INDICATOR SYSTEMS

CHAPTER 12

Establishing Multilevel Coherence in Assessment

DREW H. GITOMER AND RICHARD A. DUSCHL

The enactment of the No Child Left Behind Act (NLCB) has resulted in an unprecedented and very direct connection between high-stakes assessments and instructional practice. Historically, the *disassociation* between large-scale assessments and classroom practice has been decried, but the current irony is that the influence these tests now have on educational practice has raised even stronger concerns (e.g., Abrams, Pedulla, & Madaus, 2003) stemming from a general narrowing of the curriculum, both in terms of subject areas and in terms of the kinds of skills and understandings that are taught. The cognitive models underlying these assessments have been criticized (Shepard, 2000), evidence is still collected primarily through multiple choice items, and psychometric models still order students along a single dimension of proficiency.

However, NCLB can be viewed as an opportunity to develop a comprehensive assessment system[1] that supports educational decision making about student learning and classroom instruction consistent with theories and standards of subject matter learning. The purpose of this chapter is to propose a framework for designing coherent assessment systems, using science education as an exemplar, that provides useful information to policymakers at the same time it supports learning

Drew H. Gitomer is Distinguished Researcher at the Policy Evaluation Research Center of Educational Testing Service. Richard A. Duschl is Professor of Science Education at the Graduate School of Education and an executive member of the Center for Cognitive Science at Rutgers, The State University of New Jersey.

and teaching in the classroom. The framework is based on a review of existing literature on the nature of learning, particularly in science, emerging developments in assessment practices, and the organizational use of assessment evidence.

Developing large-scale assessment systems that can support decision making for state and local policymakers, teachers, parents, and students has proven to be an elusive goal. Yet the idea that educational assessment ought to better reflect student learning and afford opportunities to inform instructional practice can be traced back at least 50 years, to Cronbach's (1957) seminal article "The Two Disciplines of Scientific Psychology." These ideas continued to evolve with Glaser's (1976) conceptualization of an *instructional psychology* that would adapt instruction to students' individual knowledge states. Further developments in aligning cognitive theory and psychometric modeling approaches have been summarized by Glaser and Silver (1994); Pellegrino, Baxter, and Glaser (1999); Pellegrino, Chudowsky, and Glaser (2001); the National Research Council (2002); and Wilson (2004).

In this chapter, the authors propose an assessment framework for science education that is based on the idea of multilevel coherence. First, assessment systems are *externally coherent* when they are consistent with accepted theories of learning and valued learning outcomes. Second, assessment systems can be considered *internally coherent* to the extent that different components of the assessment system, particularly large-scale and classroom components, share the same underlying views of learners' academic development. The challenge is to design assessment systems that are both internally and externally coherent.[2]

We contend that while significant progress is being made in conceptualizing external coherence, the challenge to any substantial change in practice is predicated upon designing internally coherent systems that are not only consistent with theories of learning and practice, but are also pragmatic and scalable solutions in the face of very real constraints. Such designs will also need to give much more consideration to the quality and processes for interpreting assessment results across all stakeholders and decision makers in the educational system. As Coburn, Honig, and Stein (in press) have noted, the use of evidence in school districts is relatively haphazard and used to confirm existing practice, rather than used to investigate, in a disciplined manner, the validity of assumptions and practices operating in the educational system.

Coherence, like validity, is not an absolute to be attained but a goal to be pursued. Therefore, rather than defining an optimally coherent

assessment system, we attempt to outline the features of systems that maximize both internal and external coherence. We also describe challenges to establishing coherence, particularly in light of the very real constraints (e.g., cost and time available) that surround any viable assessment system. Although the focus is on science education, we believe that the basic line of argument is generalizable across content domains.

In order to support effective, assessment-based decision making, we need to consider a series of issues in the design of assessment systems. These issues guide the organization of the chapter.

1. What is the nature of the learning model on which the assessment is based?
2. How can assessments be *designed* to be externally coherent (i.e., attuned to the underlying learning model)?
3. How can assessment designs be *implemented* (for internal coherence, meaning both large-scale and classroom assessments) given practical constraints in the educational system?

A Learning Model to Guide Science Assessment

The major transformation under way in conceptualizing the learning goals for an externally coherent assessment system has been the recognition of three important perspectives: the *cognitive*, *sociocultural*, and *epistemic*. Including these three perspectives fundamentally broadens the nature of the construct underlying science assessment. This expansion of the construct means that assessment design involves more than simply improving the measurement of an existing construct.

The *cognitive perspective* focuses on knowledge and skills that students need to develop. Glaser's (1997) list of cognitive dimensions, derived from the human expertise literature, reflects a consensus among learning theorists (e.g., Anderson, 1990; Bransford, Brown, & Cocking, 1999). We add to Glaser's categories with our own commentary:

Structured, Principled Knowledge

Learning involves the building of knowledge structures organized on the basis of conceptual domain principles. For example, chess experts can recall far more information about a chessboard, not because of better memories, but because they recognize and encode familiar game patterns as easily recalled, integrated units (Chase & Simon, 1973).

Proceduralized Knowledge

Learning involves the progression from declarative states of knowledge ("I know the rules for multiplying whole numbers by fractions"), to proceduralized states in which access is automated and attached to particular conditions ("I apply the rules for multiplying by fractions appropriately, with little conscious attention," e.g., Anderson, 1983).

Effective Problem Representation

As learners gain expertise, their representations move from a focus on more superficial aspects of a problem to the underlying structures. For example, Chi, Feltovich, and Glaser (1981) showed that experts organized physics problems on the basis of underlying physics principles, while novices sorted the problems on the basis of surface characteristics.

Self-Regulatory Skills

Glaser (1992) refers to learners becoming increasingly able to monitor their learning and performance, to allocate their time, and to gauge task difficulty.

Taken together, then, assessments ought to focus on integrated knowledge structures, the efficient and appropriate use of knowledge during problem solving, the ability to use and interpret different representations, and the ability to monitor and self-regulate learning and performance.

The *socio-cultural/situative* perspective focuses on the nature of social interactions and how they influence learning. From this perspective, learning involves the adoption of socio-cultural practices, including the practices within particular academic domains. Students of science, for example, not only learn the content of science; they also develop an "intellective identity" (Greeno, 2002) as scientists by becoming acculturated to the tools, practices, and discourse of science (Bazerman, 1988; Gee, 1999; Lave & Wenger, 1991; Rogoff, 1990; Roseberry, Warren, & Contant, 1992). This perspective grows out of the work of Vygotsky (1978) and others, and posits that learning and practices develop out of social interaction and thus cannot be studied with the traditional intra-personal cognitive orientation.

Certainly, some socio-cultural theorists would argue that attempts to administer some form of individualized and standardized assessment are antithetical to the fundamental premise of a theory that is based on social interaction. Our response is that all assessments are proxies that

can only approximate the measure of much broader constructs. Given the set of constraints that exist within our current educational system, we choose to strive for an accommodation of socio-cultural perspectives by attending to certain critical domain practices in our assessment framework, while acknowledging that we are not yet able to attend to all of those social practices. Mislevy (2006) has described models of assessment that reflect similar kinds of compromise.

What, then, are some key attributes of assessment design that would be consistent with a socio-cultural perspective and that would represent a departure from more traditional assessments? We focus on the tools, practices, and interactions that characterize the community of scientific practice.

Public Displays of Competence

Productive classroom interactions mandate a much more public display of student work and learning performances, open discussion of the criteria by which performance is evaluated, and discussion among teachers and students about the work and dimensions of quality. Gitomer and Duschl (1998) have described strategies for making student thinking visible through the use of various assessment strategies that include both an elicitation of student thinking through evocative prompts and argumentation discussions around that thinking in the classroom.

Engagement With and Application of Scientific Tools

Certainly, a great deal of curriculum and assessment development has focused on the use of science tools and materials in conducting some components of science investigations. Despite limitations noted later in the chapter, assessments ought to include activities that require students to engage with tools of science and understand the conditions that determine the applicability of specific tools and practices.

Self-Assessment

A key self-regulatory skill that is a marker of expertise is the ability and propensity to assess the quality of one's own work. Assessments should provide opportunities, through practice, coaching, and modeling, for students to develop abilities to effectively judge their own work.

Access to Reasoning Practices

As Duschl and Gitomer (1997) have articulated, science assessment can contribute to the establishment and development of science practice

by students, facilitated by teachers. Certainly, the current emphasis on formative assessment and assessment for learning (e.g., Black & Wiliam, 1998; Stiggins, 2002) suggests that assessments can be designed to encourage productive interactions with students that engage them in important reasoning practices.

Socially Situated Assessment

Expertise is often expressed in social situations in which individuals need to interact with others. There is often exchange, negotiation, building on others' input, contributing and reacting to feedback, etc. (Webb, 1997, 1999). Indeed, the ability to work within social settings is highly valued in work settings and insufficiently attended to in typical schooling, including assessment.

Models of Valued Instructional Practice

Assessments exist within an educational context and can have intended and unintended consequences for instructional practice (Messick, 1989). A primary criticism of the traditional high-stakes assessment methodology is that it has supported adverse forms of instruction (Amrein & Berliner, 2002a, 2002b). By attending to the socio-cultural practices described above, assessment designs provide models of practice that can be used in instruction.

The *epistemic* perspective further clarifies what it means to learn science by situating the cognitive and socio-cultural perspectives in specific scientific activities and contexts in which the growth of scientific knowledge is practiced. There are two general elements in the epistemic perspective—one disciplinary, the other methodological. Knowledge building traditions in science disciplines (e.g., physical, life, earth and space, medical, social), while sharing many common features, are actually quite distinct when the tools, technologies, and theories each uses are considered. Such distinctions shape the inquiry methods adopted. For example, geological and astronomical sciences will adopt historical and model-based methods as scientists strive to develop explanations for the formation and structures of the earth, solar system, and universe. Causal mechanisms and generalizable explanations aligned with mathematical statements are more frequent in the physical sciences where experiments are more readily conducted. Whereas molecular biology inquiries often use controlled experiments, population biology relies on testing models that examine observed networks of variables in their natural occurrence.

Orthogonal to disciplinary distinctions, the second element of the epistemic perspective includes shared practices like modeling, measuring, and explaining that frame students' classroom investigations and inquiries. The National Research Council (NRC) report "Taking Science to School" (Duschl, Schweingruber, & Shouse, 2006) argues that content and process are inextricably linked in science. Students who are proficient in science:

1. Know, use, and interpret scientific explanations of the natural world;
2. Generate and evaluate scientific evidence and explanations;
3. Understand the nature and development of scientific knowledge; and
4. Participate productively in scientific practices and discourse.

These four characteristics of science proficiency are not only learning goals for students but they also set out a framework for curriculum, instruction, and assessment design that should be considered together rather than separately. They represent the knowledge and reasoning skills needed to be proficient in science and to participate in scientific communities, be they classrooms, lab groups, research teams, workplace collaborations, or democratic debates.

The development of an enriched view of science learning echoes 20th century developments in philosophy of science in which the conception of science has moved from an experiment-driven to a theory-driven to the current model-driven enterprise (Duschl & Grandy, 2007). The experiment-driven enterprise gave birth to the movements called *logical positivism* or *logical empiricism*, shaped the development of analytic philosophy, and gave rise to the hypothetico-deductive conception of science. The image of scientific inquiry was that of experiments leading to new knowledge that accrued to established knowledge. The justification of knowledge was of predominant interest. *How* that knowledge was discovered and refined was not part of the philosophical agenda. This early 20th century perspective is referred to as the "received view" of philosophy of science and is closely related to traditional explanations of "the scientific method," which include such prescriptive steps as making observations, formulating hypotheses, making observations, etc.

The model-driven perspective is markedly different from the experiment model that still dominates K-12 science education. In this model, scientific claims are rooted in evidence and guided by our best-reasoned beliefs in the form of scientific models and theories that frame investigations and inquiries. All elements of science—questions, methods,

evidence, and explanations—are open to scrutiny, examination, and attempts at justification and verification. *Inquiry and the National Science Education Standards* (National Research Council, 2000) identifies five essential features of such classroom inquiry:

- Learners are engaged by scientifically oriented questions.
- Learners give priority to *evidence*, which allows them to develop and evaluate explanations that address scientifically oriented questions.
- Learners formulate *explanations* from evidence to address scientifically oriented questions.
- Learners evaluate their explanations in light of alternative explanations, particularly those reflecting scientific understanding.
- Learners communicate and justify their proposed explanations.

Implications of the Learning Model for Assessment Systems

The implications for an assessment system externally coherent with such an elaborated model of learning are profound. Assessments need to be designed to monitor the cognitive, socio-cultural, and epistemic practices of doing science by moving beyond treating science as the accretion of knowledge to a view of science that, at its core, is about acquiring data and then transforming that data first into evidence and then into explanations.

Socio-cultural and epistemic perspectives about learning reshape the construct of science understanding and inject a significant and alternative theoretical justification for not only what we assess, but also how we assess. The predominant arguments for moving to performance assessment have been in terms of consequential validity, what Glaser (1976) termed instructional effectiveness, and face validity—having students engage in tasks that look like valued tasks within a discipline. But using these tasks has often been considered a trade-off with assessment quality—the capacity to accurately gauge the knowledge and skills a student has attained. For example, Wainer and Thissen (1993), representing the classic psychometric perspective, calculated the incremental costs to design and administer performance assessments that would have the same measurement precision as multiple-choice tests. They estimated that the anticipated costs would be orders of magnitude greater to achieve the same measurement quality.

When the socio-cultural and epistemic perspectives are included in our models of learning, it becomes clear that the psychometric rationale is markedly incomplete. Smith, Wiser, Anderson, and Krajcik (2006)

note that "[current standards] specify the knowledge that children should have, but not practices—what children should be able to *do* with that knowledge" (p. 4). The argument of the centrality of *practices* as demonstrations of subject-matter competence implies that assessments that ignore those practices do not adequately or validly assess the constellation of coordinated skills that encompass subject-matter competence. Thus, the question of whether multiple-choice assessments can adequately sample a domain is necessarily answered in the negative, for they do not require students to engage and demonstrate competence in the full set of practices of the domain.

The Evidence-Explanation Continuum

What might an assessment design that does account for socio-cultural and epistemic perspectives look like? The example that follows is grounded in prior research on classroom portfolio assessment strategies (Duschl & Gitomer, 1997; Gitomer & Duschl, 1998) and in a "growth of knowledge framework" labeled the Evidence-Explanation (E-E) Continuum (Duschl, 2003). The E-E approach emphasizes the progression of "data-texts" (e.g., measurements to data to evidence to models to explanations) found in science, and it embraces the cognitive, socio-cultural, and epistemic perspectives. What makes the E-E approach different from traditional content/process and discovery/inquiry approaches to science education is the emphasis on the *epistemological conversations* that unfold through processes of argumentation.

In this approach, inquiry is linked to students' opportunities to examine the development of data texts. Students are asked to make reasoned judgments and decisions (e.g., arguments) during three critical transformations in the E-E Continuum: *selecting* data to be used as evidence; *analyzing* evidence to extract or generate models and/or patterns of evidence; and *determining and evaluating* scientific explanations to account for models and patterns of evidence.

During each transformation, students are encouraged to share their thinking by engaging in argument, representation and communication, and modeling and theorizing. Teachers are guided to engage in assessments by comparing and contrasting student responses to each other and, importantly, to the instructional aims, knowledge structures, and goals of the science unit. Examination of students' knowledge, representations, reasoning, and decision making across the transformations provides a rich context for conducting assessments. The advantage of this approach resides in the formative assessment opportunities for

students and the cognitive, socio-cultural, and epistemic practices that comprise "doing science" that teachers will monitor.

A critical issue for an internally coherent assessment system is whether these practices can be elicited, assessed, and encouraged with proxy tasks in more formal and large-scale assessment contexts as well. The E-E approach has been developed in the context of extended curricular units that last several weeks, with assessment opportunities emerging throughout the instructional process. For example, in a chemistry unit on acids and bases, students are asked to reason through the use of different testing and neutralization methods to ensure the safe disposal of chemicals (Erduran, 1999).

While extended opportunities such as these are not pragmatic within current accountability testing paradigms, there have been efforts to design assessment that can be used to support instructional practice consistent with theories much more aligned with emerging theories of performance (e.g., Pellegrino et al., 2001). However, even these efforts to bridge the gap between cognitive science and psychometrics have given far more attention to the conceptual dimensions of learning than to those associated with practices within a domain, including how one acquires, represents, and communicates understanding. Nevertheless, Pellegrino et al. is rich with examples of assessments that demonstrate external coherence on a number of cognitive dimensions, providing deeper understanding of student competence and learning needs. These assessment tasks typically ask students to represent their understanding rather than simply select from presented options. A mathematics example (Magone, Cai, Silver, & Wang, 1994) asks students to reason about figural patterns by providing both graphical representations and written descriptions in the course of solving a problem. Pellegrino et al. also review psychometric advances that support the analysis of more complex response productions from students. Despite the important progress represented in their work, socio-cultural and epistemic perspectives remain largely ignored.

Two recent reports (Duschl et al., 2006; National Assessment Governing Board [NAGB], 2006) offer insights into the challenge of designing assessments that do incorporate these additional perspectives. The 2009 National Assessment of Educational Progress (NAEP) Science Framework (NAGB, 2006) sets out an assessment framework grounded in (1) a cognitive model of learning and (2) a view of science learning that addresses selected scientific practices such as coordinating evidence with explanation within specific science contexts. Both reports take up the ideas of "learning progressions" and "learning per-

formances" as strategies to rein in the overwhelming number of science standards (National Research Council, 1996) and benchmarks and provide some guidance on the "big ideas" (e.g., deep time, atomic molecular theory, evolution) and important scientific practices (e.g., modeling, argumentation, measurement, theory building) that ought to be at the heart of science curriculum sequences.

Learning progressions are coordinated long-term curricular efforts that attend to the evolving development and sophistication of important scientific concepts and practices (e.g., Smith et al., 2006). These efforts recommend extending scientific practices and assessments well beyond the design and execution of experiments, so frequently the exclusive focus of K-8 hands-on science lessons, to the important epistemic and dialogic practices that are central to science as a way of knowing. Equally important is the inclusion of assessments that examine understandings about how we have come to know what we believe and why we believe it over alternatives; that is, linking evidence to explanation.

Given the significant research directed toward improving assessment practice and compelling arguments to develop assessments to support student learning, one might expect that there would be discernible shifts in assessment practices throughout the system. While there has been an increasing dominance of assessment in educational practice brought about by the standards movement, culminating in NCLB, we have not witnessed anything that has fundamentally shifted the targeted constructs, assessment designs, or communications of assessment information. We believe that the failure to transform assessment stems from the necessary *but not sufficient* need to address issues of consistency between methods for collecting and interpreting student evidence and operative theories of learning and development (i.e., external coherence).

In addition to *external coherence*, we contend that an effective system will also need to confront issues of the *internal coherence* between different parts of the assessment system, the *pragmatics* of implementation, and the *flow of information* among the stakeholders in the system. Indeed, we argue that the lack of impact of the work summarized by Pellegrino et al. (2001) and promised by emerging work in the design of learning progressions is due, in part, to a lack of attention and solutions to the issues of internal coherence, pragmatics, and flow of information.

In the remainder of this chapter, we present an initial framework to describe critical features of a comprehensive assessment system intended to communicate and influence the nature of student learning

and classroom instruction in science. We include advances in theory, design, technology, and policy that can support such a system. We close with challenges that must be confronted to realize such a system.

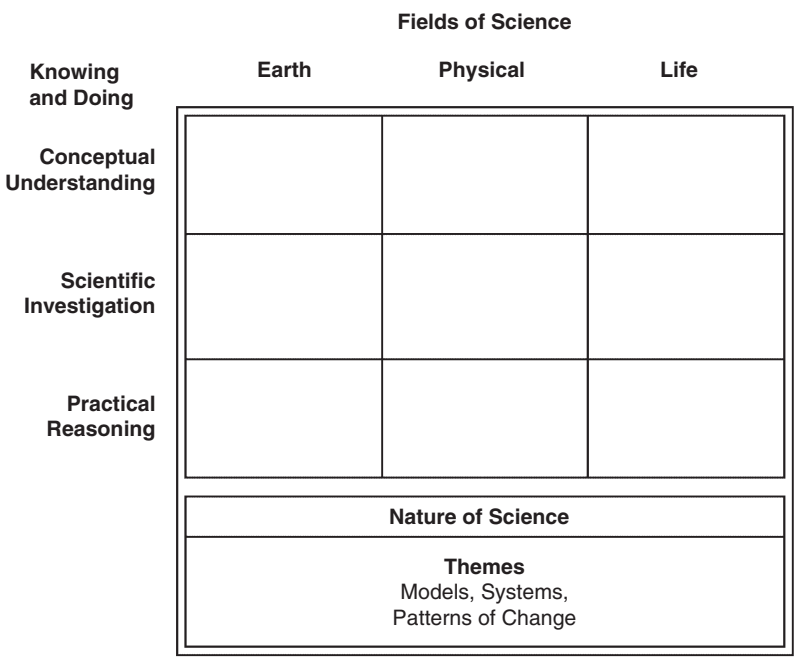
Learning Theory and Assessment Design—Establishing External Coherence

Large-scale science assessment design has faced particular challenges because of the lack of any generally accepted curricular sequence or content. The need to sample content from a very broad range of potential science concepts led to assessments largely oriented toward the recall and recognition of discrete science facts. The basic logic was that such broad sampling would ultimately be a fair method of gauging students' relative understanding of science content. This practice of assessment design was consistent with a model of science learning as the accretion of specific facts about different science concepts, with very little attention to scientific practices.

This general model of science assessment was met with dissatisfaction, particularly because of a lack of attention to practices critical to scientific understanding—most notably practices associated with inquiry, including theory building, modeling, experimental design, and data representation and interpretation. In fact, this type of assessment was in direct conflict with emerging models of science curriculum that emphasized science reasoning and deeper conceptual understanding, described in the previous section. Beginning in the 1980s, state science frameworks emphasized attention to a more comprehensive range of skills and understandings. A national consensus framework developed for the NAEP (National Assessment Governing Board, 1996) proposed a matrix that included the application of a variety of reasoning processes applied to the earth, physical, and life sciences (Figure 1).

Certainly, questions developed from these frameworks were quite a bit different from earlier questions. Assessment tasks were much more concerned with the understanding of concepts and systems rather than the recognition of definitions or recall of particular nomenclature (e.g., parts of a flower). Additional questions were developed that addressed skills associated with scientific investigation, such as the manipulation of variables in a controlled study or the interpretation of graphical data. Assessments even included what became known as "hands-on" performance tasks, in which students manipulated physical objects in laboratory-like activities to do such things as take measurements, record observations, and conduct controlled mini-experiments (e.g., Gitomer & Duschl, 1998; Shavelson, Baxter, & Pine, 1992).

FIGURE 1
NAEP ASSESSMENT MATRIX FOR 1996–2000 ASSESSMENTS



Notable about these assessments was that, despite the apparent multidimensionality of the framework, process and content were treated almost completely distinctly. Although items that addressed investigative skills were posed within a science context, the demands of the task required virtually no understanding of the content itself. For example, Pine et al. (2006) studied a set of assessment tasks taken from the Full Option Science Series (FOSS). Examining four hands-on tasks, they demonstrated that performance on these and other investigative and practical reasoning assessment tasks could be solved through the application of logical reasoning skills independent of any significant conceptual understanding from biology, physics, or chemistry, concluding that general measures of cognitive ability explained task performance far more than any other factor, including the nature of the curriculum that the student experienced.

The FOSS tasks, as well as those that have appeared in national assessments such as NAEP, reflect an approach to assessment consistent

with a view of science learning as the disaggregated acquisition of content and practices. Indeed, in many classrooms, students are taught science based on such learning conceptions. They will encounter units on "the scientific process" *or* on "earthquakes and volcanoes." The application and coordination of scientific reasoning processes and practices to understanding the concepts associated with plate tectonics, however, is a much less common experience (Duschl, 2003).

The most recent NAEP science framework for the 2009 assessment represents an attempt at a more integrated view that values both the knowing and doing of science (see Figure 2). While the content strands from the earlier framework remain stable, the process categories have been significantly restructured (NAGB, 2006). However, even this organization does not capture the coordinated and integrated cognitive, socio-cultural, and epistemic components of scientific practice. The impact of this framework ultimately will be determined by the extent

FIGURE 2
NAEP ASSESSMENT MATRIX FOR 2009 ASSESSMENT

		Science Content		
		Physical Science content statements	**Life Science content statements**	**Earth & Space Science content statements**
Science Practices	**Identifying Science Principles**	*Performance Expectations*	*Performance Expectations*	*Performance Expectations*
	Using Science Principles	*Performance Expectations*	*Performance Expectations*	*Performance Expectations*
	Using Scientific Inquiry	*Performance Expectations*	*Performance Expectations*	*Performance Expectations*
	Using Technological Design	*Performance Expectations*	*Performance Expectations*	*Performance Expectations*

to which it will lead to substantively different tasks on the next NAEP assessment.

Emerging theories of science learning have benefited from a much clearer articulation of the development of reasoning skills, suggesting radically different instructional and assessment practices. Instructional implications have been represented in learning progressions (e.g., Quintana et al., 2004; Smith et al., 2006) describing the development of knowledge and reasoning skills across the curriculum within particular conceptual areas as students engage in the socio-cultural practices of science. Clarification of these progressions is critical, as current science curricular specifications and standards are seldom grounded in any understanding of the cognitive development of particular concepts or reasoning skills. These instructional sequences are responses to science curricula that have been criticized for their redundancy across years and their lack of principled progression of concept and skill development (Kesidou & Roseman, 2002).

A more integrated view of science learning is expressed in the recent NRC report articulating the future of science assessment (Wilson & Bertenthal, 2005). The report argues that science assessment tasks should reflect and encourage science activity that approximates the practices of actual scientists by embracing a socio-cultural perspective and the idea of legitimate peripheral participation, in which learning is viewed as increasingly participating in the socio-cultural practices of a community (Lave & Wenger, 1991). The NRC committee proposes models of assessment that engage students in sustained inquiries, sharing many of the social and conceptual characteristics of what it means to "do science." Instead of disaggregating process and content, assessment designs are proposed that integrate skills and understanding to provide information about the development of both conceptual knowledge and reasoning skill.

Despite progress in science learning theory, curricular models such as learning progressions, and assessment frameworks, developing instructional practice coherent with these visions is no simple task. Coherence requires curricular choices to be made so that a relatively small number of conceptual areas are targeted for study in any given school year. If sustained inquiry is to be taken seriously, as embodied in the work on learning progressions, then large segments of the existing curricular content will need to be jettisoned. It is impossible to envision a curriculum that pursues the knowing and doing of science as expressed in learning progressions also attempting to cover the very large number of topics that are now part of most curricula (Gitomer, in press).

The implications for large-scale assessment are profound as well. Assessing constructs such as inquiry requires going beyond the traditional content-lean approach described by Pine et al. (2006). Assessing the *doing* of science requires designs that are much more tightly embedded with particular curricula. Making the difficult curricula choices that allow for an instructional and assessment focus is the only way external coherence with learning theory can be achieved.

More complex underlying learning theories require suitable psychometric approaches that can model complex and integrated performances in ways that provide useful assessment information. Rather than assigning single scale scores, psychometric models are needed that can represent the multidimensional aspects of learning embodied in the previous discussion. For this, the authors look to work on evidence-centered design (ECD) by Mislevy and colleagues (Mislevy & Haertel, 2006; Mislevy, Hamel et al., 2003; Mislevy & Riconscente, 2005; Mislevy, Steinberg, & Almond, 2002).

Evidence-Centered Design (ECD)

ECD offers an integrated framework of assessment design that builds on principles of legal argumentation, engineering, architecture, and expert systems to fashion an *assessment argument*. An assessment argument involves defining the construct to be assessed; deciding upon the evidence that would reveal those constructs; designing assessments that can elicit and collect the relevant evidence; and developing analytic systems that interpret and report on the evidence as it relates to inferences about learning of the constructs.

ECD has been applied to science assessments in the project Principled Assessment Designs for Inquiry (PADI) (Mislevy & Haertel, 2006; Mislevy & Riconscente, 2005). A key part of this effort has been to develop *design patterns*, which are assessment design templates that, like engineering design components, are intended to serve recurring needs, but have variable attributes that are manipulated for specific problems. Thus, the PADI project has developed design patterns for model-based reasoning with specific patterns for such integrated practices as model formation, elaboration, use, articulation, evaluation, revision, and inquiry. Each of the patterns has a set of attributes, some of which are characteristic of all instances and some of which vary. Design pattern attributes include the rationale; focal knowledge skills and abilities; additional knowledge skills and abilities; potential observations; and potential work products. So, for example, a template for *model elaboration* would consider the completeness of a model as one important piece

of observational evidence. Of course, how completeness is defined will vary with the science content and the sophistication of the students. ECD methods can certainly be used to examine socio-cultural claims, as tools, practices, and activity structures can be articulated in the templates. Although to date most ECD examples have focused on knowledge and skills from a traditional cognitive perspective, Mislevy (2005, 2006) has described how ECD can be applied to socio-cultural dimensions of practice such as argumentation.

This large body of work suggests that a new generation of assessments is possible, one that could address accountability needs yet also support instructional practice consistent with current models of science learning. Popham, Keller, Moulding, Pellegrino, and Sandifer (2005) propose a model that includes relatively comprehensive assessment tasks based on a two-dimensional matrix that crosses important concepts (e.g., characteristic physical properties and changes in physical science) with science-as-inquiry skills (e.g., develop descriptions, explanations, predictions; critique models using evidence). Such assessments become viable if agreements can be made on a relatively limited set of concepts to be targeted within an assessment. Persistent efforts to cover broad swaths of content with limited depth constrain the likelihood that Popham et al.'s vision will be realized.

Designing Assessment Systems—Internal Coherence

Even with an externally coherent system responsive to emerging models of how people learn science, educational systems, like other complex institutional systems, must grapple with multiple and often conflicting messages. Nowhere has this tension been more evident than in the coordination of the policies and practices of accountability systems with the practices and goals for classroom instructional practice. Honig and Hatch (2004) discuss the problem as one of *crafting coherence*, in which they provide evidence for how local school administrators contend with state and district policies that are inconsistent with other policies, as well as with the goals they have for classroom practice within their local contexts. Importantly, Honig and Hatch note that contending with these inconsistencies does not always result in a solution in which the various pieces fit together in a conceptually coherent model. Indeed, administrators often decide that an optimal solution is to avoid trying to bring disparate policies and practices into alignment. As Spillane (2004) has noted, there are also instances in which administrators simply ignore the conflict, despite its unsettling consequences for the classroom teacher.

The concept of crafting coherence can be applied generally to the coordination of assessment policies and practices. The tension between what is currently conceived of as assessment *of* learning (accountability assessment) with assessment *for* learning (formative classroom assessment) (Black & Wiliam, 1998) has been addressed by a variety of coherence models in the United States and abroad. We briefly review these models with examples, and summarize some of the outcomes associated with each of these potential solutions. We attempt to provide a perspective that characterizes prototypical features of these systems while recognizing at the same time that there have been, and will continue to be, schools and districts that have developed atypical but exemplary practices.

Independent Co-Existence

This represents what was long the traditional practice in U.S. schools, characterized by the idea that schools administered standardized assessments to meet accountability functions while not viewing them as particularly relevant to classroom learning. In fact, schools were often dismissive of these tests as irrelevant bureaucratic necessities. Certainly for many years accountability tests had very little impact on schools and educators, although the public held these tests in higher regard.

However, the lack of forceful accountability testing was not accompanied by particularly strong assessment practices in classrooms either. Whether formal classroom tests or teacher questions designed to uncover student insight, practice was characterized by questioning that required the recall of isolated conceptual fragments. Instances of eliciting, analyzing, and reporting student conceptual understanding and skill development were uncommon (see Gitomer & Duschl, 1998 for more details).

Isomorphic Coherence

With the passage of NCLB in 2001, independent co-existence was no longer viable. Isomorphic coherence builds on the idea that teaching to the test is a good thing if the test is designed to assess and encourage the development of knowledge and skills worth knowing (Frederiksen & Collins, 1989; Resnick & Resnick, 1991)—logic that has been embraced by testing and test-preparation companies and school districts alike.

The general approach involves publishers developing large banks of test items of the same format and content as items appearing on the

accountability tests. Students spend significant instructional time practicing these items and are administered *benchmark* tests during the year to help teachers and administrators gauge the likelihood of their meeting the passing (proficiency) standard set by the respective state. The net result is an internally coherent system in which the overlap between classroom practice and accountability testing is very significant.

The merit of this type of coherence has been argued vociferously. Advocates argue that such alignment provides the best opportunity for preparing all students to meet a set of shared expectations and for reducing long-standing educational inequities reflected in the achievement gap (e.g., National Center for Educational Accountability, 2006). Critics argue that this alignment has adverse effects on student learning, because of the inadequacy of the current generation of standardized tests in assessing and encouraging the development of knowledge and skills worth knowing (e.g., Amrein & Berliner, 2002a). In science education, critics are concerned that the current accountability tests reflect a limited and unscientific view and that preparing for such tests is a poor expenditure of educational resources. The socio-cultural dimensions of science learning are virtually ignored in these kinds of systems. Thus, even though they are internally coherent, these systems lack external coherence because of their lack of connection with theories of science learning.

In response to this criticism, Popham et al. (2005) propose a system, described earlier, in which accountability tests are constructed from tasks that are much more consistent with cognitive models of learning and performance. They propose tasks that are drawn from a greatly reduced set of curricular aims, are consistent with learning theory, and are transparent and readily understood by teachers. Inherent to the Popham et al. approach is an instructional system featuring a curriculum that lines up with the recommendations of Wilson and Bertenthal (2005).

Organic Accountability

Organic models are ones in which the assessment data are derived directly from classroom practice. The clearest examples of organic accountability are the variety of portfolio systems that emerged during the 1980s (e.g., Koretz, Stecher, & Deibert, 1992; Wolf, Bixby, Glenn, & Gardner, 1991). Portfolio systems were developed to respond to the traditional disconnect between accountability and classroom assessment practices. The logic behind these systems was that disciplined judgments could be made about student work products on a common set of

broad dimensions, even when the work differed significantly in content. In education, these kinds of judgments had long been applied to art shows, science fairs, and musical competitions.

Perhaps the most ambitious system was the exhibition model developed by the Coalition of Essential Schools (CES) (McDonald, 1992). In this model, high school students developed a series of portfolios to provide cumulative evidence of their accomplishment with respect to a set of primary educational objectives. One CES high school set objectives such as communicating, crafting, and reflecting; knowing and respecting myself and others; connecting the past, present, and future; thinking critically and questioning; and values and ethical decision making. For each objective, potential evidence was described. For example, potential evidence for *connecting the past, present, and future* included:

- Students develop a sense of time and place within geographical and historical frameworks.
- Students show that they understand the role of art, music, culture, science, math, and technology in society.
- Students relate present situations to history, and make informed predictions about the future.
- Students demonstrate that they understand their own roles in creating and shaping culture and history.
- Students use literature to gain insight into their own lives and areas of academic inquiry. (CES National Web, 2002).

Portfolios based on these objectives were then shared, and an oral presentation was made to an audience of faculty, other students, and external observers. Often, students needed to further develop their portfolio to satisfy the criteria for success. Quite apparent in these portfolio requirements is the dominant focus on the socio-cultural dimensions of learning.

Ironically, the strength of the organic system also led to its virtual demise as an accountability mechanism. When assessment evidence is derived from classroom practice, student achievement cannot be partitioned from the opportunities students have been given to demonstrate learning. Portfolio data provides a window into what teachers expect from students and what kinds of opportunities students have had to learn. To many, true accountability requires an examination of opportunity to learn (Gitomer, 1991; Shepard, 2000). LeMahieu, Gitomer, and Eresh (1995) demonstrated how district-wide evaluations of portfolios could shed light on educational practice in writing classrooms.

Koretz et al. (1992) concluded that statewide portfolios were more valuable in providing information about educational practice than they were in satisfying the need for making judgments about whether a particular student had achieved at a particular level.

Indeed, the variability in student evidence contained in the portfolios made it very difficult to make judgments about the relative learning and achievement of individual students. Had a student been asked to provide different evidence or held to different expectations by the teacher, the portfolio of the very same student might have looked radically different. And the fact that the portfolio made these differences in opportunity so much more transparent than did traditional "drop-in from the sky" (Mislevy, 1995) assessments also challenged the ability to provide assessment information that met psychometric standards.

The desirability of organic systems has much to do with perceptions of accountability (cf. Shepard, 2000), as well as whether there is sufficient trust in the quality of information yielded by the organic system (e.g., Koretz et al., 1992). Certainly the dominant perspective today is to provide individual scores that meet standards of psychometric quality. This has led, in the age of NCLB, to the virtual abandonment of organic models as a source of accountability.

Organic Hybrids

These hybrid models are ones in which accountability information is drawn from both classroom performance and external high-stakes assessments. Major attempts at operational hybrids include the California Learning Assessment System (California Assessment Policy Committee, 1991), the New Standards Project (1997), and the Task Group on Testing and Assessment in the United Kingdom (Nuttall & Stobart, 1994). These efforts all included classroom generated portfolio evidence along with more standardized assessment components.[3] The impetus was to combine the broad evidence captured by the portfolio with more psychometrically defensible traditional assessments in order to represent both the cognitive and socio-cultural dimensions of learning.

In each case, the portfolio effort withered for a combination of reasons. First, as was true for organic approaches, the "opportunity to learn" impact on portfolio outcomes made inferences about the student inescapably problematic (Gearhart & Herman, 1998). Second, when there was conflicting information from the two sources of evidence, standardized assessment evidence inevitably trumped portfolio evidence

(e.g., Koretz, Stecher, Klein, & McCaffrey, 1994). Despite the fact that the two evidence sources were oriented toward different types of information, the quality of evidence was judged as if they were offering different lenses on the same information. This inevitably put the portfolio in a bad light, because it is a much less effective mechanism for determining whether students know specific content and/or skills, although it has the potential to reveal how well students can perform legitimate domain tasks while making use of content and skills. Finally, the portfolio emphasis decreased because of financial, operational, and sometimes political constraints (Mathews, 2004).

An Alternative: The Parallel Model

Taken together, each of the models discussed above has failed to become a scalable assessment system consistent with desired learning goals because it fell short on at least one but typically several of the criteria that are critical for such a system:

- theoretical symmetry or external coherence (models with an impoverished view of the learner);
- internal coherence between different parts of the assessment system (models in which the summative and formative components of the system are not aligned);
- pragmatics of implementation (models that are unwieldy and too costly); and
- flow of information among the stakeholders in the system (models in which inconsistent messages about what is valued are communicated between stakeholders).

In this section, we outline the characteristics of a system that can be externally and internally coherent, which aligns with the conceptual work that has been presented in Wilson and Bertenthal (2005), Popham et al. (2005), and Pellegrino et al. (2001). Their work, among others, describes assessment systems that can be externally coherent by including cognitive structures, scientific reasoning skills, and socio-cultural practices in integrated assessment activities.

However, we argue that, in order for such assessment systems to be internally coherent and scalable, far more attention needs to be paid to issues of pragmatics and information flow than has been the case in discussions of future assessment design. Pragmatic aspects of assessment refer to tractable solutions to existing constraints. The model we propose does not assume a radical restructuring of schools or policy.

Our attempt is to put forth a system that can significantly improve assessment practice within the current educational environment.

We begin with a set of assumptions about the design of an assessment system that includes components to be used for both accountability purposes and in classrooms. While this is sometimes referred to as a summative/formative dichotomy, it is our intention that information for policymakers ought to be used to shape instructionally related policy decisions and therefore, serve a formative role at the district and state levels as well.

The two components are separate, yet parallel in nature. By separate, we accept the premise (e.g., Mislevy et al., 2002), that different assessments have different purposes, and that those purposes should drive the architecture of the assessment. Trying to satisfy both formative and summative needs is bound to compromise one or both systems. Accountability instruments are designed to provide summary information about the achievement status of individuals and institutions (e.g., schools) and are not well suited for supporting particular diagnoses of students' needs, which ought to be the province of classroom-based assessments and formative classroom tools.

Requirements

Nevertheless, the systems need to be parallel in two important ways. They need to be built on the same underlying theory of learning. In science, this means a theory that takes into account cognitive, sociocultural, and epistemic aspects of learning. They also need to share, in large part, common task structures. The summative assessment ought to provide models of assessment tasks that are designed to support ambitious models of learning.

A further assumption is that the majority of assessment tasks will be constructed-response. If the goal is to gauge students' abilities to generate explanations, provide representations, model data, and otherwise engage in various aspects of inquiry, they must show evidence of "doing science."

The next assumption is that there will be an agreed upon focus on major scientific curricular goals, as argued by Popham et al. (2005)—a circumstance requiring substantial changes in educational practice in the United States. There does seem to be an emerging consensus for the first time, however, that this narrowing and deepening of the curriculum is the appropriate road for the future of science education (e.g., Wilson & Bertenthal, 2005).

A final assumption is that the assessment design, psychometric analysis, and reporting of results will be consistent with the underlying learning models; that is, that they will provide information to all stakeholders to make the model of science learning transparent. Reports will go beyond providing a scalar indicator to providing descriptions of student performance that are meaningful status reports with respect to identified learning goals.

Constraints

Even if richer theories of science learning were embraced, and curricular objectives became more widely shared and focused, there remain two powerful constraints that can inhibit the development of a coherent assessment system. The first is time. While accountability testing time varies across grades and states, the typical practice is that subject matter testing consists of a single event of one to three hours. Once such a constraint is in place, the options for assessment design decrease dramatically. If one moves to a large proportion of constructed-response tasks, it becomes highly problematic to sample the entire domain.[4]

The second constraint is cost. Most systems that use constructed-response tasks rely on human raters, which has made the cost of scoring these tasks very daunting (Office of Technology Assessment, 1992; Wainer & Thissen, 1993; Wheeler, 1992). If we are to move to an assessment system with a very high preponderance of constructed-response tasks, the cost issue must be confronted.

Researchers at the Educational Testing Service (ETS) are currently working on an accountability system model that addresses these two constraints directly. Time issues are mitigated by multiple administrations of the accountability assessment during the school year. Each administration consists of an assessment module involving integrated tasks that are externally coherent. With multiple administrations, it now becomes possible to include complex tasks consistent with models of learning that will also yield psychometrically defensible information.

Of course, this model also involves significantly more testing, which is apt to be criticized. Acknowledging the concern about overtesting our youth, there are several important potential advantages of proceeding in this way. First, if the assessment tasks are truly worthy of being targets of instruction, then the assessments and preparation for them can be valuable. The second advantage to the distributed model is that students and teachers are able to gauge progress over the course of the year, rather than wait for results from a one-time, end-of-year admin-

istration. A third advantage being considered is the opportunity for students to retake alternate forms of particular modules to demonstrate accomplishment. If educational policy calls for a model in which students truly do not get left behind, then it seems reasonable for students to continue to work to meet the performance objectives set forth by the system.

We plan to address the cost constraint through rapid progress being made in the development of automated scoring engines for constructed-response tasks (e.g., Foltz, Laham, & Landauer, 1999; Leacock & Chodorow, 2003; Shermis & Burstein, 2003; Williamson, Mislevy, & Bejar, 2006), which offer the potential to drastically decrease the cost differential between item formats that is primarily attributable to the cost of human scoring. It is important to note that although automated tools can be used to support teachers in classrooms, these scoring approaches are concentrated primarily in supporting accountability testing. We envision teachers using good assessment tasks to structure classroom interactions to provide rich information about student understanding. However, the teacher would be responsible for management and analysis of this assessment information—control would not be handed off to any automated systems. The current state of technology requires that automatically scored assessments be administered via computer, typically increasing test administration costs. But as computing resources become ubiquitous in schools, and as administration occurs over the Internet, those cost differentials should continue to decline, even to the point where computer delivery is less costly than all of the logistical costs associated with paper-and-pencil testing.

With these constraints addressed, we envision the accountability portion of the assessment to be structured as seen in Figure 3. Several aspects are worthy of note. Over the course of the school year, the accountability assessment is administered under relatively standardized conditions in a series of periodic assessments. These assessments are designed in light of a domain model that is defined by learning research, as well as their intersection with state standards. Results from these tasks are reported to various stakeholders at appropriate levels of granularity. Students, parents, and teachers receive information that reflects specific profiles of individual students. Different levels of aggregated information are provided to teachers and school and district administrators to support their respective decision making requirements, including decisions about professional development and instructional/curricular policy. The results are then aggregated up to meet state-level accountability

FIGURE 3
The Accountability Component of a Coherent Assessment System

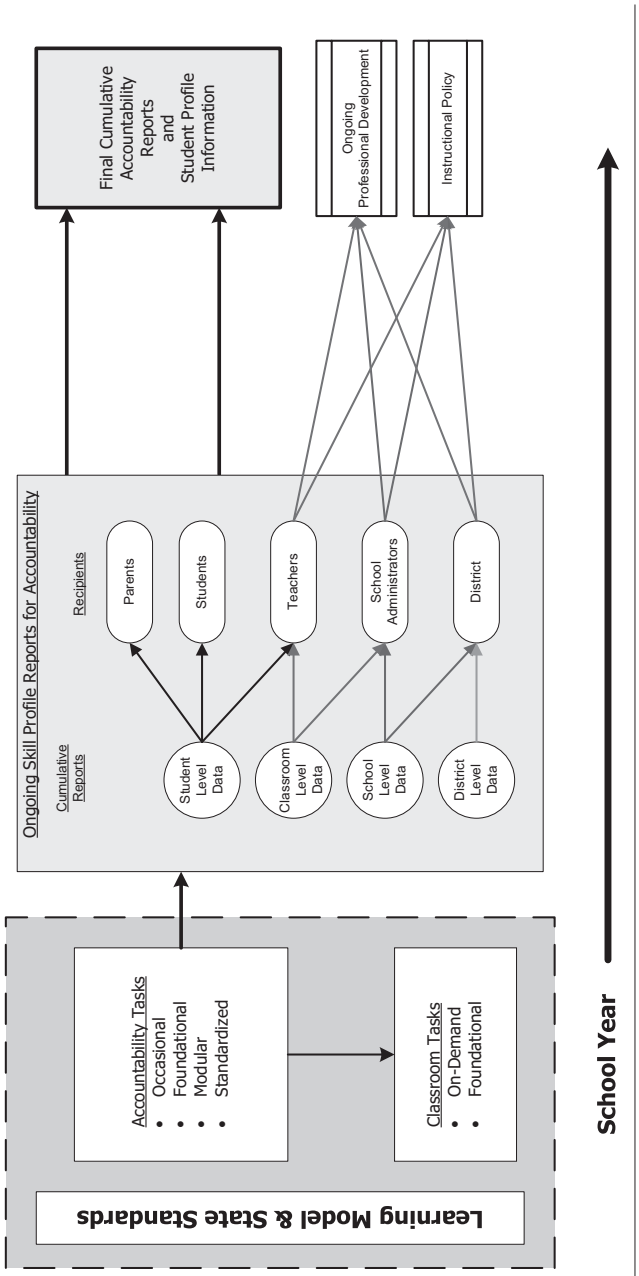
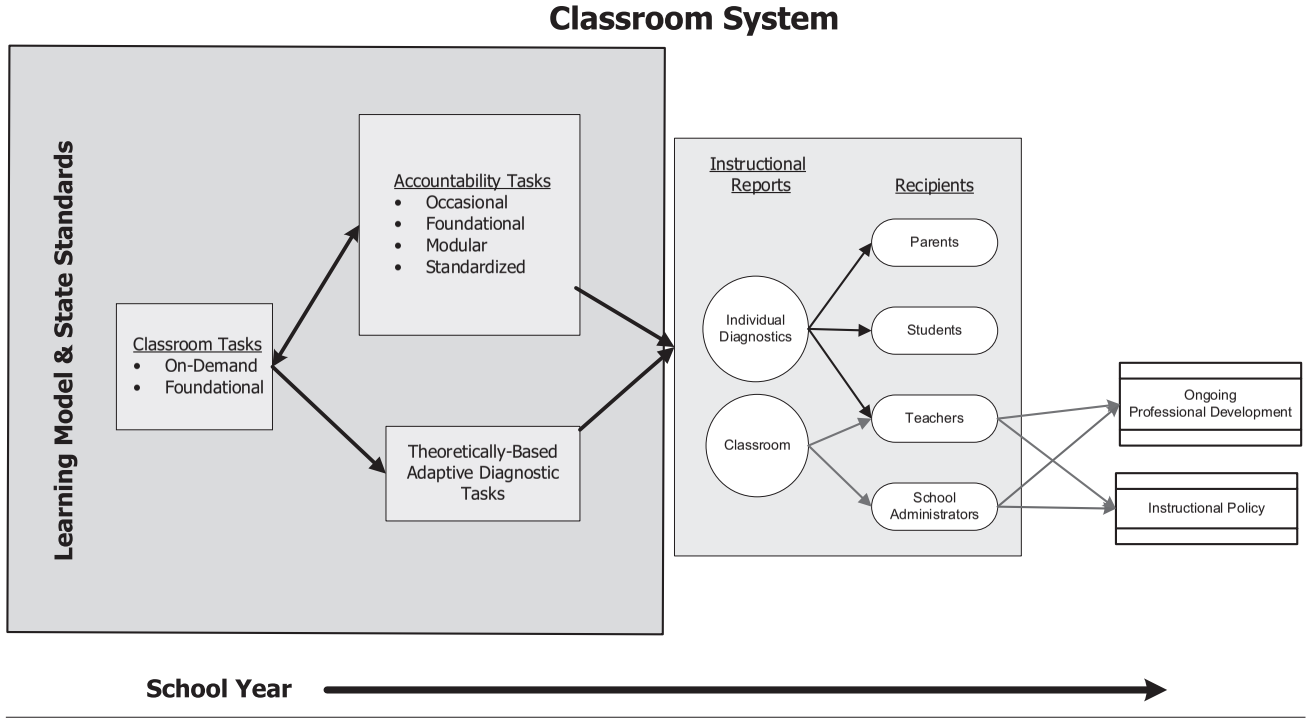


FIGURE 4
THE CLASSROOM COMPONENT OF A COHERENT ASSESSMENT SYSTEM



demands. At all levels of the system, however, the same underlying learning model, in consideration of state standards, is operative. Reports will be designed to enhance the likelihood that educators at all levels of the system are working within the same framework of student learning, a condition that is not typically found in schools (Spillane, 2004) or supported by evidence in the system (Coburn et al., in press).

The parallel classroom system is presented in Figure 4. The same underlying model of learning, contributing to internal coherence, also drives this system. However, specific classroom tasks are invoked for particular students, as determined by the teacher on the basis of accountability test performance as well as his or her professional judgment. Tasks include integrated tasks that are foundational to the domain, as well as tasks that may be targeted at clarifying specific aspects of student understanding or performance. The information from the formative system is used only to support local instructional decision making—it provides no information to the parallel but separate accountability system.

Challenges to the Parallel System

Certainly, realizing the vision of the parallel system presents numerous challenges, many of which have been identified throughout the chapter. These include clarification of the underlying learning model and making deliberate curricular choices for focus. Fully solving the pragmatic constraints will be nontrivial as well. Implementing a distributed system will require substantial changes for teachers, schools, and districts. In order to make this work, the perceived payoff will have to seem worth the effort. Solving the cost issue for scoring is not a given either.

While tremendous progress has been made in automated processing of text and other representations, there is still much progress to be made in order to have a fully defensible and acceptable automated scoring system that can be used in high-stakes accountability settings. There are numerous psychometric issues, as well, involved in the aggregation of assessment information over time, the impact of curricular implementation on assessment module sequencing, the interpretation of results under different sequencing conditions, and the handling of retesting. However, if we can successfully address these issues, we have the potential to support decision making throughout the educational system that is based on valid assessments of valued dimensions of student learning.

AUTHORS' NOTE

The authors are grateful for the very helpful reviews from Pamela Moss, Phil Piety, Valerie Shute, Iry Katz, and several anonymous reviewers.

NOTES

1. Our approach is to accept the basic assumptions of NCLB and propose a system that can meet those assumptions, while also contributing to effective teaching and learning. Therefore, we do not challenge the idea of each student receiving an individual score in the assessment system. Nor do we challenge the basic premise of large-scale standardized testing as the primary instrument in the accountability process. Certainly, provocative challenges and alternatives have been raised, but we do not pursue those directions in this chapter.

2. Research and development work in building these systems is currently being pursued at Educational Testing Service.

3. Note that systems such as those used in Queensland, Australia (Queensland School Curriculum Council, 2002) include classroom-generated information in judgments of educational achievement. However, these models conduct audits of schools that sample performance to ensure that standards are being interpreted as intended. This type of model does not attempt to merge the different sources of information about achievement into a unified assessment program.

4. Another strategy to reduce cost and testing time is to use matrix sampling, in which any one student is tested on a relatively small portion of the assessment design. While matrix sampling is useful for making inferences about groups of students, it cannot be used to assign unique scores to individuals and is not acceptable under the provisions of NCLB.

REFERENCES

Abrams, L.M., Pedulla, J.J., & Madaus, G.F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice*, *42*(1), 8–29.

Amrein, A.L., & Berliner, D.C. (2002a, March 28). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18). Retrieved September 12, 2006, from http://epaa.asu.edu/epaa/v10n18/.

Amrein, A.L., & Berliner, D.C. (2002b, December). An analysis of some unintended and negative consequences of high-stakes testing. Education Policy Research Unit, Arizona State University, Tempe. Retrieved September 6, 2006, from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-125-EPRU.pdf.

Anderson, J.R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison: University of Wisconsin Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–73.

Bransford, J., Brown, A., & Cocking, R. (Eds.). (1999). *How people learn: Brain, mind, experience and school*. Washington, DC: National Academy Press.

California Assessment Policy Committee (1991). *A new student assessment system for California schools* (Executive Summary Report). Sacramento, CA: Office of the Superintendent of Instruction.

CES National Web (2002). A richer picture of student performance. Retrieved October 2, 2006, from Coalition of Essential Schools web site http://www.essentialschools.org/pub/ces_docs/resources/dp/uhhs.html.

Chase, W.G., & Simon, H.A. (1973). The mind's eye in chess. In W.G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152.

Coburn, C.E., Honig, M.I., & Stein, M.K. (in press). What is the evidence on districts' use of evidence? In J. Bransford, L. Gomez, N. Vye, & D. Lam (Eds.), *Research and practice: Towards a reconciliation*. Cambridge, MA: Harvard Educational Press.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Duschl, R. (2003). Assessment of scientific inquiry. In J.M. Atkin & J. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41–59). Arlington, VA: NSTA Press.

Duschl, R., & Gitomer, D. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Education Assessment*, *4*(1), 37–73.

Duschl, R., & Grandy, R. (Eds.). (2007). *Establishing a consensus agenda for K-12 science inquiry*. The Netherlands: SensePublishers.

Duschl, R., Schweingruber, H., & Shouse, A. (Eds.). (2006). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academy Press.

Erduran, S. (1999). *Merging curriculum design with chemical epistemology: A case of teaching and learning chemistry through modeling*. Unpublished doctoral dissertation, Vanderbilt University, Nashville, TN.

Foltz, P.W., Laham, D., & Landauer, T.K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2). Retrieved January 8, 2006, from imej.wfu.edu/articles/1999/2/04/index.asp.

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

Gearhart, M., & Herman, J.L. (1998). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability. *Educational Assessment*, *5*(1), 41–55.

Gee, J. (1999). *An introduction to discourse analysis: Theory and method*. New York: Routledge.

Gitomer, D.H. (1991). The art of accountability. *Teaching Thinking and Problem Solving*, *13*, 1–9.

Gitomer, D.H. (in press). Policy, practice and next steps for educational research. In R. Duschl & R. Grandy (Eds.), *Establishing a consensus agenda for K-12 science inquiry*. The Netherlands: SensePublishers.

Gitomer, D.H., & Duschl, R. (1998). Emerging issues and practices in science assessment. In B. Fraser & K. Tobin (Eds.), *International handbook of science education* (pp. 791–810). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Glaser, R. (1976). Components of a psychology of instruction: Toward a science of design. *Review of Educational Research*, *46*, 1–24.

Glaser, R. (1991). The maturing of the relationship between the science of learning and cognition and educational practice. *Learning and Instruction*, *1*(2), 129–144.

Glaser, R. (1992). Expert knowledge and processes of thinking. In D.F. Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics* (pp. 63–75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R. (1997). Assessment and education: Access and achievement. *CSE Technical Report 435*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Glaser, R., & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 20, pp. 393–419). Washington, DC: American Educational Research Association.

Greeno, J.G. (2002). *Students with competence, authority, and accountability: Affording intellective identities in classrooms*. New York: College Board.

Honig, M., & Hatch, T. (2004). Crafting coherence: How schools strategically manage multiple, external demands. *Educational Researcher*, *33*(8), 16–30.

Kesidou, S., & Roseman, J.E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching*, *39*(6), 522–549.

Koretz, D., Stecher, B., & Deibert, E. (1992). *The reliability of scores from the 1992 Vermont portfolio assessment program*. Los Angeles, CA: RAND Institute on Education and Training.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5–16.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short answer questions. *Computers and the Humanities*, *37*(4), 389–405.

LeMahieu, P.G., Gitomer, D.H., & Eresh, J.T. (1995). Large-scale portfolio assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice*, *14*, 11–28.

Magone, M., Cai, J., Silver, E.A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research*, *12*(3), 317–340.

Mathews, J. (2004). Whatever happened to portfolio assessment? *Education Next*, *3*. Retrieved October 12, 2006, from http://www.hoover.org/publications/ednext/3261856.html.

McDonald, J. (1992). *Teaching: Making sense of an uncertain craft*. New York: Teachers College Press.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Mislevy, R.J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis*, *17*(4), 419–437.

Mislevy, R.J. (2005). *Issues of structure and issues of scale in assessment from a situative/sociocultural perspective* (CSE Report 668). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education/Praeger.

Mislevy, R.J., & Haertel, G. (2006). *Implications of evidence-centered design for educational testing* (Draft PADI Technical Report 17). Menlo Park, CA: SRI International.

Mislevy, R.J., Hamel, L., Fried, R., Gaffney, T., Haertel, G., Hafter, A., et al. (2003). *Design patterns for assessing science inquiry*. Menlo Park, CA: SRI International.

Mislevy, R.J., & Riconscente, M.M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). Menlo Park, CA: SRI International.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

National Assessment Governing Board (NAGB) (1996). *Science framework for the 1996 and 2000 National Assessment of Educational Progress*. U.S. Department of Education. Washington, DC: The Department. Retrieved October 22, 2006, from http://www.nagb.org/pubs/96-2000science/toc.html.

National Assessment Governing Board (2006). *NAEP 2009 science framework*. Washington, DC: Author.

National Center for Educational Accountability (2006). Available at http://www.just4kids.org/jftk/index.cfm?st=US&loc=home.

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

National Research Council (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academy Press.

National Research Council (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. Committee on Programs for Advanced Study of Mathematics and Science in American High Schools. J.P. Gollub, M.W. Bertenthal, J.B. Labov, & P.C. Curtis (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

New Standards Project (1997). *New standards performance standards* (Vol. 1, Elementary School; Vol. 2, Middle School; Vol. 3, High School). Washington, DC: National Center on Education and the Economy and the University of Pittsburgh.

Nuttall, D.L., & Stobart, G. (1994). National curriculum assessment in the U.K. *Educational Measurement: Issues and Practice*, *13*(2), 24–27.

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*. OTA-SET-519. Washington, DC: U.S. Government Printing Office.

Pellegrino, J.W., Baxter, G.P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P.D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 307–353). Washington, DC: American Educational Research Association.

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

Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C. et al. (2006). Fifth graders' science inquiry abilities: A comparative study of students in hands-on and textbook curricula. *Journal of Research in Science Teaching*, *43*(5), 467–484.

Popham, W.J., Keller, T., Moulding, B., Pellegrino, J., & Sandifer, P. (2005). Instructionally supportive accountability tests in science: A viable assessment option? *Measurement: Interdisciplinary Research and Perspectives*, *3*(3), 121–179.

Queensland School Curriculum Council (2002). *An outcomes approach to assessment and reporting*. Queensland, Australia: Author.

Quintana, C., Reiser, B.J., Davis, E.A., Krajcik, J., Fretz, E., Duncan, R.G., et al. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences*, *13*(3), 337–386.

Resnick, L.B., & Resnick, D.P. (1991). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37–75). Boston: Kluwer.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. New York: Oxford University Press.

Roseberry, A., Warren, B., & Contant, F. (1992). Appropriating scientific discourse: Findings from language minority classrooms. *The Journal of the Learning Sciences*, *2*, 61–94.

Shavelson, R., Baxter, G., & Pine, J. (1992). Performance assessment: Political rhetoric and measurement reality. *Educational Researcher*, *21*, 22–27.

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

Shermis, M.D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Smith, C., Wiser, M., Anderson, C., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, *4*(1&2), 1–98.

Spillane, J. (2004). *Standards deviation: How local schools misunderstand policy*. Cambridge, MA: Harvard University Press.

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

Vygotsky, L.S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103–118.

Webb, N.L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. National Institute for Science Education and Council of Chief State School Officers Research Monograph No. 6. Washington, DC: Council of Chief State School Officers.

Webb, N.L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research monograph No. 18). Madison: University of Wisconsin-Madison, National Institute for Science Education.

Wheeler, P.H. (1992). *Relative costs of various types of assessments*. Livermore, CA: EREAPA Associates (ERIC Document No. ED 373074).

Williamson, D.M., Mislevy, R.J., & Bejar, I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability. The one hundred and third yearbook of the National Society for the Study of Education*, Part II. Chicago: National Society for the Study of Education.

Wilson, M., & Bertenthal, M. (Eds.). (2005). *Systems for state science assessment*. Washington, DC: National Academies Press.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of educational research* (Vol. 17, pp. 31–74). Washington, DC: American Educational Research Association.