Developing Assessments to Inform Teaching and Learning

CSE Report 628

Kristin M. Bass and Robert Glaser
CRESST/Learning Research and Development Center,
University of Pittsburgh

May 2004

Center for the Study of Evaluation (CSE)
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education and Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.2 Classroom and Teachers' Assessment, Strand 1: Descriptors for Assessing Classroom Progress, Robert Glaser, Project Director, Learning Research and Development Center, University of Pittsburgh.

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education. Additional Support by James S. McDonnell Foundation, 1034 South Brentwood Boulevard, Suite 1610, St. Louis, MO 63117.

DEVELOPING ASSESSMENTS TO INFORM TEACHING AND LEARNING

Kristin M. Bass and Robert Glaser

CRESST/Learning Research Development Center,

University of Pittsburgh

Abstract

The centrality of assessment for facilitating thinking, reasoning, and problem solving is well-documented and indisputable. Less apparent is how to create informative, yet practical measures for classroom use. Clearly, the changing of assessments alone will not in and of itself improve learning; teachers' beliefs and practices will need to be altered with various levels of support. The design of assessment situations can nevertheless have a substantial impact on the quality of information provided to teachers and students for instructional decision-making and meaningful learning.

This report considers principles of informative assessments that improve teaching and learning by communicating learning goals, interpreting student performance, tracking progress over time, and suggesting appropriate corrective actions. In the report, we describe several properties of assessment design that enable teachers and students to describe progress in terms of cognitive features of performance, and then act on that information to improve learning. We review classroom assessment programs across subject matters and grade levels in order to suggest essential design elements for tasks, score forms, and interpretive materials that maximize the information provided by assessment of performance and competence. These principles are not intended to be comprehensive, but are meant to highlight some promising areas for informative assessment research.

Introduction

The centrality of assessment for facilitating thinking, reasoning, and problem solving is well-documented and indisputable (Black & Wiliam, 1998; Glaser & Silver, 1994; National Research Council [NRC], 2001; Shepard, 2000a). Less apparent is how to create informative, yet practical measures for classroom use. Clearly, the changing of assessments alone will not in and of itself improve learning; teachers' beliefs and practices will need to be altered with various levels of support (Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Shepard, 2000b). The design of assessment situations can nevertheless have a substantial impact on the quality of information provided to teachers and students for instructional decision-making and meaningful learning.

This report considers principles of informative assessments that improve teaching and learning by communicating learning goals, interpreting student performance, tracking progress over time, and suggesting appropriate corrective actions. Presently, the drive to identify and develop such tools is more compelling than ever before. Demands for standards and accountability (No Child Left Behind Act of 2001) have brought assessment to the forefront of educational policy concerns. Renewed interest in uniting the fields of cognitive psychology and psychometrics has also generated conversation and commitment for the improvement of national, state, and classroom testing (Pellegrino, Baxter, & Glaser, 1999; NRC, 2001). Together, these shifts in political and intellectual climates have created unprecedented opportunities for the exploration of informative assessment techniques.

Our approach is grounded on several assumptions. We presume that informative assessment, like formative measures, can provide "a clear view of the learning goals, information about the present state of the learner, and action to close the gap" (NRC, 2001, p. 229). The term "formative," however, carries a chronological connotation and emphasizes the placement of assessments during the course of an instructional unit (Black & Wiliam, 1998; Bloom, Hastings, & Madaus, 1971). We use the label "informative" to draw attention to the instructional purpose for improving student learning (New South Wales Department of Education and Training (NSWDET), 1998). Further, assessments can be informative of various aspects of achievement for different audiences. All assessments are informative in some way; the key issues are *of what, for whom,* and *how* those measures inform. We concentrate on the information given to teachers and students to facilitate the teaching and learning of thinking, reasoning, and problem solving as advocated by various sets of national standards (e.g., National Council of Teachers of Mathematics [NCTM], 1995; NRC, 1996).

We also adhere to the general framework for assessment design put forth by the Committee on the Foundations of Assessment (NRC, 2001). The committee identifies the three key elements of assessment as: (a) *cognition,* or theories about learning, performance, and targets for assessment; (b) *observation,* or tasks used to elicit information about learning; and (c) *interpretation,* or methods for scoring and validating assessment results. We examine how these elements are presented to teachers and students in clear, coherent, and instructionally meaningful ways.

The sections to follow describe several properties of assessment design that enable teachers and students to describe progress in terms of cognitive features of performance, and then act on that information to improve learning. We review classroom assessment programs across subject matters and grade levels in order to suggest essential design elements for tasks, score forms, and interpretive materials that maximize the information provided by assessment of performance and competence. These principles are not intended to be comprehensive, but are meant to highlight some promising areas for informative assessment research.

Properties of Assessment to Inform Teachers

A key aspect of teaching has always been monitoring students' progress. Teachers traditionally do this by giving curriculum-based classroom tests and judging the number of correct responses. Unfortunately, this usual approach to assessment often does not provide the information that teachers could use in order to improve student proficiency. In this report, we provide examples and instances of approaches to assessment that can effectively elicit and display information about student achievement (i.e., not simply their knowledge of subject matter, but their ability to use that knowledge to solve problems and reason about novel situations). We anticipate changes in classrooms of the future will occur as assessments of thinking, reasoning, and problem solving are integrated with instruction to inform teaching and learning.

The unique, fundamental nature of informative assessment is its ability to prepare teachers for effective instructional activity based on detailed knowledge of student accomplishment. The specification of cumulative objectives is a powerful component of good educational programs and, in this context, informative assessment supports the teachers' pedagogical skill and judgment. Taking full advantage of informative assessment requires a particular use of adaptive teaching techniques; in other words, information about the student's learning process and the nature of accomplishment sets teachers up for appropriate instructional modifications. To do so, assessment must effectively communicate instructional and curriculum changes that enable teachers to see the student's displayed performance as connected to details of classroom practice.

We discuss four properties of informative assessment for teachers that display cognitive components of performance and suggest instructional interventions. First, we suggest that a *cognitive language* is essential for helping teachers interpret assessment results and promote complex problem solving. Second, we demonstrate how *rubrics* can distinguish and communicate various qualities of thinking and reasoning. Third, we explore the importance of *explicit relationships between assessments and instruction* that reinforce consistent expectations for teaching and learning. Finally, we discuss how assessments can suggest *guidelines for instruction.*

A Cognitive Language for Teachers

During informative assessment, descriptions of learning goals and performances are couched in a language compatible with teachers' practices, experiences, and beliefs. Teachers tend to communicate to one another in a "language of the particular" (Leinhardt, 1990), or of how to teach under specific circumstances. At the same time, there is a growing body of cognitive theory and its application that can inform instructional situations (e.g., Carver & Klahr, 2001; McGilly, 1994). Making this knowledge accessible to teachers is essential to the creation and dissemination of informative assessments. What is necessary is a language of learning and cognition for teacher practice that steers away from "molecular" learning theory, and toward a level of discourse that relates learning processes to instructional performances.

Consider Cognitively Guided Instruction (CGI), a program that establishes a cognitive language through frameworks for the assessment of elementary grades' mathematical problem-solving (Carpenter, Fennema, & Franke, 1996; Carpenter, Fennema, Franke, Levi, & Empson, 1999). Their language consists of a taxonomy of common problem types and solution strategies. The researchers observed that across computational problems (i.e., addition/subtraction and multiplication/division), numerical literacy progresses from rigid, exclusive dependence on manipulatives (e.g., base-10 blocks) to more flexible, abstract mental representations of numbers. Moreover, students use different strategies based on two features of the question: (a) the operation to be performed between the two numbers (e.g., joining, separating), and (b) the location of the unknown value within the problem (e.g., 2 + 3 = ? vs. 2 + ? = 5). The interaction of these two features sets up expectations for typical student behaviors and alternative problem-solving strategies. When teachers are able to classify problem types and solution strategies, they can encourage students to consider various ways to approach questions.

CGI illustrates several key features of an informative, cognitive language for educators. Terminology (e.g., problem types and strategies) is derived from observable student practices and illustrated with multiple classroom examples (Carpenter et al., 1999). Further, the frameworks are extremely flexible in that they are not linked to specific curricula or prescribed teaching protocols, but instead give teachers leeway to select problems to assess understanding. Teachers can then modify instruction through the use of progressively simpler or more difficult problems, and the discussion and modeling of alternative strategies (Carpenter et al., 1996, 1999). In this sense, CGI frameworks provide a language that is highly adaptable; teachers are given tools for identifying student performance without constraining instructional practices.[1]

In other programs, metaphors provide a basis for dialogue between teachers and instructional designers (e.g., Martínez, Sauleda, & Huber, 2002). The Classroom Assessment as the Basis for Teacher Change (CATCH) project is dedicated to helping teachers improve their instructional practices by way of innovative, informative assessments (CATCH, 2002). Its keystone is a pyramid metaphor of problem types (Verhage & de Lange, 1997). In this scheme, tasks are fit into a three-tiered, three-dimensional pyramid. The rows, or tiers, represent increasingly complex levels of thinking: (a) simple rote reproduction of facts and algorithms; (b) connections across mathematical problems or domains (e.g., generating combinations of numbers which, when added or subtracted, equal five); and (c) analyses of situations that require novel, self-generated models or solution strategies. The vertical dimension of the pyramid refers to the various domains of mathematics, such as geometry, algebra, and probability. Finally, the depth dimension describes relative problem difficulty.

The pyramid metaphor communicates selected principles of assessment implementation. In particular, its shape (wide at the bottom, narrow at the top)

---

[1] The reader is encouraged to consult CGI's handbook (Carpenter et al., 1999) for extended examples of its principles in classroom practice.

illustrates the relative number of problems that should appear on a balanced test of mathematics skills, with higher level tasks at the narrow top of the pyramid taking more time to complete and appearing less frequently than rote recall questions. Ideally, teachers should strive to fill the pyramid, eventually testing their students in all content dimensions at all problem levels (Dekker, Querelle, & van den Boer, 2000). Missing from this metaphor is a model of knowledge and skill development, a hallmark of CGI. It might be assumed that skills build upon one another and that students must possess a firm foundation in basic algorithms before proceeding to more complex problems, but this is not confirmed in project literature. Nevertheless, the CATCH project and assessment pyramid demonstrate the potential effectiveness of metaphors and concrete imagery for representing complex mathematic concepts for assessment and instruction.

Rubrics Emphasizing Cognitive Elements of Performance

Performance rubrics are convenient, readily available tools for characterizing student learning (Arter & McTighe, 2001; Luft, 1999; Shafer, Swanson, Bené, & Newberry, 2001). Rubrics can be defined as systems for rating the quality of a particular assessment performance. Essential to a rubric is the notion of levels that can distinguish different quality performances (Arter & McTighe, 2001). When properly designed, they can distinguish and communicate various qualities of thinking, reasoning, and problem solving. At issue is specifying the parameters or principles for exemplary, informative classroom rubrics.

At their best, rubrics can make students' thinking explicit and highlight areas for growth. Balanced Assessment for the Mathematics Curriculum (1999) ranks tasks on a four level scale: "the student needs significant instruction," "the student needs some instruction," "the student's work needs to be revised," and "the student's work meets the essential demands of the task." Each task is accompanied with specific objectives, performance descriptions, and examples of each level. Teachers can then use this information to monitor students' performance under a variety of formal and informal assessment conditions (Balanced Assessment, 1999).

In one of the elementary grades' tasks, for instance, students are asked to determine the ages of three dogs based on their combined ages and the ages of two of the dogs relative to the third (e.g., "Jason is 5 years older than Boy Blue," (Balanced Assessment, 1999, p. 166). Scores reflect the correctness of the response and attention to the problem details (i.e., difference between ages, combined ages of the dogs). Students who need significant instruction failed to consider the relative ages of the dogs or their combined ages, while students who need to revise their work obtained the correct answer but did not explain how they did so. The rubrics point out specific errors in students' logical reasoning and in so doing, articulate criteria for improvement.

Formulating effective rubrics demands an awareness of cognitive theory as realized in student performance (e.g., the developmental progression of students' ideas of a particular concept or subject matter), combined with attention to context and detail. Among suggested guidelines for rubric development is that scores must emphasize quality of performance over quantity of information provided. In doing so, rubrics can challenge teachers' beliefs that performance can be defined exclusively by the number of details students include (Goldberg & Roswell, 1999). Tradeoffs must also be made between analytic and holistic rubrics (i.e., rubrics that analyze multiple aspects of performance versus overall quality), and generalized versus specific guidelines (i.e., using the same rubric for multiple assessment situations versus using a different rubric each time). While analytic and holistic rubrics are similar in technical qualities such as reliability (Arter & McTighe, 2001; Klein et al., 1998), analytic rubrics take longer to score yet are also judged more informative by teachers (Waltman, Kahn, & Koency, 1998). Likewise, generalized rubrics may be easier to learn because of repeated practice opportunities, but may leave valuable information that only specifically attuned guidelines can provide.

Future consideration will need to be given to operationalizing the alignment of instructional situations and rubric complexity in order to optimize rubric use. If the teacher's goal is to obtain an overall impression of the class, for example, a holistic rubric may suffice. If, instead, that aim is the diagnosis of students' particular strengths and weaknesses, an analytic rubric may be more appropriate (Waltman et al., 1998). Conversely, attention to the individual details of a response may be misleading if the number of details is less important than the overall quality of information included (Arter & McTighe, 2001; Klein et al., 1998). Ultimately, discussions of rubrics' informational value are incomplete without concomitant analysis of the instructional context.

Explicit, Consistent Relationships Between Assessments and Instruction

Informative assessments are administered in a carefully planned sequence detailing students' past, present, and future goals and achievements. Assessments are not conducted in isolation, but instead build upon each other toward an overarching purpose. Synchrony between assessments and instruction reinforces consistent expectations for teaching and learning. These relationships must be made explicit to teachers, however, if they are to influence classroom learning. Several examples of how these linkages are established and made apparent to teachers are reported here.

KIDMAP software (NSWDET, 2001) demonstrates how technology can help teachers visualize relationships between assessments, and then craft individual learning trajectories for their students. The software functions as a comprehensive database containing national standards, performance rubrics, sample assessments and lesson plans, and comparison data from students across the country. This enables teachers to plan their instruction around predetermined standards by charting objectives, instruction, assessment, and available resources. The compact display of information facilitates its review and draws teachers' attention to lessons where they are not presently collecting a broad range of assessment evidence (NSWDET, 2001). On an individual student level, KIDMAP helps teachers record performance on specific syllabus outcomes or national indicators, and then create profiles over time and in comparison to other students. This in turn allows teachers to easily identify areas for individual and class improvement, with sample lesson plans and assessments providing options for follow-up activities (NSWDET, 2001).

Assessments linked within units can also be beneficial. One approach exploits the potential of curriculum-embedded assessments to assess and develop skills needed for a summative, end-of-unit task. Consider Mystery Powders (Baxter, Elder, & Glaser, 1995; Baxter, Elder, & Shavelson, 1997; Baxter & Glaser, 1998), a series of assessments designed for a hands-on science unit of the same name in which elementary students study the reactions of five white powders to various indicators (e.g., water, vinegar, iodine, heat). In the end-of-unit assessment, students must identify the composition of six samples of white powder based on a list of five options. Students are given water, vinegar, iodine, and a hand lens, and may also consult their science laboratory notebooks for information on reactions between powders and indicators. To identify the powders, students must collect confirming and disconfirming evidence. For example, two of the powder options are cornstarch only, and cornstarch and baking soda. To identify the cornstarch only sample, students should test it with iodine (which turns purple in the presence of starch) to confirm that it contains cornstarch, but should *also* use vinegar (which fizzes only with baking soda) to disconfirm, or rule out, the presence of baking soda.

Data collected from the end-of-unit assessment revealed that students had difficulty gathering appropriate evidence to support their claims. Students rarely, if ever, used information to rule out powders. In response, two embedded assessments were developed to promote the skills students were lacking (Baxter et al., 1997). The tasks mirrored the general purpose and structure of the end-of-unit assessment in that they asked students to identify the composition of one or more white powder samples. They gave students practice in determining when they had gathered sufficient evidence to conclusively identify the samples, and alerted them to the need to rule out powders by a process of elimination. The associated score forms and score interpretation guidelines provided teachers with a means to understand the nature of evidence students used to identify the powders, and their ability to draw conclusions from their findings. This heightened awareness could form the basis of instructional interventions that reinforced the principle of adequate and definitive evidence.

The Mystery Powders tasks demonstrate an important property of informative assessment development: that formative, or curriculum-embedded measurements are created *after* establishing an end-of-unit assessment or outcome. A more typical development routine is to insert assessments at regular intervals during instruction. This process tends to test the material that has just been covered without any regard to how the assessment relates to the rest of the unit. A more beneficial approach, illustrated in the Mystery Powders assessments and others (e.g., Roberts, Wilson, & Draney, 1997; Wiggins & McTighe, 1998), is to work backward from the end-of-unit assessment or desired learning results, so that all activities are directed toward a central outcome. In the process, this integrated system of assessments consistently reinforces unit goals and fosters the development of skills necessary to acquire proficiency.

Interpretive Guidelines for Teaching

Our exploration of the properties of informative assessment for teachers concludes with a consideration of the ways in which task feedback summarizes student performance and facilitates instructional changes. These guidelines can take on a variety of forms. This section highlights a few programs representative of the array of informative assessment designs.

Facets of Thinking (Minstrell, 2000, 2001) generates "prescriptions" for teachers and students to improve learning. Its foundation is a series of hierarchically-organized clusters of ideas, or facets, that students may hold about a particular phenomena or subject. Facets are ranked in order of their expected development, with lower numbers representing accepted conventions in that particular field, and higher values indicative of a less principled understanding (Minstrell, 2000, 2001).

Students' facets are evaluated through, among other ways, DIAGNOSER, a web-based assessment (DIAGNOSER, 2002; Hunt & Minstrell, 1994). The program presents students with a series of multiple choice questions asking to predict observations and justify their answers. Teachers can review the distribution of answers and decide how to adapt instruction using "prescriptive activities" provided on DIAGNOSER's web site (DIAGNOSER, 2002). The ranking of the facets within a cluster (those ending in 0 or 1 being the most conceptually appropriate) allow teachers to further evaluate the level of students' understanding and the types of assumptions that must be corrected.

Consider the assessment of students' understanding of the effects of gravity and media (e.g., air) on objects. Teachers may elicit students' prior knowledge through a series of "elicitation questions" suggested on the DIAGNOSER website (DIAGNOSER, 2002). One such question asks students to predict the weight of an object in space given that it weighs five pounds on Earth. Teachers may choose to correct misunderstandings with a class demonstration of the reading of a scale in an airless environment and other related activities, accompanied by the assignment of a DIAGNOSER problem set (DIAGNOSER, 2002; Minstrell, 2000). The web site provides follow-up activities for students according facet codes for their assessment responses (DIAGNOSER, 2002). In time, students should be able to see that the principles they first applied to the elicitation questions do not hold up in reality.

The Mystery Powders assessments (Baxter et al., 1995, 1997; Baxter & Glaser, 1998), described in the previous section illustrate how an assessment system can orient teachers to unit goals and student performance in a way which shapes their teaching practices. Recall that the Mystery Powders assessment system consists of three performance assessments (two curriculum-embedded, one end-of-unit), and that all tasks are designed to draw awareness of students' generation and interpretation of necessary and sufficient evidence for identifying the composition of unknown white powders. Teachers are given five types of information to help them implement and score the assessments: (a) *a sequence of assessment and instruction*, which outlines the unit lessons and points at which the performance assessments should be administered; (b) *instructions* for each task, including objectives, materials needed, time limits, grouping arrangements, and a script for giving students directions; (c) *a score form*, with blank spaces for checking students' performance (i.e., identification of powders, and supporting evidence) against the correct answers; (d) *scoring instructions* with examples; and (e) *a score summary* with instructions for assessment interpretation. The latter two documents are particularly worthy of discussion for their informational value.

The scoring instructions begin by identifying the evidence used for scoring and the types of scores students will receive. In the Mystery Powders assessments, students receive separate scores for the identification of the powders and the evidence provided to support claims. Next follows a scoring rationale that explains that students need to gather appropriate combinations of confirming and disconfirming evidence to identify the powders. Teachers are taken step-by-step through the process of marking the score form and assigning points. Answers and evidence are scored on separate scales, with one point given for every correct combination of powders identified, and zero to four points given per powder for supplying complete and correct evidence. The score form includes two examples of evidence at each point level (i.e., 0 through 4) to clarify scoring criteria.

The score summary allows teachers to examine a random sample of ten students in more depth to identify specific areas for instructional change. The summary consists of a table where teachers list students' answer and evidence scores, along with the number of times they identified confirming, disconfirming, and irrelevant evidence. It also includes a series of questions to guide teachers' interpretation (e.g., "Look at Column III. How many students used confirming evidence?") This process allows teachers to break scores down into specific components to further direct their teaching efforts. For example, if most students used confirming evidence to "rule in" powders but failed to use disconfirming evidence to "rule out" alternative choices, the teacher should draw students' attention to the characteristics of necessary and sufficient powder identification evidence.

Unlike Facets, the Mystery Powders score form and summary do not provide specific activities for remedying performance discrepancies. Rather, instructional modifications are left to the teachers' discretion, and may be guided by district professional development opportunities for integrating assessment with instruction

(Baxter et al., 1997). The scoring system, as it presently stands, nevertheless provides teachers with the kinds of information necessary to alter and expand their practices.

Properties of Assessment to Inform Students

Informative assessment for the learner considers how students can use assessment results to improve their performance in the course of instruction. Two such types of information for the student are the steps or strategies for a problem's solution, or the explanation of the principles underlying a correct response. The approach to informative assessment for learners that we consider here is one that makes learning goals and performance quality explicit so that learners can easily monitor and improve their thinking and reasoning.

This section reviews three design principles that make assessments informative to students, as a means of identifying some of the elements of effective programs. First, exemplary *models of competence* are described that familiarize students with the criteria for successful assessment performance. Second, the value of *graphical tools to track progress* is discussed with an emphasis on the principles underlying informative graphs. Third, the value of *structured opportunities for reflection and revision* is explored as a principle for maximizing the benefits of self-assessment.

Models of Competence

Practical, clear grade-level appropriate standards are essential for improving student performance (Arter & McTighe, 2001). While this idea is not new (see the discussions of assessment transparency in Frederiksen & Collins, 1989), it takes an especially prominent role in the context of informative assessment for students. The programs described here illustrate various forms of classroom goals or standards, and the roles of teachers and peers in establishing performance expectations.

One example of performance standards is ThinkerTools, which incorporates peer and self-assessment into inquiry-based science instruction (White, 1993; White & Frederiksen, 1998). As students conduct investigations, they learn to evaluate their performance according to several criteria, such as "Understanding the science," "Reasoning carefully," and "Writing and communicating well." Throughout the various phases of inquiry (e.g., question, prediction, experiment), students apply at least two such criteria to their work and others'. These reflective experiences, in turn, give students a language for discussing their thinking and reasoning during inquiry, leading to the identification and improvement of problem areas for learning.

Other programs incorporate more implicit teacher- and student-mediated assessment to model desired learning behavior. Fostering Communities of Learners, for example, capitalizes on group activity to elicit evidence of student thinking and strengthen comprehension and communication. In one such program, guided writing, students compose written and illustrated research reports with help from a more knowledgeable expert (e.g., teacher, researcher, or older student). As students work in small groups, the expert asks probing questions like "Do you think the reader will be able to understand that?" (Brown, Ellery, & Campione, 1998). Students' activities, the experts' assessment and feedback, and the students' ensuing revision are immediate and interactive, and directed towards improved writing proficiency.

In addition to expert adults who are capable of modeling learning, students are equally instrumental in shaping their classmates' understanding. This is evidenced by the role of reciprocal teaching (RT) (Palincsar & Brown, 1984) in FCL classrooms. Students work in small groups to understand information (e.g., article, video) pertinent to their research topic. One student leads the discussion by summarizing the material, asking questions about it, or prompting for predictions about future research. The others contribute by clarifying their understanding and correcting comprehension problems (Brown & Campione, 1996; Brown et al., 1998).

Reciprocal teaching exercises help students discern how well they understand material by virtue of their ability to ask and answer questions about it. Peers also provide feedback to one another by reacting to those questions and answers, and demonstrating how well they understood the material. In turn, the more competent students model comprehension for those at a lower level of understanding (Brown et al., 1998). Two related activities, jigsaw and crosstalk, have students gather and share research in order to teach their peers. Questions during sessions direct students to the information they must obtain in order to clarify their peers' understanding (Brown et al., 1998).

The effectiveness of models of competence depends on the quality of teacher and peer support, as demonstrated by the preceding examples. Studies of middle schoolers' creation of hypermedia documents (Erickson & Lehrer, 1998) further corroborate this assumption. Over the course of two years, students in an urban middle school learned to design multimedia presentations and, in so doing, generate expansive research questions (i.e., topics that encouraged inquiry) and successfully communicate ideas. Throughout instruction, students evaluated peers' work and internalized "critical standards" for questions and project designs. Teachers scaffolded these discussions by elaborating on student ideas, identifying examples of exemplary work, and articulating performance criteria. With time and practice, students took more control over class discussions, appropriated the teachers' guidelines into their own discourse, and developed a peer critique sheet for assessing their hypermedia documents (Erickson & Lehrer, 1998). In this case, as in the ThinkerTools and Fostering Communities of Learners projects, assessment was made informative to students through the introduction of standards and the diminishment of teacher reinforcement.

Graphical Tools to Track Progress

Informative assessments for students incorporate methods for visualizing progress over time. Like teachers, students also need concise, comprehensible records of past and present accomplishments in order to set future learning goals. Some assessment programs address this need through student-friendly graphical displays. Such efforts must be tempered by research on students' difficulties with interpreting graphs and other inscriptions, as will be noted shortly.

KIDMAP (NSWDET, 2001) and the Berkeley Evaluation and Assessment Research group (BEAR) (Roberts et al., 1997; Wilson, Draney, & Kennedy, 2001; Wilson & Sloane, 2000) graph students' achievement against subject-specific criteria and/or classroom norms for the purposes of reflection and review. These reports typically take the form of percentile rankings across students, or a chart or graph of performance on several variables rated on the same ordinal scale. The BEAR assessment system for a middle-school science curriculum, for example, rates students on a scale from zero to four on several inquiry variables such as designing and conducting investigations and understanding concepts. A student's performance is charted with respect to these variables accompanied with suggestions for improving performance (e.g., "Explain any unexpected results," "Think about other ways you could use the scientific information") (Wilson et al., 2001). KIDMAP likewise displays performance on a five-point scale (from "Progressing Towards" to "Working Beyond") on selected objectives, supplemented with more detailed comments. The BEAR and KIDMAP reports can then be used in conferences with the student and/or parents to discuss progress and set appropriate learning goals (NSWDET, 2001; Roberts et al., 1997; Wilson & Sloane, 2000).

Graphing student progress may well prove fruitful for optimizing assessment informativeness. The promise of graphing, however, must be tempered with evidence of students' problems in interpreting such representations (Bowen & Roth, 2002; Leinhardt, Zaslavsky, & Stein, 1990; Roth & McGinn, 1998; Vekiri, 2002). Progress maps and other graphs could play a significant role toward the goal of fostering interpretative skills, but they must be designed properly. Students are most likely to understand graphs that require the least amount of cognitive processing, either because of a simplified spatial organization (i.e., all related information is grouped together), or auditory or written explanations identifying the graph's purpose and directing students' attention to key interpretive details (Vekiri, 2002). BEAR and KIDMAP incorporate these principles by including concrete suggestions for future learning, and advocating discussions of graphed progress among teachers, students, and parents. If interpretation is properly scaffolded in manners like these, graphs can be extremely valuable tools for informative assessment among students.

Structured Opportunities for Reflection and Revision

Informative assessment is instrumental in the development of self-monitoring expertise and establishes routine opportunities to reflect on performance. Initially, learning involves a significant degree of external support that is controlled by the teacher or the tutor. As learning proceeds, there are increasing opportunities for the acquisition of self-regulatory skills, and the identification and discrimination of criteria for high levels of performance. As this stage progresses, the design of the instructional situation becomes increasingly under the control of the student as a developing expert. There is a selective use of external support with students observing the performance of other students and calling on the advice of the teacher only as needed.

A number of features can be incorporated into assessment situations so that the student can work with others to observe performance and receive feedback, with the opportunity to refer to supporting materials or the instructor as necessary. This report has already discussed the roles of models of competence and graphs in facilitating self-assessment. Another critical feature we discuss here is the need for deliberate, structured opportunities to reflect on performance.

Consider the lessons to be learned from selected programs that promote self-assessment. One is that is that self-assessment must be accompanied with suggestions for improvement if it is to facilitate learning. ThinkerTools (White, 1993; White & Frederiksen, 1998), as noted before, provides habitual opportunities for students to assess their progress on selected aspects of science inquiry. Class discussions, teacher feedback, and opportunities to evaluate anonymous reports ensure that students understand the assessment criteria and their appropriate application (White & Frederiksen, 1998). An equally important component of reflective assessment is information on how to succeed on projects, as communicated through evaluation criteria, conversations with higher-performing peers, class discussions, and teacher feedback. If students only evaluate their performances without understanding how to improve them, they run the risk of being demoralized by low scores and convinced that they lack the ability to do well (White & Frederiksen, 1998).

Suggestions for promoting reflection through assessment must be taken with some caution. Designing prompts for productive reflection is a complex process that must take learners' beliefs and abilities, and learning environment characteristics into consideration. Programs such as ThinkerTools (White & Frederiksen, 1998) demonstrate the types of activities that help students design and understand standards for performance, and then use those standards to evaluate their own learning. At the same time, it cannot be assumed that all prompts or reflective activities will be equally effective for all students. For example, giving middle-school students specific reflective prompts during inquiry-based science instruction (e.g., "To do a good job on this project, we need to …") can actually be detrimental to students who have moderate trouble taking responsibility for their own learning. Generic prompts (e.g., "Right now we're thinking about …"), on the other hand, seem to allow those students the opportunity to reflect on ideas of their own choosing, which in turn leads to better performance (Davis, 2003). This is not to suggest that generic prompts are suitable for all classrooms, as they have primarily been examined in highly scaffolded computer-supported learning environments (Davis, 2003).

Studies of individual differences in reflection suggest that there is no ideal template for assessment design that is going to elicit productive monitoring and idea revision for all learners in all instructional situations. Rather, self-assessment opportunities place extra responsibilities on teachers to monitor students' performance, identify ongoing difficulties (e.g., overestimation of abilities), and correct misunderstandings. Opportunities for reflection and revision may need to be tailored to the needs of individual learners, with future research focusing on how this might be accomplished (Davis, 2003).

One way to accommodate individual differences in monitoring ability is through one-on-one conversations about student performance. Although these conversations are typically associated with private conferences about portfolios (e.g., Courtney & Abodeeb, 1999; Klimenkov & LaPick, 1996; Schipper & Rossi, 1997), they can also be conducted during class discussions of subject matter (Clarke, McCallum, & Lopez-Charles, 2001). Effective interactions relate students' self-evaluation to specific criteria, direct questions to particular students (e.g., asking for evaluations of certain criteria with which a student has had problems in the past), probe for details, provide feedback about the quality of self-evaluations, and generate goals and plans for future work (Courtney & Abodeeb, 1999; Clarke et al., 2001). For instance, one second grade teacher worked with students to create three individualized learning goals, which she recorded on index cards and taped to students' desks as reminders (Courtney & Abodeeb, 1999). Because personal interactions permit the immediate recognition and response to student difficulties (Bell & Cowie, 2001; Cowie & Bell, 1999), they might be more amenable to promoting self-assessment than are structured, group-administered assessments alone.

Conclusion

> Evaluation used to improve the course while it is still fluid, contributes more to the improvement of education than evaluation used to appraise a product already placed on the market … Hopefully, evaluation studies will go beyond reporting on this or that course and help us understand educational learning. Such insight will, in the end, contribute to the development of all courses rather than just the course under test. (Cronbach, 1963, p. 675).

This quotation is a commentary on the progress that has been made in uniting instruction and assessment for the improvement of classroom learning. In the past four decades since Cronbach wrote about formative evaluation, a central assessment goal has remained the same: to develop optimally informative measures of understanding that productively redirect teacher and student behavior. During that time, strides have been taken to achieve this goal, including a reconceptualization of what it means to "know" a discipline, the merger of psychometrics with cognitive science, new observational and statistical models for documenting learning, and the emergence of models of exemplary assessment practice (NRC, 2001). Efforts must now also be made to operationalize assessment development in as practical, detailed, and comprehensive a manner as possible, until informative tools can become commonplace in classroom learning environments.

This report has identified several principles of assessment programs that inform teachers and students of performance on relevant aspects of achievement and suggest directions for future learning. Principles fell into three general categories: (a) descriptions of cognitive activity, (b) presentation of assessment goals and feedback, and (c) transformation of feedback into recommendations for teaching and learning. Our efforts, while highlighting a selection of desirable assessment properties, also refer to one effort to "accumulate, synthesize, and disseminate existing knowledge" (NRC, 2001, p. 299) for the purposes of articulating assessment development, particularly with regards to analyses that "cut across effective exemplars with the goal of identifying and clarifying the new science of assessment design" (NRC, 2001, p. 304).

This paper has concentrated on the interpretive elements of assessment design and supporting materials. Similar reports could discuss, for instance, the qualities of teacher feedback that sustain learning (e.g., Black, Harrison, Lee, Marshall, & Wiliam, 2002; Black & Wiliam, 1998; Suffolk County Council, 2001), or the elements of task design in various fields (e.g., Baxter & Glaser, 1998; Solano-Flores & Shavelson, 1997; Sugrue, 1995). Analyses of particular forms of assessments such as performance assessments or short-answer questions could be especially fruitful. Haladyna, Downing, & Rodriguez's (2002) taxonomy of multiple-choice guidelines, derived from a synthesis of research studies and educational measurement textbooks, represents one approach with promising implications for future item creation guidelines.

Informative assessment can also benefit from a design research approach (Brown, 1992; Edelson, 2002) whereby task development is systematically documented, with results generalized to various classes of assessment contexts (e.g., across subject matters, units within the same subject, age groups, assessment types, etc.). This information is a rare commodity; assessment programs are generally presented as final products, with little discussion of the creation process, design evaluation and revision, and lessons learned along the way. If such details were made available for public consideration (e.g., taking the form of design cases; Edelson, 2002), they would make invaluable contributions to defining the "design space" (NRC, 2001), establishing the parameters and principles for the science and technology of informative classroom assessment design.

A Working Language of Cognition for Educators

In addition to research on the development of informative assessments, we urge that a language be developed that can serve as a source of ready communication between researchers and educators. Examples of such a language have already been presented in this paper in the discussion of assessment properties to inform teachers. More work is needed, however. This discourse would be a merger of learning concepts and cognitive processes that guide and develop practices in the classroom in terms of the feedback provided by informative assessment. The teacher would use this feedback to influence student performance in learning various subject matter. A general set of principles would be incorporated for practices involved in different subject matters, so a major recommendation is the development of language of learning and cognition for teacher practice toward a level of discourse that would relate learning processes to instructional performances. Consideration would need to be given to both general concepts cross-cutting domains and to specific kinds of performance required in a variety of subject matters and school situations.

The effort to accomplish this objective would reference a consistent vocabulary that would consider the context and description of performance, as well as the processes of learning. Working back and forth between descriptions of instructional situations and learning terminology would enable the refinement and emergence of discourse conventions. The goal of this effort would be developed on a basis of a shared usage in a variety of situations. Sources and test situations for this working language for informative assessment would involve an integration of disciplinary concepts and learning theory, an analysis of the educational context, and descriptions of various levels of performance and developing proficiency.

A Classroom Environment of the Future

Efforts to construct informative assessments will significantly impact classroom environments that promote active knowledge. In these classrooms, teachers and students model practices encouraged by cognitive and situative learning theories, such as attention to prior knowledge, concentration on learning with understanding, and establishment of a community of learners (NRC, 2000, 2001). Assessments have been and will continue to be an integral part of this process by providing information that facilitates the design of adaptive learning environments. To teach for understanding, teachers must receive insights about students' thinking, reasoning, and problem solving that they can use to modify subsequent learning opportunities. To learn *with* understanding, students must obtain feedback about their performance that helps them revise their ideas and evaluate future work.

The classroom thus becomes an environment of interchange interspersed by periods of reflection and problem solving. The teachers' reflection comes about as a result of interpretation of student performance; the students' reflection is the result of their reactions to feedback presented over the course of learning. There are periods in which both teachers and students appear to come together in appreciation that their activity has been successful, with the teachers planning for next steps and the students anticipating the development of new knowledge.

This report has identified principles of informative assessment development taken primarily from localized, researcher-supported programs. These practices can now be studied and expanded to the mainstream. Projects that identify development principles and disseminate those ideas to teachers will be particularly helpful in this regard. There is no question that effective, practical, informative assessments can make substantial inroads in improving classroom teaching and learning. The challenge that lies ahead is to capitalize on the principles underlying exemplary assessment programs and undertake development efforts to make informative assessment opportunities accessible to all.

REFERENCES

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance.* Thousand Oaks, CA: Corwin Press, Inc.

Balanced Assessment for the Mathematics Curriculum (1999). *Elementary Grades Assessment Package 1.* White Plains, NY: Dale Seymour Publications.

Baxter, G. P., Elder, A. D., & Glaser, R. (1995). *Cognitive analysis of a science performance assessment.* CSE Technical Report 398. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Baxter, G. P., Elder, A. D., & Shavelson, R. J. (1997). *Effect of embedded science assessments on students' end-of-unit performance.* Unpublished manuscript, University of Michigan.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*(3), 37-45.

Bell, B., & Cowie, B. (2001). *Formative assessment and science education.* Dordrecht, The Netherlands: Kluwer.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box: Assessment for learning in the classroom.* London: School of Education, King's College.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7-74.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13,* 259-78.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges to creating complex interventions in classroom settings. *The Journal of Learning Sciences, 2,* 141-178.

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289-323). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, A. L., Ellery, S., & Campione, J. C. (1998). Creating zones of proximal development electronically. In J. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 341-367). Mahwah, NJ: Lawrence Erlbaum Associates.

Bowen, G. M., & Roth, W. M. (2002). Why students may not learn to interpret scientific inscriptions. *Research in Science Education, 32,* 303-327.

Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *Elementary School Journal, 97*, 3-20.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction.* Portsmouth, NH: Heinemann.

Carver, S. M., & Klahr, D. (Eds.). (2001). *Cognition and instruction: Twenty-five years of progress.* Mahwah, NJ: Lawrence Erlbaum Associates.

CATCH (2002). *About CATCH.* Retrieved August 7, 2002 from http://www.fi.uu.nl/catch/

Clarke, S, McCallum, B., & Lopez-Charles, G. (2001). *Gillingham partnership: Formative assessment project 2000-2001.* London: University of London, Institute of Education.

Courtney, A. M., & Abodeeb, T. L. (1999). Diagnostic-reflective portfolios. *Reading Teacher, 52,* 708-714.

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education, 6,* 101-116.

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record, 64*, 672-683.

Davis, E. A. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *Journal of the Learning Sciences, 12,* 91-142.

Dekker, T., Querelle, N., & van den Boer, C. (2000). *The Great Assessment Picture Book.* Retrieved August 7, 2002 from http://www.fi.uu.nl/catch/

DIAGNOSER (2002). The Diagnoser project: Instructional tools for science and mathematics. Retrieved November 7, 2002 from http://tutor.psych.washington.edu/

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *The Journal of the Learning Sciences, 11,* 105-121.

Erickson, J., & Lehrer, R. (1998). The evolution of critical standards as students design hypermedia documents. *The Journal of Learning Sciences, 7*, 351-386.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Glaser, R., & Silver, E. (1994). Assessment, testing and instruction: Retrospect and prospect. *Review of Research in Education, 20,* 393-419.

Goldberg, G. L., & Roswell, B. S. (1999). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment, 6,* 257-290.

Haladyna, T. M., Dowling, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15,* 309-334.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice.* (pp. 51-74). Cambridge, MA: MIT Press/ Bradford Books.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ornseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11,* 121-137.

Klimenkov, M., & LaPick, N. (1996). Promoting student self-assessment through portfolios, student-facilitated conferences, and cross-age interaction. In R. C. Calfee & P. Perfumo (Eds.), *Writing portfolios in the classroom: Policy and practice, promise and peril.* (pp. 239-260). Mahwah, NJ: Lawrence Erlbaum Associates.

Leinhardt, G. (1990). Capturing craft knowledge in teaching. *Educational Researcher, 17* (2), 18-25.

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*, 1–64.

Luft, J. A. (1999). Rubrics: Design and use in science teacher education. *Journal of Science Teacher Education, 10,* 107-121.

Martínez, M. A., Sauleda, N., & Huber, G. L. (2002). Metaphors as blueprints of thinking about teaching and learning. *Teaching and Teacher Education, 17,* 965-977.

McGilly, K. (Ed.). (1994). *Classroom lessons: Integrating cognitive theory and classroom practice.* Cambridge, MA: MIT Press/ Bradford Books.

Minstrell, J. (2000). Student thinking, instruction and assessment in a facet-based learning environment. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 44-73). Washington, DC: National Academy Press.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley, C. Schunn & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 415-443). Mahwah, NJ: Erlbaum.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: Author.

National Research Council (1996). *National science education standards.* Washington, DC: National Academy Press.

National Research Council (2000). *How people learn: Brain, mind, experience, and school.* Committee on Developments in the Science of Learning and Committee on Learning Research and Educational Practice. J. D. Bransford, A. L. Brown, and R. R. Cocking (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser, (Eds.), Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

New South Wales Department of Education and Training (1998). *Principles for effective and informative assessment and reporting.* Retrieved April 2, 2003, from http://www.schools.nsw.edu.au/learning/k-6assessments/principles.php

New South Wales Department of Education and Training (2001). *Using KIDMAP to support effective and informative assessment and reporting.* Retrieved July 26, 2002 from http://www.oten.edu.au/kidmap/pdf/KIDMAP_AR.pdf

No Child Left Behind Act of 2001, Public L. No. 107-110, 115 Stat. 1425 (2002).

Palincsar, A., & Brown, A. (1984). Reciprocal teaching of comprehension fostering and comprehension monitoring activities. *Cognition and Instruction, 1,* 117-175.

Pellegrino, J., Baxter, G. P., & Glaser, R. (1999). Addressing the "Two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24,* 307-353.

Roberts, L., Wilson, M., & Draney, K. (1997). *The SEPUP assessment system: An overview.* University of California, Berkeley: BEAR Report Series, SA-97-1.

Roth, W. M., & McGinn, M. K. (1998). Inscriptions: A social practice approach to "representations." *Review of Educational Research, 68*, 35–59.

Schipper, B., & Rossi, J. (1997). *Portfolios in the classroom: Tools for learning and instruction.* York, ME: Stenhouse Publishers.

Shafer, W. D., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education, 14,* 151-170.

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

Shepard, L. A. (2000b). *The role of classroom assessment in teaching and learning.* CSE Technical Report 517. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16-25.

Suffolk County Council (2001). *"How am I doing?"—Assessment and feedback to learners.* Suffolk County, England.

Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem-solving ability. *Educational Measurement: Issues and Practice, 14*(3), 29-35.

Vekiri, I. (2002). What is the value of graphical displays of learning? *Educational Psychology Review, 14,* 261-312.

Verhage, H., & de Lange, J. (1997). Mathematics education and assessment. *Pythagoras, 42*, 14-20.

Waltman, K., Kahn, A., & Koency, G. (1998). *Alternative approaches to scoring: The effect of using different scoring methods on the validity of scores from a performance assessment.* CSE Technical Report 488. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

White, B. Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition and Instruction, 10,* 1-100.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16,* 3-118.

Wiggins, G., & McTighe, J. (1998). *Understanding by design.* Alexandria, VA: Association for Supervision and Curriculum Development.

Wilson, M., Draney, K., & Kennedy, C. (2001). *GradeMAP* [computer program]. Berkeley, CA: BEAR Center, University of California.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181-208.