations, is also consistent with the research-based expectation in state and national standards that students should be flexible in moving across representations (tables, graphs, expressions). Moreover, developing and accessing multiple representations supports deeper understanding (e.g., Shafrir, 1999). Designing informative assessments with these three research-based features (timely feedback, tailored content, and multiple representations) is a reasonable response to counter some major educational problems. That is, with traditional education, by the time the results of high-stakes accountability tests are disseminated, it is usually too late to effect change in the classroom to address weak areas or misconceptions. We want to develop tasks that have been designed not only to provide feedback about the correctness of the response, but also provide guidance on areas of misconception, and are presented in a timely manner (usually immediately, with our solutions). Examples will be provided later in this section describing a prototype system called MIM (Mathematics Intervention Module).

This kind of educational support system—with immediate diagnostic feedback, multiple and varied tasks, and tailored to a learner's specific and current needs—is expected to significantly help students overcome procedural errors and areas of misconceptions. Furthermore, summary data provided to the teacher can allow her to modify the instructional approach and suggest further activities for a student or class based on targeted problem areas. The feedback can also be used by students to guide self-study and reflection. Over the long term, such an approach should help students understand the material better and improve their performance on high-stakes tests (Mory, 2004).

Methods for Developing a Prototype Solution

The research considerations and methods that we are combining in our prototype solution include (a) individual differences, (b) diagnostic assessments, and (c) instructionally rich learning environments. As part of this process, we are extending the scope of ECD, as originally formulated with its assessment design focus, to embrace learning as well.

Individual Differences Individual differences are typically defined as persistent and measurable aptitudes or attributes that distinguish people from one another. These variables may be used to predict performance on some learning tasks (see Shute, Lajoie, & Gluck, 2000). Disparities among students that are relevant to education can be cognitive, affective, perceptual, or demographic, or can involve other characteristics. We need to accurately identify variables that affect learning, and then offer appropriate supports, as needed. A key word here is "appropriate," as we need to ensure that accommodation for overcoming accessibility barriers, for example, does not also invalidate assessment results. The point is that students come to any new learning task with differing profiles. As educators, we want to take what we already know about students and add to that an understanding of what they are doing in real time in the learning environment. We can then combine that information with knowledge about strategies for bringing individuals to a higher level of knowledge, and adapt instruction to carry out those strategies. Valid and reliable cognitive diagnoses, then, are essential to learning environments that adapt to users' needs. According to Bass and Glaser (2004), taking full advantage of informative assessments requires the use of adaptive teaching techniques that yield information about the student's learning process and outcomes. This allows teachers to take appropriate instructional actions and make meaningful modifications to instruction. Two approaches to adaptation are described below.

One way in which content can be customized for a student is through *microadaptation*, the real-time selection of content in response to a learner's inferred knowledge and skill state (Shute & Towle, 2003; Vassileva & Wasson, 1996). Microadaptation occurs during the learning process and is sometimes referred to as domain-dependent adaptation. It can also be thought of as a set of small, ongoing formative assessments. Decisions about content selection are typically based on performance and on subsequent inferences of student knowledge and skill states.

The other approach to adapting content is through *macroadaptation*—the customization of content in line with learner qualities, such as stable cognitive or perceptual abilities. In contrast with microadaptation, macroadaptive decisions are domain-independent and are based on learner information that is usually, but not always, collected before instruction begins (see Shute, Graf, & Hansen, 2005; Snow, 1992, for more on this topic). Macroadaptation relates to decisions about the format and sequence of the content presented to the learner. For a review of some specific macroadaptive examples from the literature, see Shute, Lajoie, and Gluck (2000).

These two forms of adaptation are not necessarily incompatible and may, in fact, improve learning even more when combined. Microadaptation is typically applied to the problem of *what* to present and when to present it, while macroadaptation is applied to the issue of *how* it should be presented. The success of either type of adaptation, however, is a function of the validity and reliability of the underlying assessments.

Assessment Design Evidence-centered design (e.g., Mislevy, Steinberg, & Almond, 2000, 2003) provides (a) a way of reasoning about assessment design, (b) a way of reasoning about examinee performance, and (c) the means to unify and extend probability-based reasoning to assessment (e.g., to traditional standardized tests, classroom tests/quizzes, simulations, gaming environments, and portfolios). The basic idea of ECD is to specify the structures and supporting rationales for the evidentiary argument of an assessment. By making the evidentiary argument explicit, the argument becomes easier to examine, share, and refine. Argument structures encompass, among other things, the claims (inferences) one wishes to make about a student, the observables (performance data) that provide support for those claims, the task performance situations that elicit the observables from the students, and the rationales for linking it all together. The three main models used in ECD are

- *Proficiency Model*: Establishes claims about a particular piece of knowledge, skill, or ability. The proficiency model describes what is to be measured, conditions under which the ability is demonstrated, and the range and relations of proficiencies in the content area.
- *Evidence Model*: Defines the evidence needed to support the claims. Evidence models describe what is to be scored, how to score it, and how to combine scores to support claims. These models thus establish the boundaries of performance and identify observable actions that are within those boundaries.
- *Task Model:* Identifies tasks that are able to elicit that evidence. Task models specify the inputs required to perform the observable actions as well as the outputs (work products) that result from performing the observable actions.

Cognitive Diagnosis To determine student strengths and weaknesses, and figure out the nature and extent of difficulties in a student's problem solving efforts, we need to design tasks such that this information can be disentangled and interpreted in valid and reliable ways (see Hunt & Minstrell, 1996; Minstrell, 1992, 2001, for more on this topic). A good diagnostic assessment system should be able to infer proficiency estimates accurately for a student, at various grain sizes.[12] This process begins with the design of a reasonable (i.e., accurate and informative) proficiency model, which provides the basis for task-level (i.e., real-time, formative) and overall (i.e., summative) level diagnoses to occur. This is a very challenging undertaking, and we are currently exploring ways to use cognitive models to integrate evidence of student knowledge gathered from a variety of formative and summative sources. Information from student interactions with tasks or problems is automatically analyzed based on preestablished scoring rules, to inform and update relevant proficiencies. Task-level diagnoses can provide immediate support to the student via task-specific feedback; estimates of more general proficiencies provide the basis for decisions concerning what to do next, such as selecting a new task or offering other content to the student, providing practice, or some other instructionally helpful activity. This is all accomplished behind-the-scenes, on the computer, via selection rules and/or algorithms. Alternatively, diagnostic results can be handed off to the teacher in the form of instructional prescriptions or suggestions about what to do next, for the student or for the entire class.

Proficiency estimates can assume a variety of forms, from simple percent-correct data to probabilistic estimates of mastery of knowledge/skills using regression equations, to item response theory (IRT), multidimensional IRT models, or Bayesian networks (e.g., Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Mislevy, Almond, Yan, & Steinberg, 1999; Reckase, 1997; Shute, 1995). In all these cases, diagnostic assessment requires students to do something (i.e., to produce a "work product") in order to demonstrate knowledge/skill capability on specific tasks. The more student data collected, the more accurate the inference. Thus, it is very important in assessment design to ensure an array of activities with which a learner can interact, receive targeted feedback, and demonstrate his or her level of performance. Interpretation of proficiency is a function of the quality of the evidence collected. In a valid proficiency model, each piece of knowledge, skill, and ability is linked to more than one task so that evidence of a student's performance can be accumulated in a number of different contexts. In a hierarchical proficiency model, evidence of one skill's mastery can also feed into mastery estimations for related skills. An example of a proficiency model is presented later in the context of our prototype system, MIM.

Putting It All Together

To diagnose student status at the task level, and to infer student status at the proficiency level, we employ a variety of technological solutions in our assessments, such as automated scoring of different constructed response types, automatic item generation, adaptive testing, and the capability to present or simulate "authentic" problem-solving contexts. Again, it is important to ensure that each of these are weighed against concerns for construct validity, equity, and access (Bennett & Bejar, 1998; Shute, Graf, & Hansen, 2005).

For implementation of these ideas—which can run the gamut from paper-and-pencil to computer delivery—consider the 4-process model shown in Figure 6.1 (Almond, Steinberg, & Mislevy, 2002). This model specifies the following cycle, shown by the four circles (i.e., main processes) at the corners of the figure: (1) select a task (using a linear, adaptive, or other sequencing algorithm), (2) administer the task, (3) collect evidence and score the response, (4) update the student model,[13] and return to the first step (i.e., select the next task). This process continues until a termination criterion is met (e.g., some preestablished threshold is exceeded, time runs out, or there are no more tasks).

In summary, student responses to assessment tasks, as well as patterns of responses, serve as the primary evidence of proficiencies. This information is culled directly from student behaviors and work products as they interact with and complete items within an assessment task (or task set). Based on exactly what the student produces in response to a given problem-solving task (i.e., the

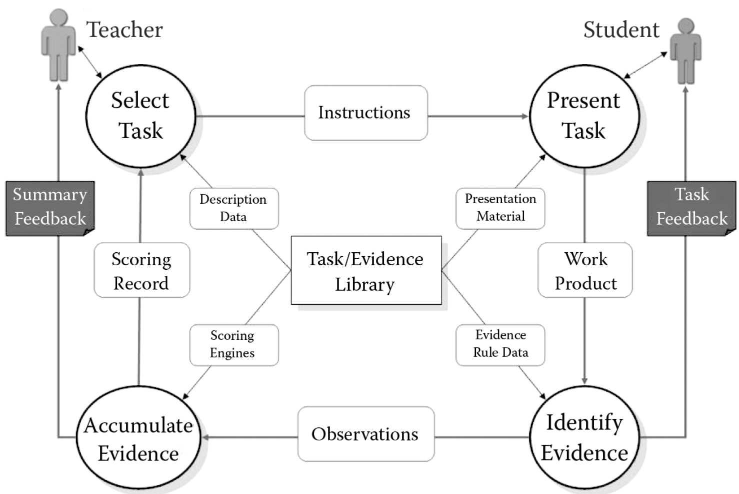

Figure 6.1 Four-process model [modified from "Enhancing the Design and Delivery of Assessment Systems: A Four-Process Architecture" by R. G. Almond, L. S. Steinberg, and R. J. Mislevy, 2002, *Journal of Technology, Learning, and Assessment,* 1(5), 1-63. Copyright 2002 by the Journal of Technology, Learning, and Assessment. Adapted with permission].

evidence), inferences can be made about the source of the problem or strength of a set of skills. Open-ended tasks typically invoke more varied evidence than do multiple-choice responses. ETS has been developing tools to analyze and evaluate various open-ended response types, discussed next.

Mathematics Intervention Module (MIM)

We are currently developing a mathematics intervention prototype, MIM, using ideas and methods described earlier in this chapter. The general topic was selected after consulting with teachers who identified Algebra I as a consistent obstacle for students; and within Algebra I, identified a few particularly difficult learning objectives or standards. We chose one of the most difficult objectives for our initial module: *Translate word expressions to symbolic expressions or equations and then solve and/or graph*.[14]

What Is MIM and How Does It Work?

MIM is an online application designed to help students become proficient in state mathematics standards. The initial focus is on Algebra I, but it may be extended to other subjects in subsequent releases. The module is based on a proficiency model that describes the skills that must be mastered for a student to be judged proficient in that standard. Each module presents students with open-ended questions dealing with the various skills identified in the proficiency model. These questions require the student to respond with a number, an expression or an equation, a graph, or text,[15] all of which are automatically scored.

Diagnostic Feedback All responses in the intervention module are automatically evaluated, with immediate and helpful feedback provided to the student. Feedback is directed at the error that the student has made, and is not simply, "Wrong. Please try again." Similar to a human tutor, MIM attempts to give some indication of why the student's answer was wrong. The student is given three attempts to answer each question correctly, with progressively more detailed feedback provided along the way if the answers are incorrect. The correct answer, with an associated rationale, is presented if the student answers incorrectly three times. In addition, if the student is judged to be in need, the module presents a short (2–4 minute) instructional video that covers the problematic skill. These "instructional objects" reinforce the learning that is taking place as the student works through the questions and reads the feedback.

Instructional Objects A specific instructional object (IO) is presented in the case where a student has gone through all three levels of feedback for a given problem. There are about 16 IOs that have been developed for the current MIM prototype. Within an IO, the flow of instruction proceeds as follows: (a) introduce the topic using concrete and engaging context, (b) state a particular problem that needs solving, (c) provide relevant definitions, (d) illustrate the concept within different examples (both prototypical and counter-examples), (e) provide sufficient practice and interactivity, and (f) conclude with summary and reflection screens. Reflection activities can also be used to gather evidence of student knowledge, assuming that these activities are interactive. Figure 6.2 shows a screen capture from an IO on the topic of Use Properties of Equality to Simplify Equations. The IO begins by using a scale as an analogy to "balancing both sides of an equation." A definition is presented, which explains why, mathematically, the scale is balanced. Following screens in the IO show examples of what happens—to the scale and the equation—when weights are added and removed.

Practice Opportunities Depending on classroom needs and other factors, the teacher has the option of assigning multiple-choice questions for additional practice on each skill. The teacher can (a) require these practice questions of all students who seem not to have mastered the skill, (b) make the practice questions optional, or (c) configure the module so that the practice questions are not delivered.

Integrating Knowledge and Skills The final section of the intervention module is a set of integrated open-ended questions that deal with a common theme or contextual situation. These questions reflect the standard as a whole. Like the open-ended questions earlier in the module, these integrated questions involve responses

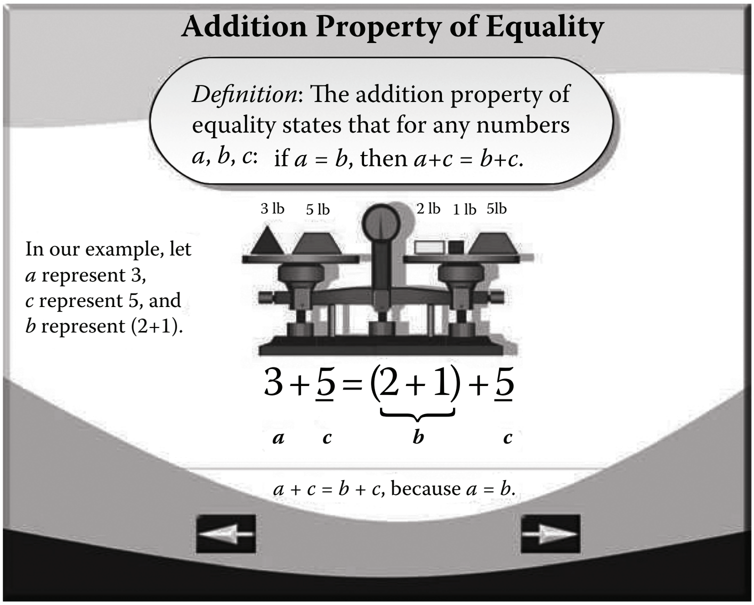

Figure 6.2 Screen capture from a MIM instructional object.

that require the entry of a number, an expression or an equation, a graph, or text.

Information to the Teacher After the student completes the intervention module, the teacher receives a summary report. In addition, the teacher can review the student's entire session, viewing the student's responses to each question. Classroom summaries are also available, so that teachers can see at a glance how their students are progressing on the target standard.

Proficiency Model A proficiency model generally describes the skills that must be mastered to be judged "proficient" in relation to a specific standard, and displays the relationships among these skills. The simplified proficiency model shown in Figure 6.3 analyzes the standard: Translate word expressions to symbolic expressions or equations and then solve and/or graph. By working down the model, one can see how the component skills can be isolated. In this standard, "word expressions" means the information contained in a story, a contextual description, or some other real-life situation. At a high level, this standard can be divided into three parts, each corresponding to a separate skill, and each represented by a node (three white ovals) on the model. The first skill is to translate the information given in the story into an equation or graph or some other symbolic expression. The second skill is to solve the equation, and the third is to graph the equation and obtain useful information from the graph. For the purposes of this model, we assume that the equations and graphs are linear.

The first skill (translate context to equations and/or graphs) can be further divided into several sub-skills. To translate contextual information into an equation or graph, one must first identify the variables, and then translate the operations (addition, multiplication, and so on) that connect the variables, and, finally, put it all together correctly to form the relevant equation. Each of these three skills is represented by a node within the model, and each node is connected to its parent node, *Translate context to equations and/or graphs*. In addition, dotted lines connect the third sub-skill with the first two because the third sub-skill requires the proper application of the first two. As shown in the proficiency model, these nodes are faded. Due to constraints in the current project, we could not fully implement the mathematical content for these skills. Instead, we teased out part of this content area and displayed it as a separate skill—entering contextual information into a table and then translating the table into a linear equation or graph. This skill is displayed as a gray node, indicating that this is one of the skills implemented in the current release of the intervention module.

A similar analysis applies to the second high-level skill (solve linear equations). This skill can be divided into three sub-skills: (1) use the rules of algebra to simplify expressions, (2) use the rules of algebra to simplify equations, and (3) combine the first two skills to solve equations. Again, each of the three skills is connected to the parent skill. In addition, the third skill (*Apply algebraic properties to solve equations*) is connected by dotted lines to the first two skills as it represents a proper application of the first two. All three of these nodes are displayed as gray because all three are implemented in the intervention module.

The third high-level skill (graph linear equations) is subdivided into three component sub-skills: understand intercepts, understand slope, and use knowledge of intercepts and slope to graph equations and determine equations from graphs. In

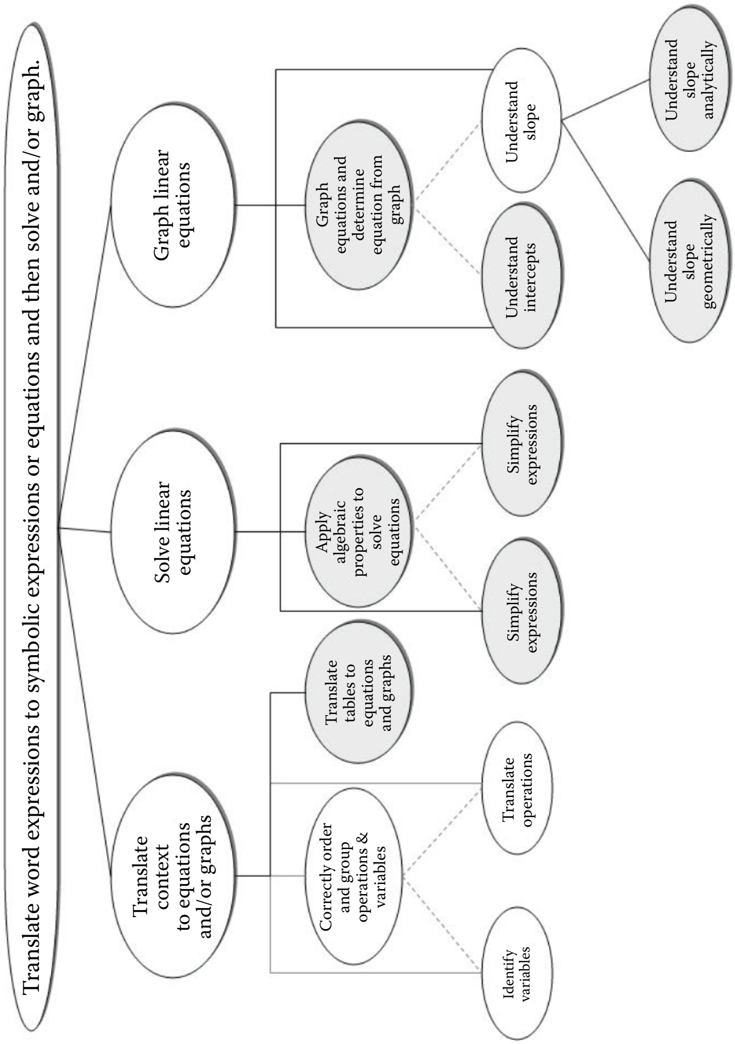

Figure 6.3 Simplified proficiency model for MIM (Fife, Shute, Underwood, & Graf, 2005).

addition, the *Understand slope* skill is further divided into two parts: *Understand slope geometrically* and *Understand slope analytically*. The "leaf nodes" (i.e., nodes with no children, or lower levels) are displayed as gray and are implemented in the intervention module.

Example of an Integrated Task Set in MIM

The following example is an isomorph of a problem from a set of ETS-owned content (Marquez, 2003). This integrated task set, as mentioned earlier, is presented at the end of the module, and its function is to assess the conjoined knowledge and skill elements. Finding a solution to the task requires the student to graph a line, find the equation of the line, identify the *y*-intercept and slope, state their significance in the context of the problem, and extrapolate data.

In the Music World Task, each node in the proficiency model may be linked, via different evidence models, to a number of tasks. As the student interacts with the system and answers questions, evidence is accumulated and the student model is updated. If a student demonstrates that she can calculate the slope using points on a graph and interpret what it means in the context of the problem, the corresponding nodes in the proficiency model will show higher estimates of mastery. Moreover, because of the hierarchical nature of the proficiency model, the parent node, "Understands slope," may also automatically increase slightly. The converse is true for failing to solve the problem correctly. In general, proficiency information in the student model can highlight specific areas that need more instructional support.

Diagnosis To further facilitate the diagnosis of student performance, the system knows about a number of common misconceptions in relation to the skills in the proficiency model. To illustrate, in relation to the calculation and interpretation of the slope, some of the salient misconceptions and errors include inaccurate symbolic and graphical modeling of data, misunderstanding of slope as a rate of change, misinterpretation of slope and y-intercept in real contexts, and inability to use the equation of a line as a tool to predict linear behavior (i.e., extrapolation). These are used as indicators to help diagnose the problems with the knowledge and skills in the

Music World Task You found a new Web site that claims to offer the best deal around for buying music CDs. The Web site isn't clear about the cost for each CD or the cost of shipping and handling (except to say shipping is a flat fee), but it does give you the following information:

Number of CDs Ordered	**1**	**2**	**3**
Total Cost (with Shipping & Handling)	\$9	\$14	\$19

1. Plot the data in the table on the graph (provided). Draw the line that contains the data points.
2. Assume that total cost is a linear function of number of CDs ordered.
   a. Write an equation of the line that contains the data points. Show your work.
   b. What is the slope of the line that contains the data points?
   c. What does that slope represent in the context of this problem?
   d. What is the y-intercept of the line that contains the data points?
   e. What does that y-intercept represent in the context of this problem?
3. Your friend says that he can get 15 CDs from the Web site for \$64.00. Is your friend correct? Explain.

proficiency model. A teacher or instructional module, armed with this information, can be considerably more effective in providing a targeted intervention.

Scoring Following are some general requirements for a student to get a maximum score per item element in the Music World example.

1. Graphs points correctly with respect to the axes.
2. a. Writes a correct equation for the line based on an accurate reading of the graph or correct calculations using a linear form.
2. b. Gives the correct slope based on the graph or the equation written in part 2a.

2. c. *Gives clear and correct interpretation of slope in context.*
2. d. Gives the correct *y*-intercept based on the graph or the equation written in part 2a.
2. e. Writes clear and correct interpretation of *y*-intercept in context.
3. Writes an answer and justification that are correct, based on the equation given in question 2 or based on the graph in question 1.

Let's look at requirement 2c in more detail. The learning objective is that the student can give clear and correct interpretation of slope in the context of the problem. The work product is a written (typed) response to an assessment item. The three levels are as follows:

- *Low*: Student describes something that does not relate to the contextual variables related to slope (i.e., something other than CD price and shipping and handling).
- *Medium*: (a) Student describes slope in correct definitional terms (rise/run), but with no link to the context; or (b) student describes the correct contextual variables, but with an incorrect relationship.
- *High*: Student describes the correct contextual variables with the correct relationship (total cost of each CD including shipping and handling).

Now suppose that a student types in the response, "Slope is the rise over the run," which the system recognizes as correct but having no context. The system displays feedback appropriate to the inferred (common) error.[16] For example: *"You've told me the correct definition of slope, but you need to explain it in terms of the problem. For example, what do the rise and run in the graph have to do with the cost of CDs and shipping and handling?"* The student then tries again, and the system uses progressive levels of feedback for scaffolded support of learning.

Updating the Student Model After each response, or some other defined interval, the system updates the relevant nodes in the student model. Thus estimates of relevant proficiencies would be updated according to the evidence model. The example above showcased an ETS tool called C-RATER that can capture and analyze text input. Another ETS tool can "read" points and lines on a graph, and compare values to scoring rules (Bennett, Morley, Quardt, & Rock, 2000). Diagnostic feedback can similarly be embedded in xml files for the task, and linked to different responses. See Figures 6.4 and 6.5 for an example of graph analysis and feedback.

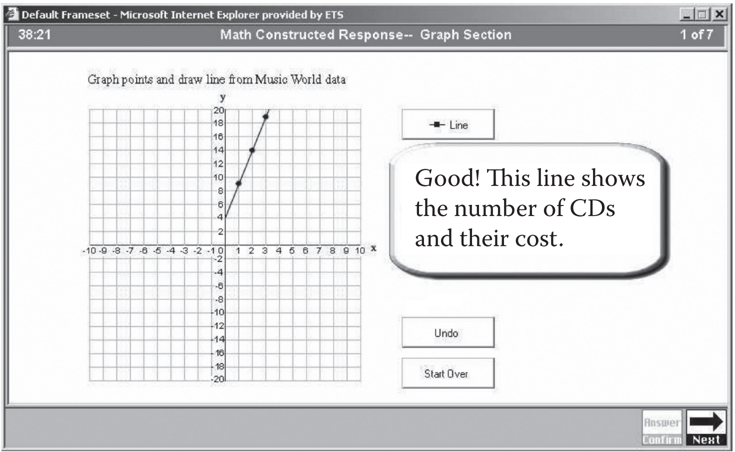

Figure 6.4 Graph analysis with diagnostic feedback shown superimposed on the work product. Additionally, the program evaluates the expressions and equations that a student types (see Figure 6.5) for mathematical accuracy/equivalence. For more information on the various automated scoring methods, see Bennett, Morley, Quardt, and Rock (2000) and Bennett, Morley, and Quardt (2000).

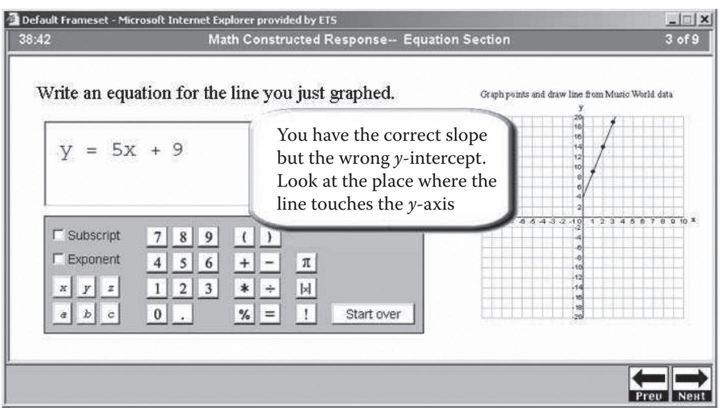

Figure 6.5. Equation analysis with diagnostic feedback shown on the work product.

Instructional Design The various elements of an intervention module—the open-ended questions, instructional videos, and multiple-choice practice questions—are presented to the student according to a carefully planned instructional design, based on principles of assessment and instruction that have been developed by researchers at ETS (Kuntz et al., 2005). We used the principles underlying ECD to develop the underlying proficiency model, scoring rules, and informative assessment tasks, and incorporated into MIM the three research-based features to support learning discussed in this chapter: timely diagnostic feedback, tailored content, and multiple representations of concepts. Finally, we plan to pilot test the first MIM module, employing three learning conditions: control (classroom instruction only), practice (classroom instruction and practice problems on relevant topics), and treatment (classroom instruction and the MIM prototype). This will be administered to several hundred students in school districts in southern California. Of interest will be the value-added of MIM over the other two conditions in relation to student learning.

Summary and Conclusions

> If we take no action to improve teaching and learning, we will just be using children as "extras" in a high profile political drama while undermining the social and economic prospects of the nation in the process.
>
> Kurt Landgraf, 2001

The chasm between traditional and progressive educational philosophies, described in the beginning of this chapter, is real. And support on both sides is fervent. Neither position is an educational panacea—both have enormous strengths and serious limitations. I have suggested merging the best features from each into a unified and more powerful educational approach. *There are two gifts we can give our students—one is roots, the other wings.* The traditional approach provides the roots, and the progressive approach provides the wings. Table 6.1 characterizes four assessment variables (main role in the classroom, frequency of administration, typical format, and feedback) characteristic of each of the three approaches: traditional, progressive, and unified.

Evidence-based learning, an extension of evidence-centered design for assessment, forms the foundation of the unified approach

TABLE 6.1 Assessment Variables Across Three Educational Approaches[a]

	Traditional	Progressive	Unified
Role of Assessment	Assessment of learning, to quantify fixed and measurable aspects of learner knowledge, skills, and abilities. Used for accountability purposes, often with norm-referenced tests. Produces a static/snapshot of the student	Assessment for learning, to characterize important aspects of the learner. Focus is on aspects of student growth, employing criterion-referenced tests, used to help learners learn, and teachers teach better	Both assessments of and for learning have important roles in education. Need to know where the student started, where currently is, where heading, how the journey is progressing, and ultimately degree to which destination is attained
Frequency of Assessment	Infrequent, summative assessments using standardized tests. Focus is on product or outcome (achievement) assessment. Typically conducted at the end of a major event (e.g., unit, marking period, school year)	Intermittent, formative assessment. The focus is more process oriented (but needn't exclude outcomes). Assessments of this type are administered as often as desired and feasible; monthly, weekly, or even daily. Administration is informal	Because assessments are embedded into the curriculum, a constant flow of evidence (student performance data) informs teachers and students. Data include both product (what) and process (how) assessment, as well as collaborative, negotiated, and/or self-assessment
Format of Assessment	Objective assessments, often selected responses. Focus on whether test is valid and reliable more than the degree to which it supports learning, per se	Constructed responses and authentic context, collected from multiple sources (e.g., quizzes, portfolios, self-appraisals, presentations)	Different task types and performance data are acceptable, from selected to constructed responses. Possible to extract data from problem-solving tasks, simulations, and other novel environments. Multiple representations used
Feedback	Correct or incorrect responses to test items and quizzes, or just overall score. Support of learning is not the intention	Global (proficiency) diagnoses attempted, with ways to improve (learning and teaching) suggested. Feedback is crafted to be helpful, rather than judgmental	Task-level and general diagnoses from item to proficiency level; procedural errors and misconceptions addressed and supported with immediate and timely help. Customized feedback is on the horizon

[a] This characterization is intended to convey general aspects of each approach in terms of these assessment variables, and should not be viewed as definitive categorizations.

proposed in this chapter for the design and development of informative assessments that can contribute toward improved teaching and learning. The ETS tools and approaches described herein collect and analyze a variety of evidence from the student across extended periods of time. These data collectively serve as the basis for estimates of proficiency status. This approach for developing informative assessments involves explicitly linking performance data to claims about learner proficiencies via an evidentiary chain, and therefore is more likely to be valid for multiple intended purposes.

Given the range of technology and tools at our disposal, at ETS and elsewhere, assessment tasks can now handle a variety of representations as input and output (e.g., graphics, equations, and text responses). Even more input/output options are on the educational horizon, along with new models and technologies to support learning. Using these tools for assessment *of* and *for* learning, as in the unified approach, can support our teachers and students, and at the same time, satisfy the requirements of NCLB.

Lessons Learned As noted earlier, Dewey's innovative educational reform ideas did not pan out. What can we learn from that? First, the "school as a factory" metaphor undervalues and undermines teachers. For example, teachers have the very important responsibility of educating future generations of citizens, but their salaries are not nearly commensurate with their responsibilities, leading to a growing shortage of quality teachers. McCoy (2003) surveyed teachers in their first 3 years of teaching, to analyze reasons for teachers leaving the profession. The following categories were identified: societal attitude toward teachers, financial issues, time scarcity, workload, working conditions, and relationships with students and parents. Informative assessments cannot directly help with the first two issues (attitudinal and financial), but they can help with the last four—freeing up more time for teachers to do their jobs, reducing workload, improving working conditions, and fostering better communication and relationships among teachers, students, and parents. The second reason cited for the failure-to-thrive of Dewey's ideas relates to the zeitgeist of practical education, and the consequent restriction of subject matter that occurred at the time. NCLB is threatening a similar shrinkage with its primary focus on mathematics, reading, and, soon, science. But so many other subject areas (e.g., history and art) comprise a well-rounded education. Another ramification of NCLB is the current trend of teaching to the test. Informative assessments can help reverse that trend by providing ongoing information about the student (to the teacher, student, parent, and so on), thus reducing our currently heavy reliance on formal standardized tests (see Pellegrino, 2004). This, in turn, could re-focus education on its primary mission, which is ensuring that our children learn the things they need to learn to contribute as well-adapted, effective members of society.

The third reason that Dewey's ideas did not become widely implemented concerns the use of measurement in his era. Although student abilities and intelligence were extensively "measured," it was not done to help them learn better or otherwise to progress. Instead, the main purpose of testing was to track students into appropriate paths, with the understanding that their aptitudes were inherently fixed. Thankfully, we have evolved in our thinking since then. We also have considerably more tools and techniques to promote learning, as described in this report. Students and teachers are both expected to benefit from (a) a unified approach to education, and more specifically, (b) informative assessments. For students, *tailored content* means that they receive subject matter based on their specific needs. Needs are determined from prior performance data from the student. Content is tailored to individual proficiency levels—not too easy or too hard. Other types of adaptations are possible as well, as discussed earlier. In addition, *diagnostic feedback* is believed to enhance learning by providing immediate diagnosis, assistance, and challenge, in relation to problematic and successful areas. Finally, working with *multiple representations* of concepts promotes flexible and deep comprehension. All these features, and others as well, are expected to increase learning, but must be subjected to rigorously controlled evaluations.

From the teacher's perspective, *timely and flexible reporting* of informative assessments permits the teacher to generate and view reports that show performance of students—as individuals, as groups, and as a whole class, and so on. These reports-on-demand can be used to modify instruction, exactly when it really counts—when students reach an impasse or when they display clear misconceptions. Reports can also show progress over time, as opposed to just a snapshot, for individual students as well as for the class. When coupled with instructional prescriptions or suggestions about what to do next, the reports would be even more valuable to teachers.

As a country, we are poised, with our current collection of research, approaches, and tools, to make a substantial, positive sea change to education. This chapter illustrates the pros and cons of different educational approaches and philosophies and advocated the integration of the "best of both worlds" in a unified approach. This needs to begin with a rational understanding of what we value in terms of proficiencies to be instructed and assessed, now and with an eye toward the future. Knowing what a student knows comes from obtaining quality evidence, which in turn is obtained from carefully designed assessment tasks. The approach described in this report is intended to be powerful, for students and teachers, especially when joined with sufficient practice opportunities and targeted feedback. The next step is to systematically test these ideas, and others that follow, in a series of controlled evaluations. The key to accomplishing our sea change goal is to work in a unified manner, toward a shared vision of excellent education for all.

Author Notes

Valerie Shute works at Educational Testing Service, Princeton, New Jersey, in the Center for Assessment Innovation and Technology Transfer.

I'd like to thank my wonderful colleagues for earlier reviews of this chapter: Eric Hansen (who had the tough job of reviewing the first and thus very rough draft), Aurora Graf, Diego Zapata, Jim Fife, Malcolm Bauer, Dave Kuntz, Carol Dwyer, and Evelyn Fisch. I'd like to additionally acknowledge Jim Fife's summary of MIM and analysis of its proficiency model, described in this chapter.

References

Adams, G. (1996). Project follow through: In-depth and beyond. In G. Adams & S. Engelmann (Eds.), *Research on direct instruction*. Retrieved August 25, 2005, from http://darkwing.uoregon.edu/~adiep/ft/adams.htm

Aleven, V. A. & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, *26*(2), 147-179.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5), 1-63. Retrieved August 30, 2005, from http://www.jtla.org

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5-11.

Bajraktarevic, N., Hall, W., & Fullick, P. (2003, May). Incorporating learning styles in hypermedia environment: Empirical evaluation. Paper presented at the workshop on Adaptive Hypermedia and Adaptive Web-based Systems, Budapest, Hungary. Retrieved August 26, 2005, from http://wwwis.win.tue.nl/ah2003/proceedings/paper4.pdf

Barton, P. E. (2005). One-third of a nation: Rising dropout rates and declining opportunities (Research Report No. PIC-ONETHIRD). Princeton, NJ: Educational Testing Service. Retrieved August 29, 2005, from http://www.ets.org/Media/Research/pdf/PICONETHIRD.pdf

Bass, K. M., & Glaser, R. (2004). Developing assessments to inform teaching and learning (CSE Report 628). Los Angeles: Graduate School of Education and Information Studies, UCLA. Retrieved April 20, 2007, from http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/00000006/80/2b/b3/ac.pdff

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9-17.

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests. *Applied Psychological Measurement, 24,* 294-309.

Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (2000). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education, 13*, 303–322.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment.* London: School of Education, King's College.

Bobbitt, J. F. (1912). The elimination of waste in education. *The Elementary School Teacher*, *12*, 259-271.

Bracey, G. W. (2004). *Setting the record straight: Responses to misconceptions about public education in the U.S*. (2nd ed.). Portsmouth, NH: Heinemann.

Brna, P., Self, J., Bull, S., & Pain, H. (1999). Negotiated collaborative assessment through collaborative student modelling. In R. Morales, H. Pain, S. Bull, & J. Kay (Eds.), *Proceedings of the workshop: Open, interactive, and other overt approaches to learner modelling* (pp. 35-42). Presented at AIED 1999, Le Mans, France.

Brown, A. L., & Campione, J. C. (1990). Communities of learning and thinking, or a context by any other name. *Contributions to Human Development, 21*, 108-126.

Brusilovsky, P. (2003). Adaptive navigation support in educational hypermedia: The role of student knowledge level and the case for meta-adaptation. *British Journal of Educational Technology, 34*(4), 487-497.

Brusilovsky, P., & Vassileva, J. (2003). Course sequencing techniques for large-scale web-based education. *International Journal of Continuing Engineering Education and Lifelong Learning, 13*(1-2), 75-94.

Bunt, A., Conati, C., Huggett, M., & Muldner, K. (2001). On improving the effectiveness of open learning environments through tailored support for exploration. Retrieved August 18, 2005, from http://www.cs.ubc.ca/~conati/my-papers/aied2001.pdf

Carrier, C. A., & Williams, W. D. (1988). A test of one learner control strategy with students of differing levels of task persistence. *American Educational Research Journal, 25*(2), 285-306.

Civil Society Institute. (2005, August). *NCLB left behind: Understanding the growing grassroots rebellion against a controversial law.* Retrieved August 19, 2005, from the NCLBGrassroots.org Web site: http://reliableanswers.com/hs/nclb_open_revolt.asp

Cohen, V. B. (1985). A reexamination of feedback in computer-based instruction: Implications for instructional design. *Educational Technology, 25*(1), 33-37.

Conlan, O., & Wade, V. (2004). Evaluation of APeLS – An adaptive eLearning service based on the multi-model, metadata-driven approach. In P. De Bra & W. Nejdl (Eds.), *Proceedings of Third International Conference on Adaptive Hypermedia and Adaptive Web-based Systems* (pp. 291-295). Eindhoven, Netherlands: Springer LNCS.

Corbett, A. T., & Anderson, J. R. (1989). Feedback timing and student control in the Lisp Intelligent Tutoring System. In D. Bierman, J. Brueker, & J. Sandberg (Eds.), *Proceedings of the Fourth International Conference on Artificial Intelligence and Education* (pp. 64-72). Springfield, VA: IOS.

Corno, L., & Snow, R. E. (1986) Adapting teaching to individual differences among learners. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 605-699). New York: Macmillan.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Cuban, L. (2004). The open classroom. *Education Next*. Retrieved August 28, 2005, from the Hoover Institute Web site: http://www.educationnext.org/20042/68.html

Darling-Hammond, L. (1997). *The right to learn: A blueprint for creating schools that work*. San Francisco: Jossey-Bass.

Darling-Hammond, L., Griffin, G., & Wise, A. (1992). *Excellence in teacher education: Helping teachers develop learner-centered schools.* Washington, DC: National Education Association.

De Bra, P., Aerts, A., Berden, B., De Lange, B., Rousseau, B., Santic, T., et al. (2003, August). AHA! The Adaptive Hypermedia Architecture. *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia* (pp. 81-84). Nottingham, UK.

Deci, E. L., & Ryan, R. (1985). *Instrinsic motivation and self-determination in human behavior.* New York: Plenum.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *The Educational Psychologist*, *74*, 852-859.

de Jong, T., van Joolingen, W., Scott, D., deHoog, R., Lapied, L., & Valent, R. (1994). SMILSLE: System for multimedia integrated simulation learning environments. In T. de Jong & L. Sarti (Eds.), *Design and production of multimedia and simulation based learning material*. Dordrecht, Netherlands: Kluwer.

Dewey, J. (1916). *Democracy and education*. New York: Macmillan. Retrieved August 22, 2005, from http://www.ilt.columbia.edu/publications/dewey.html

Dewey, J. (1938). *Experience and education*. New York: Simon & Schuster.

Engelmann, S., Becker, W.C., Carnine, D., & Gersten, R. (1988). The direct instruction follow through model: Design and outcomes. *Education and Treatment of Children*, *11*, 303-317.

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., et al. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record, 52,* 187-201.

Fife, J., Shute, V. J., Underwood, J. S., & Graf, E. A. (2005). Proficiency model for Mathematics Intervention Model (MIM). Unpublished manuscript and model. Educational Testing Service, Princeton, NJ.

Flanagan, F. M. (Speaker). (1994, May 9). The great educators, first series. *John Dewey.* [Radio broadcast]. Dublin, Ireland: RTE, Radio 1. Retrieved August 18, 2005, from http://www.ul.ie/~philos/vol1/dewey.html

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *MMSS: Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hart, P. D., & Winston, D. (2005). *Ready for the real world? Americans speak on high school reform*. Executive summary. Retrieved August 30, 2005, from http://ftp.ets.org/pub/corp/2005execsum.pdf

Haycock, K. (2001). Closing the achievement gap. *Educational Leadership, 58*(6), 6-11.

Heffernan, N. T., & Koedinger, K. R. (2002). An intelligent tutoring system incorporating a model of an experienced human tutor. *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems*, Biarritz, France.

Hirsch, E. D. (1996). *The schools we need and why we dont have them*. New York, NY: Doubleday.

Hunt, E., & Minstrell, J. (1996). Effective instruction in science and mathematics: Psychological principles and social constraints. *Issues in Education*, *2*(2), 123-162.

Kahl, S. (2003, July). Implementing NCLB assessments and accountability requirements into an imperfect world. Paper presented at the Tenth Annual Education Law Conference, Portland, ME. Retrieved April 20, 2007, from http://www.measuredprogress.org/resources/education/nclb.html

Katz, I., Lipps, A., & Trafton, J. (2002). Factors affecting difficulty in the generating examples item type (GRE Board Report No. 97-18P). Princeton, NJ: Educational Testing Service.

Keller, J. M. (1979). Motivation and instructional design: A theoretical perspective. *Journal of Instructional Development, 2*, 26-34.

Keller, J. M., & Kopp, T. W. (1987). An application of the ARCS model of motivational design. In C. M. Reigeluth (Ed.), *Instructional theories in action: Lessons illustrating selected theories and models* (pp. 289-320). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kliebard, H. (1987). *The struggle for the American curriculum, 1893-1958*. New York: Routledge and Kegan Paul.

Koedinger, K. R., & Anderson, J. R. (1998). Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. *Interactive Learning Environments*, *5*, 161-180.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

Koedinger, K. R., & Tabachneck, H. J. M. (1994, April). *Two strategies are better than one: Multiple strategy use in word problem solving.* Presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kuntz, D., Fife, J., Shute, V. J., Graf, E. A., Supernavage, M., Marquez, E., et al. (2005). MIM: Mathematics Intervention Module 1. Unpublished computer program/prototype. Educational Testing Service, Princeton, NJ.

Landgraf, K. (2001, March). *Using assessments and accountability to raise student achievement*. Testimony presented at the Education Reform Subcommittee of the House Committee on Education and the Workforce on Measuring Success, Washington, DC.

Laurillard, D. (1984). Interactive video and the control of learning. *Educational Technology, 23*, 7-15.

Laurillard, D. (1991). Computers and the emancipation of students: Giving control to the learners. In O. Boyd-Barrett & E. Scanlon (Eds.), *Computers and learning* (pp. 64-80). Wokingham, UK: Addison-Wesley.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31*, 3-12.

Lemann, N. (1999, March 9). The IQ meritocracy: Our test-obsessed society has Binet and Terman to thank—or to blame. *Time, 153,* 115-116. Retrieved August 15, 2005, from http://www.time.com/time/time100/scientist/other/iq.html

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., et al. (2004). *International outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. Perspective* (NCES Publication No. 2005-003). Retrieved July 28, 2005, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005003

Lepper, M. R., & Chabay, R. W. (1985). Intrinsic motivation and instruction: Conflicting views on the role of motivational processes in computer-based education. *Educational Psychologist, 20*(4), 417-430.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mac Iver, M. A., & Kemper, E. (2002). Research on direct instruction in reading [Guest editors' introduction]. *Journal of Education for Students Placed at Risk, 7*(2), 107-116. Retrieved August 23, 2005, from http://www.louisville.edu/edu/jespar/vol_7_no_2_editors.htm

Marquez, E. (2003). *Algebra guide four: Exploring relationships between symbolic expressions and graphs of lines* [Pathwise series: Teacher assistance package]. Princeton, NJ: Educational Testing Service.

Mathan, S. A., & Koedinger, K. R. (2002). An empirical assessment of comprehension fostering features in an intelligent tutoring system. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems,* Lecture notes in computer science, Vol. 2363 (pp. 330-343). London: Springer-Verlag.

McCalla, G. I., & Greer, J. E. (1994). Granularity-based reasoning and belief revision in student models. In J. E. Greer & G. I. McCalla (Eds.), *Student modelling: The key to individualized knowledge-based instruction,* NATO ASI Series F: Computer and Systems Sciences, Vol. 125 (pp. 39-62). New York: Springer-Verlag.

McCombs, B., & Whistler, J. S. (1997). *The learner-centered classroom and school: Strategies for increasing student motivation and achievement.* San Francisco: Jossey-Bass.

McCoy, L. P. (2003, March 28). It's a hard job: A study of novice teachers' perspectives on why teachers leave the profession. *Current Issues in Education*, *6*(7). Retrieved August 23, 2005, from http://cie.ed.asu.edu/volume6/number7/

McMillan, J. H. (2007). *Classroom assessment: Principles and practice for effective instruction* (2nd ed.). Boston : Allyn & Bacon.

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 110-128). Kiel, Germany: IPN.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications for professional, instructional, and everyday science*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K.B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 437-446). San Francisco: Morgan Kaufmann.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2000, January). Evidence-centered assessment design: A submission for the NCME Award for Technical or Scientific Contributions to the Field of Educational Measurement. Retrieved November 12, 2004, from http://www.ncme.org/about/awards/mislevy.html

Mislevy, R. J., Steinberg, L.S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective, 1*(1) 3-62.

Mitrovic, A., & Martin, B. (2002). Evaluating the effects of open student models on learning. In P. de Bra, P. Brusilovsky, & R. Conejo (Eds.), *Proceedings of the Second International Conference on Adaptive Hhypermedia and Adaptive Web-based Systems* (pp. 296-305). Malaga, Spain: Springer-Verlag LCNS 2347.

Mora, J. K. (2001a, April 4). Language, literacy and content learning: Being accountable FOR and accountable TO English language learners. Keynote address at the Tenth Annual Administrators Conference, Sonoma County Office of Education, Santa Rosa, CA. Retrieved August 13, 2005, from http://coe.sdsu.edu/people/jmora/Prop227/accountabiliySCOE.htm

Mora, J. K. (2001b). Effective instructional practices and assessment for literacy and biliteracy development. In S. R. Hurley & J. V. Tinajero (Eds.), *Literacy assessment of second language learners* (pp. 149-166). Boston: Allyn & Bacon.

Mory, E. H. (2004). Feedback research review. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745-783). Mahwah, NJ: Lawrence Erlbaum Associates.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegemann, D. Leutner, & R. Brünken (Eds.), *Instructional design for multimedia learning* (pp. 181-195). New York: Waxmann.

National Commission of Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Research Council. (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment.* J. A. Koenig & L. F. Bachman (Eds.), Committee on Participation of English Language Learners and Students with Disabilities in NAEP and Other Large-Scale Assessments. Washington, DC: National Academies Press.

Ng, M. M., Guthrie, J. T., Van Meter, P., McCann, A., & Alao, S. (1998). How do classroom characteristics influence intrinsic motivation for literacy? *Reading Psychology*, *19*, 319-398.

Paige, R. (2001, October). Math summit speech excerpt. [U.S. Secretary of Education, keynote speaker at the Summit on Math Education]. Retrieved August 30, 2005, from http://www.publishers.org/school/mathsummit.cfm

Papanikolaou K. A., Grigoriadou M., Magoulas G. D., & Kornilakis, H. (2002). Towards new forms of knowledge communication: The adaptive dimension of a Web-based learning environment. *Computers and Education, 39* (4), 333-360.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas.* New York: Basic Books.

Pea, R. D. (1994). Seeing what we build together: Distributed multimedia learning environments for transformative communications. *Journal of the Learning Sciences, 3*(3), 283-298.

Pellegrino, J., Chodowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press.

Pellegrino, J. W. (2004). *The evolution of educational assessment: Considering the past and imagining the future* (Angoff Lecture No. 6, Policy Information Center Research Report). Princeton, NJ: Educational Testing Service.

Phi Beta Kappa. (2005). The 37th annual Phi Delta Kappa/Gallup Poll of the public's attitudes toward the publics schools. Retrieved September 20, 2005, from http://www.pdkintl.org/kappan/k0509pol.htm#exec

Proper, E. C., & St. Pierre, R. G. (1980). *A search for potential new Follow Through approaches: Executive summary*. Cambridge, MA: Abt Associates. (ERIC Document Reproduction Services No. ED 187 809).

Ravitch, D. (2000). *Left back: A century of failed school reforms*. New York: Simon & Schuster.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., et al. (2005). The ASSISTment project: Blending assessment and assisting. *Proceedings of the Twelfth Annual Conference on Artificial Intelligence in Education* (pp. 555-562). Amsterdam, Netherlands.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. *The Journal of the Learning Sciences, 3,* 265–283.

Scheirer, M. A., & Kraut, R. E. (1979). Increasing educational achievement via self-concept change. *Review of Educational Research, 49*, 131-150.

Schimmel, B. (1988). Providing meaningful feedback in courseware. In D. Jonassen (Ed.), *Instructional designs for microcomputer courseware* (pp. 183-195). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shafrir, U. (1999). Representational competence. In I. E. Sigel (Ed.), *The development of mental representation: Theory and applications* (pp. 371-389). Mahwah, NJ: Lawrence Erlbaum Associates.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shute, V. J. (1993). A comparison of learning environments: All that glitters... In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 47-74). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shute, V. J. (1995). SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction, 5*, 1-44.

Shute, V. J., Gawlick, L. A., & Gluck, K. A. (1998). The effects of practice and learner control on short- and long-term gain and efficiency. *Human Factors, 40*, (2), 296-310.

Shute, V. J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 279-326). New York: W. H. Freeman.

Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for sighted and visually-disabled students. In L. Pytlikzillig, R. Bruning, & M. Bodvarsson (Eds.). *Technology-based education: Bringing researchers and practitioners together* (pp. 169-202). Greenwich, CT: Information Age Publishing.

Shute, V. J., Lajoie, S. P., & Gluck, K. A. (2000). Individualized and group approaches to training. In S. Tobias & J. D. Fletcher (Eds.), *Training and retraining: A handbook for business, industry, government, and the military* (pp. 171-207). New York: Macmillan.

Shute, V. J., & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist*, *38*(2), 105-114.

Sleeman, D. H., Kelly, A. E., Martinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive Science, 13*(4), 551-568.

Smith, P. L., & Ragan, T. J. (1999). *Instructional design* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist, 27*(1), 5-32.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. (1977). *Education as experimentation: A planned variation model: Vol. 4.-A. An evaluation of Follow Through*. Cambridge, MA: Abt Associates.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, *83*(10), 758-765.

Stone, J. E., & Clements, A. (1998). Research and innovation: Let the buyer beware. In R. Spillan & P. Regnier (Eds.), *The superintendent of the future* (pp. 59-97). Gaithersburg, MD: Aspen Publishers.

Tabachneck, H. J. M., Koedinger, K. R., & Nathan, M. J. (1994). Toward a theoretical account of strategy use and sense-making in mathematics problem solving. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Terman, L. M. (1916). *The measurement of intelligence*. Cambridge, MA: Riverside Press.

U.S. Department of Education. (2005). *Ten Facts about K-12 education funding*. Retrieved August 25, 2005, from http://www.ed.gov/about/overview/fed/10facts/index.html

Vassileva, J., & Wasson, B. (1996). Instructional planning approaches: From tutoring towards free learning. *Proceedings of EuroAIED '96* (pp. 1-8), Lisbon, Portugal.

Weber, G., & Brusilovsky, P. (2001). ELM-ART: An adaptive versatile system for Web-based instruction. *International Journal of Artificial Intelligence in Education, 12*(4), 351-384.

Zapata-Rivera, J. D. (2003). *Learning environments based on inspectable student models*. Unpublished doctoral dissertation, University of Saskatchewan. Department of Computer Science, Saskatchewan, Canada.

Zapata-Rivera, J. D., & Greer, J. (2004). Interacting with Bayesian student models. *International Journal of Artificial Intelligence in Education*, *14*(2), 127-163.

Appendix: Definitions of Different Types of Assessments

Assessment can be conducted at various times throughout the school year or instructional program. Moreover, the format and purpose of the assessment can differ. Following are definitions of different assessments, as used in the context of this chapter.

- *Formative Assessment*. Formative assessment is usually done at the beginning or during a program, providing the opportunity for immediate evidence for student learning in a particular course or at a particular point in a program. The purpose of formative assessment is to improve quality of student learning and should not be evaluative or involve grading students.
- *Summative Assessment*. Summative assessment is comprehensive, provides accountability, and is used to check the level of learning at the end of the program. Program goals and objectives often reflect the cumulative nature of the learning that takes place (or should occur) in a program. Summative assessment is conducted at the end of the program to ensure students have met the program goals and objectives.
- *Diagnostic Assessment*. Although some educators view diagnostic assessment as a component of formative assessment, most consider it a distinct form of measurement (e.g., McMillan, 2001). In practice, the purpose of diagnostic assessment is to determine, prior to instruction or during the course of learning, each student's strengths, weaknesses, knowledge, and skills. Determining this information allows the teacher to remediate students and adjust the curriculum to meet each student's specific needs.
- *Criterion-Referenced Testing*. Criterion-referenced testing (CRT) is based on a well-specified domain with items appropriately sampled, and with the intention of making an inference about the degree of mastery a student attains in relation to the domain. Scores on criterion-referenced tests indicate what individuals *can* do—not how they have scored in relation to the scores of particular groups of persons, as with norm-referenced tests.
- *Norm-Referenced Testing*. Norm-referenced testing (NRT) compares a person's score against the scores of a group of people who have already taken the same exam, called the "norming group." Scores are usually reported as percentile ranks. Most achievement NRTs are multiple-choice tests, although some also include open-ended, short-answer questions. The questions on these tests mainly reflect the content of nationally-used textbooks, not the local curriculum. NRTs are designed to "rank-order" test takers to compare student scores.

Endnotes

1. The expression "sea change" in general refers to a profound change in the nature of something. The phrase appears to have originated in Shakespeare's *The Tempest* (I,ii).
2. See the Appendix for definitions of various types of assessments as used in the context of this chapter.
3. Addressing the basis for this ideological war over the best ways to teach, Cuban (2004) provides this interesting perspective, that the enduring quarrels are "proxies for deeper political divisions between conservatives and liberals on issues ranging from environmental protection to foreign policy. There are, of course, liberals who believe in traditional education and conservatives who embrace progressive ideas, but the lines are fairly well drawn" (p. 71).
4. The nine models are (1) Direct Instruction Model (University of Oregon); (2) Behavior Analysis Model (University of Kansas); (3) Language Development (Bilingual) Model (Southwest Educational Developmental Laboratory); (4) Cognitively-Oriented Curriculum (High Scope Foundation); (5) Florida Parent Education Model (University of Florida); (6) Tucson Early Education Model (University of Arizona); (7) Bank Street College Model (Bank Street College of Education); (8) Open Education Model (Education Development Center); and (9) Responsive Education Model (Far West Laboratory).
5. Briefly, "Direct Instruction" refers to a highly structured instructional approach, designed to accelerate the learning of at-risk students. Curriculum materials and instructional sequences attempt to move students to mastery at the fastest possible pace. Teachers follow scripts and the focus is on basic skills.
6. Some currently-popular terms related to progressive education have been summarized by Hirsch (1996), including lifelong learning, developmentally appropriate instruction, situated learning, cooperative/collaborative learning, multiple intelligences, discovery learning, portfolio assessment, constructivism, hands-on/experiential learning, project method, integrated curriculum, higher-order thinking/learning, and authentic assessment.
7. Diagnoses in this context refer to accurate analyses (measurement and reporting) of what the student knows and does not know, and to what degree.
8. While many ideas in constructivism come from cognitive psychology, it also embodies ideas from developmental psychology and anthropology.
9. ECD adheres to the guidelines for assessment design established by within a recent report by the National Research Council, "Knowing What Students Know" (Pellegrino, Chudowsky, and Gluser, 2001),

which identify three key, interconnected elements for assessments: (1) *cognition*: a theory of what students know and how they develop competence in a subject domain; (2) *observation*: tasks or situations used to collect evidence about student performance; and (3) *interpretation*: a method for drawing inferences from those observations.

10. "Counterpart" refers to similar students, based on age and grade, who reside in different countries. For more on international standing and comparisons among NAEP, TIMSS, and PISA results, see http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf. The main difference between the two international analyses (TIMSS and PISA) is that TIMSS is the U.S. source for internationally comparative information on mathematics and science achievement in the primary and middle grades, while PISA is the U.S. source for internationally comparative information on the mathematical and scientific literacy of students in the upper grades at an age that, for most countries, is near the end of compulsory schooling.
11. NAEP scores range from 0–500, and are divided into four categories: Below Basic (0–261), Basic (262–298), Proficient (299–332), and Advanced (333–500).
12. In this context, grain size refers to the scope or generality of a proficiency. For instance, a large grain size, and hence general proficiency would correspond to, say, the course level (e.g., Algebra I concepts and skills). A small grain size, thus more specific proficiency may be a particular skill (e.g., can calculate slope from points). Between these extremes are additional levels of aggregation and generality. For more on the topic, see McCalla and Greer (1994).
13. "Student model" refers to a proficiency model that has been instantiated with information (estimations of mastery) in relation to a particular student.
14. This comes from an internal ETS effort to map alignments among state standards, and while it is not a specific state standard, it aligns well with actual state standards, such as Nevada: Translate among verbal descriptions, graphic, tabular, and algebraic representations of mathematical situations; West Virginia: Translate word phrases into algebraic expressions or word sentences into equations and inequalities; and Texas: Translates among and uses algebraic, tabular, graphical, or verbal descriptions of linear functions.
15. An example of a question requiring a textual response is, "Explain in words how you know that…."
16. The common errors, per item, were identified after we reviewed answers to about 500 paper-and-pencil tests covering all eight proficiencies, in each of the four variants (graph, numeric, expression/equation, and text), and with two difficulty levels (easy and hard).

After the tests were scored, incorrect responses, per item, were examined and tallied in a spreadsheet. The more frequent errors were further analyzed to infer misconceptions or procedural bugs underlying them.