Center for
K–12 Assessment
& Performance Management

An independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Exploratory Seminar:

Measurement Challenges Within
the Race to the Top Agenda

December 2009

The Design of an Assessment System for the Race to the Top: A Learning Sciences Perspective on Issues of Growth and Measurement

James W. Pellegrino

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

The Design of an Assessment System for the Race to the Top: A Learning Sciences Perspective on Issues of Growth and Measurement

James W. Pellegrino

University of Illinois at Chicago

This paper was presented by James W. Pellegrino at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of other papers presented at the seminar at http://www.k12center.org/publications.html.

This paper discusses some of the most important conceptual, practical, and policy issues surrounding the *what* and *how* of assessment in K–12 education, especially in the context of current federal and state initiatives related to the Race to the Top. The general argument developed in this paper is that the learning sciences have made great strides in understanding the nature of learning and knowing, in ways that can and must inform the development of systems of assessments and the selection and use of models for measuring the status and growth of academic achievement in core areas of the curriculum. In developing this argument, the paper discusses a set of interconnected and critical issues regarding the nature and function assessment in the larger educational enterprise.

The first section of the paper is concerned with two critical issues in the development, implementation, and use of assessments for any group of K–12 students. The first issue is that assessment should never be the tail that wags the educational dog. Rather, assessment should be integrated with curriculum and instruction, with all three guided by theories and research on the nature of learning and knowing in academic content domains. A second fundamental issue is that we can never really know what a student knows. Thus, assessment is best conceptualized as a rigorous and carefully structured process of reasoning from evidence that should be driven by theories and data on student cognition and learning.

The second section of the paper then elaborates a key aspect of the argument developed in section one—it focuses on some of what theory and research have to say about student knowledge and performance in subject matter domains. A discussion is provided of knowledge development in instructional domains with a particular emphasis on the concept of *learning progressions*. The latter should be used to shape the design of assessments of student learning and they have implications for contexts of application that span classrooms to state and national testing programs. Illustrations are provided of work on learning progressions in mathematics and science.

Some of the broader implications of research and theory from the learning sciences for assessment design and use are then explored in the third section of this paper. Topics discussed include the functions and purposes of assessment and the timescales of learning which they are intended to represent. Specific attention is given to implications for the design of classroom assessments and large-scale assessments, as well as the application of specific models of measurement.

The final section of this paper considers why a coordinated and balanced system of assessments is needed to accomplish the goals of educational improvement. To meet the information needs of individuals whose responsibilities range from classroom teaching and learning to district, state, and national policymaking, different types of assessments must be developed and implemented within a larger systemic structure. Assessments at each level of the system should use approaches aligned to the scientific knowledge base on student cognition and learning and they should be appropriately designed for particular levels of use, with real clarity about the functions served and information needs of users at each level.

Two Critical Issues in Conceptualizing Student Assessment

The Curriculum-Instruction-Assessment Triad

Assessment does not and should not stand alone in the educational system. Rather, it is one of three central components—curriculum, instruction, and assessment—as shown in Figure 1. The three elements of this triad are linked, although the nature of their linkages and reciprocal influence is often less explicit than it should be. Furthermore, the separate pairs of connections are often inconsistent in practice, which can lead to an overall incoherence in the educational enterprise.

Curriculum consists of the knowledge and skills in subject matter areas that teachers teach and students are supposed to learn. The curriculum generally consists of a scope or breadth of content in a given subject area and a sequence for learning. Content standards in a subject matter area typically outline the goals of learning, whereas curriculum sets forth the more specific means to be used to achieve those ends. *Instruction* refers to methods of teaching and the learning activities used to help students master

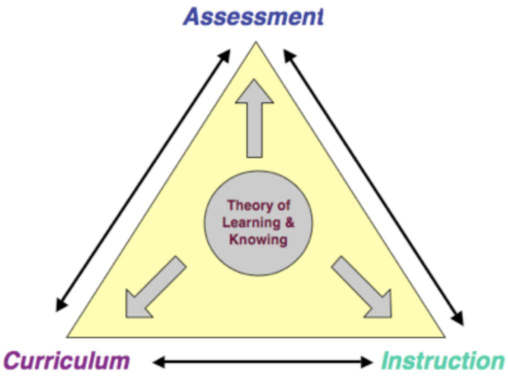

Figure 1. Representation of the interconnections among curriculum, instruction, and assessment and pivotal role of theories of learning.

the content and objectives specified by a curriculum. Instruction encompasses the activities of both teachers and students. It can be carried out by a variety of methods, sequences of activities, and topic orders. *Assessment* is the means used to measure the outcomes of education and the achievement of students with regard to important competencies. Assessment may include both formal methods, such as large-scale state assessments, or less formal classroom-based procedures, such as quizzes, class projects, and teacher questioning.

A precept of educational practice is the need for alignment among curriculum, instruction, and assessment (e.g., National Council of Teachers of Mathematics [NCTM], 1995, 2000; Webb, 1997). Alignment, in this sense, means that the three functions are directed toward the same ends and reinforce each other, rather than working at cross-purposes. Ideally, an assessment should measure what students are actually being taught, and what is actually being taught should parallel the curriculum one wants students to master. If any of the functions is not well synchronized, it will disrupt the balance and skew the educational process. Assessment results will be misleading, or instruction will be ineffective. Alignment is difficult to achieve, however. Often what is lacking is a central theory about the nature of learning and knowing, around which the three functions can be coordinated, as shown in Figure 1.

Most current approaches to curriculum, instruction, and assessment are based on theories and models that have not kept pace with modern knowledge of cognition and how people learn (e.g., Bransford, Brown, & Cocking, 1999; Bransford, Brown, Cocking, Donovan, & Pellegrino, 2000; Donovan & Bransford, 2005; Donovan, Bransford, & Pellegrino, 1999; Pellegrino, Chudowsky, & Glaser, 2001: Pellegrino, Jones, & Mitchell, 1999; Shepard, 2000). They have been designed on the basis of implicit and highly limited conceptions of cognition and learning. Those conceptions tend to be fragmented, outdated, and poorly delineated for domains of subject matter knowledge. Alignment among curriculum, instruction, and assessment could be better achieved if all three are derived from a scientifically credible and shared knowledge base about cognition and learning in subject matter domains. The model of learning would provide the central bonding principle, serving as a nucleus around which the three functions would revolve. Without such a central core, and under pressure to prepare students for high-stakes accountability tests, teachers may feel compelled to move back and forth between instruction and external assessment and teach directly to the items on a state test. The latter approach, in which assessment serves as the tail wagging the educational dog, can result in an undesirable narrowing of the curriculum and a limiting of learning outcomes. Such problems can be ameliorated if, instead, decisions about both instruction and assessment are guided by models of learning in academic domains that represent the best available scientific understanding of how people learn (Bransford et al., 2000; Donovan & Bransford, 2005).

Assessment as a Process of Reasoning From Evidence

Educators assess students to learn about what they know and can do, but assessments do not offer a direct pipeline into a student's mind. Assessing educational outcomes is not as straightforward as measuring height or weight; the attributes to be measured are mental representations and processes that are not outwardly visible. Thus, an assessment is a tool designed to observe students' behavior and produce data that can be used to draw reasonable inferences about what students know. Deciding what to assess and how to do so is not as simple as it might appear.

The process of collecting evidence to support inferences about what students know represents a chain of reasoning from evidence about student learning that characterizes all assessments, from classroom quizzes and standardized achievement tests, to computerized tutoring programs, to the conversation a student has with her teacher as they work through a math problem or discuss the meaning of a text. In the 2001 report, *Knowing What Students Know: The Science and Design of Educational Assessment*, issued by the National Research Council, the process of reasoning from evidence was portrayed as a triad of three interconnected elements: the *assessment triangle* (Pellegrino et al., 2001). The vertices of the assessment triangle (see Figure 2) represent the three key elements underlying any assessment: a model of student *cognition* and learning in the domain of the assessment; a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies; and an *interpretation* process for making sense of the evidence. These three elements may be explicit or implicit, but an assessment cannot be designed and implemented without consideration of each. The three are represented as vertices of a triangle because each is connected to and dependent on the other two. A major tenet of the *Knowing What Students Know* report is that for an assessment to be effective and valid, the three elements must be in synchrony. The assessment triangle provides a useful framework for analyzing the underpinnings of current assessments to determine how well they accomplish the goals we have in mind, as well as for designing future assessments and establishing validity (e.g., see Marion & Pellegrino, 2006).

The *cognition* corner of the triangle refers to theory, data, and a set of assumptions about how students represent knowledge and develop competence in a subject matter domain (e.g., fractions). In any particular assessment application, a theory of learning in the domain is needed to identify the set of knowledge and skills that is important to measure for the context of use, whether that be characterizing the competencies students have acquired at some point in time to make a summative judgment, or for making a formative judgment to guide subsequent instruction so as to maximize learning. A central

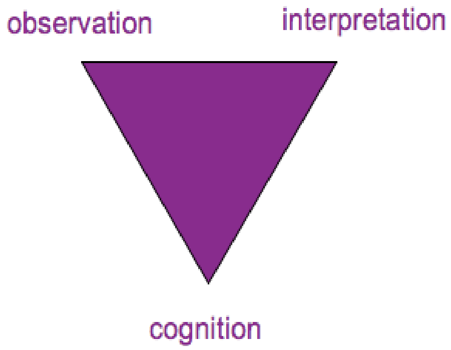

Figure 2. The three elements involved in conceptualizing assessment as a process of reasoning from evidence.

premise is that the cognitive theory should represent the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in a domain. More will be said in the next section about ways in which we currently think about cognition and the development of subject matter competence.

Every assessment is also based on a set of assumptions and principles about the kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates important knowledge and skills. The tasks to which students are asked to respond on an assessment are not arbitrary. They must be carefully designed to provide evidence that is linked to the cognitive model of learning and to support the kinds of inferences and decisions that will be made on the basis of the assessment results. The *observation* vertex of the assessment triangle represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students. In assessment, one has the opportunity to structure some small corner of the world to make observations. The assessment designer can use this capability to maximize the value of the data collected, as seen through the lens of the underlying assumptions about how students learn in the domain.

Every assessment is also based on certain assumptions and models for interpreting the evidence collected from observations. The *interpretation* vertex of the triangle encompasses all the methods and tools used to reason from fallible observations. It expresses how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed. In the context of large-scale assessment, the interpretation method is usually a statistical model, which is a characterization or summarization of patterns one would expect to see in the data given varying levels of student competency. In the context of classroom assessment, the interpretation is often made less formally by the teacher, and is usually based on an intuitive or qualitative model rather than a formal statistical one.

A crucial point is that each of the three elements of the assessment triangle not only must make sense on its own, but also must connect to each of the other two elements in a meaningful way to lead to an effective assessment and sound inferences. Thus to have an effective assessment, all three vertices of the triangle must work together in synchrony. Central to this entire process, however, are theories and data on how students learn and what students know as they develop competence for important aspects of the curriculum.

Student Cognition and Domain-Specific Learning

As part of studying the nature of knowledge and learning, researchers have probed deeply the nature of competence and how people acquire large bodies of knowledge over long periods of time. Studies have revealed much about the kinds of mental structures that support problem-solving and learning in various domains; what it means to develop competence in a domain; and how the thinking of high achievers differs from that of novices and low achievers (e.g., Bransford et al., 2000; Chi, Feltovich, & Glaser, 1981). What distinguishes high from low performers is not simply general mental abilities or general problem-solving strategies. High performers have acquired extensive stores of knowledge and skill in a particular domain. But perhaps most significant, their minds have organized this knowledge in ways that a make it highly retrievable and useful. Because their knowledge has been encoded in a way that closely links it with the contexts and conditions for its use, high achievers do not have to search through the vast repertoire of everything they know when confronted with a task or problem. Instead, they can readily activate and retrieve the subset of their knowledge that is relevant to the task at hand (Simon, 1980; Glaser, 1992). Such findings suggest that teachers should place more emphasis on the conditions for applying the facts or procedures being taught, and that assessment should address whether students know when, where, and how to use their knowledge.

Considerable effort has also been expended on understanding the characteristics of persons and of the learning situations they encounter that foster the development of expertise. Much of what we know about the development of expertise has come from studies of children as they acquire competence in many areas of intellectual endeavor, including the learning of school subject matter. From a cognitive standpoint, *development* and *learning* are not the same thing. Some types of knowledge are universally acquired in the course of typical development, while other types are learned only with the intervention of deliberate teaching (which includes teaching by any means, such as apprenticeship, formal schooling, or self-study). Infants and young children appear to be predisposed to learn rapidly and readily in some domains, including language, number, and notions of physical and biological causality. Infants who are only 3 or 4 months old, for example, have been shown to understand certain concepts about the physical world, such as the idea that inanimate objects need to be propelled in order to move (Massey & Gelman, 1988). By the time children are 3 or 4 years old, they have an implicit understanding of certain rudimentary principles for counting, adding, and subtracting cardinal numbers (Gelman, 1990; Gelman & Gallistel, 1978).

In math, the fundamentals of ordinality and cardinality appear to develop in all nondisabled human infants without instruction. In contrast, however, such concepts as mathematical notation, algebra, and Cartesian graphing representations must be taught. Similarly, the basics of speech and language comprehension emerge naturally from millions of years of evolution, whereas mastery of the alphabetic code necessary for reading typically requires explicit instruction and long periods of practice (Geary, 1995). Much of what we want to assess in educational contexts is the product of such deliberate learning in specific curricular and instructional domains. Accordingly, every domain of knowledge and skill needs to be understood in terms of the body of concepts, factual content, procedures, and other components that together constitute the knowledge that we intend for students to learn across a span of time such as single year, a cluster of grades such as middle school, or the entire K–12 grade range. In virtually every curricular and content domain this knowledge is complex and multifaceted, requiring sustained effort and focused instruction to master. Developing deep knowledge of a domain such as that exhibited by high achievers, along with conditions for its use, takes time and focus and requires opportunities for distributed practice with feedback.

Whether considering the acquisition of some highly specific piece of knowledge or skill such as the process of adding two numbers, or some larger schema for representing and solving a mathematics or physics problem or understanding a genre of literature, certain laws of knowledge acquisition always apply. The first of these is the *power law of practice:* acquiring knowledge takes time, often requiring hundreds or thousands of instances of practice in retrieving a piece of information or executing a procedure. This law operates across a broad range of tasks, from typing on a keyboard to solving geometry problems (Anderson, 1981; Rosenbloom & Newell, 1987). According to the power law of practice, the speed and accuracy of performing a simple or complex cognitive operation increase in a systematic non-linear fashion over successive attempts (see Figure 3 for an example of typical data). This pattern is characterized by an initial rapid improvement in performance as shown in Figure 3, followed by subsequent and continuous improvements that accrue at a slower and slower rate. Such nonlinear, monotonic changes, can be readily modeled quantitatively by power or exponential functions. Thus, as shown in the lower panel of Figure 3, the change is linear as a function of the log of attempts. Even so, these simple quantitative patterns are typically best explained by substantial qualitative changes in the nature of what is represented in long-term memory and how that knowledge is accessed and deployed.

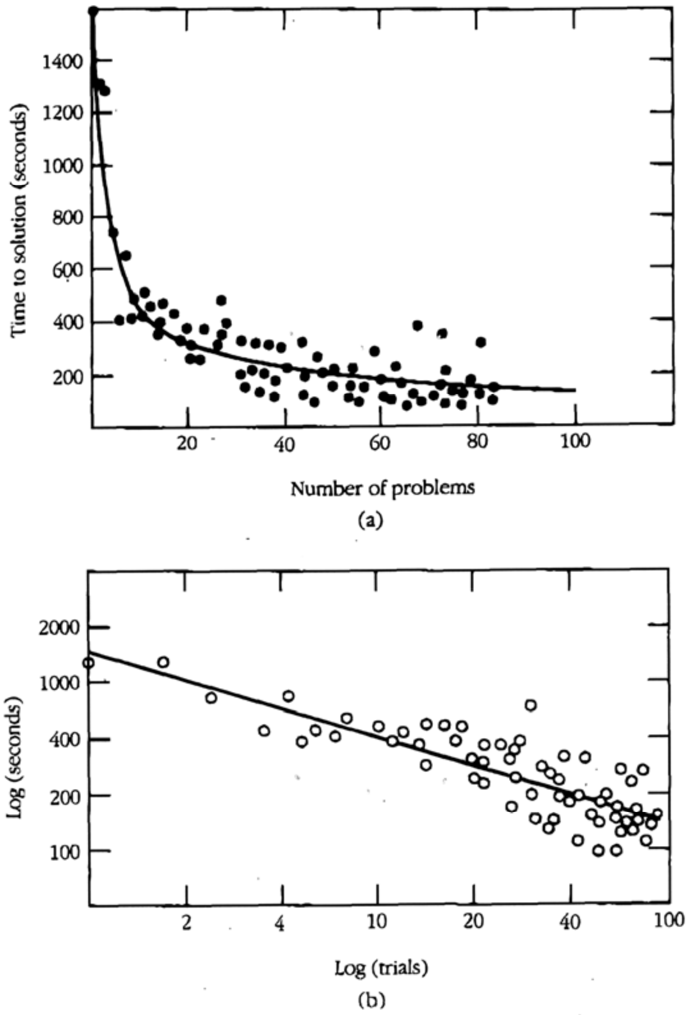

Figure 3. An illustration of the power law of knowledge acquisition.

Consistent with the above, one of the most important findings from detailed observations of children's learning and performance is that children do not move simply and directly from a state of *not knowing* to one of *knowing* (Kaiser, Proffitt, & McCloskey, 1985). Instead, their performance may exhibit several different but locally or partially correct understandings and strategies (Fay & Klahr, 1996). They also may use less advanced strategies even after demonstrating that they know more advanced ones, and the process of acquiring and consolidating robust and efficient knowledge and strategies may be quite protracted, extending across many weeks and months and hundreds of problems and examples (Siegler, 1998). These studies have also found, moreover, that short-term transition strategies often precede more lasting approaches and that generalization of new approaches often occurs very slowly. Thus, it is useful to remember that growth and change in knowledge is seldom a simple matter and it does not often lend itself to simple linear scales with equal interval or ratio measurement properties. Rather, growth and change may often resemble a nominal or ordinal scale, both of which have very different implications for measurement and quantification.

We also know that repeated exposure to content and practice in performing tasks are not enough to ensure that knowledge and skill will be acquired appropriately and/or efficiently. The conditions of learning and practice are also important. The second major law of knowledge acquisition involves *knowledge of results*. Individuals acquire knowledge much more rapidly and appropriately if they receive feedback about the correctness of what they have done. If incorrect, they need to know the nature of their mistake (Thorndike, 1931). One of the persistent dilemmas in education is that students often spend time practicing incorrect skills with little or no feedback. Furthermore, the feedback they ultimately receive is often neither timely nor informative. Unguided practice (e.g., homework in math) can be for the less able student, practice in doing tasks incorrectly. One of the most important roles for assessment is the provision of timely and informative feedback to students (and their teachers) during instruction and learning so that the practice of a skill and its subsequent acquisition will be effective and efficient (Black & Wiliam, 1998; Sadler, 1989; Wiliam, 2007).

Domain-Specific Learning: The Concept of Learning Progressions

A central thesis of this paper is that the targets of inference for any given assessment should be largely determined by models of cognition and learning that describe how people represent knowledge and develop competence in the domain of interest (the cognition element of the assessment triangle). Starting with a model of learning is one of the main features that distinguishes the proposed approach to assessment design from typical current approaches. The model suggests the most important aspects of student achievement about which one would want to draw inferences, and provides clues about the types of assessment tasks that will elicit evidence to support those inferences (see also Pellegrino, 1988; Pellegrino, Baxter, & Glaser, 1999; Pellegrino et al., 2001).

A model of learning that informs assessment design should have as many as possible of the following key features:

1. Be based on empirical studies of learners in the domain of interest.
2. Identify performances that differentiate beginning and expert performance in the domain.

3. Provide a developmental perspective, laying out typical progressions from novice levels toward competence and then expertise, and noting landmark performances along the way.

4. Allow for a variety of typical ways in which children come to understand the subject matter.

5. Capture some, but not all, aspects of what is known about how students think and learn in the domain. Starting with a theory of how people learn the subject matter, the designers of an assessment will need to select a slice or subset of the larger theory as the targets of inference.

6. Lend itself to being aggregated at different grain sizes so that it can be used for different assessment purposes (e.g., to provide fine-grained diagnostic information as well as coarser-grained summary information).

Consistent with these ideas, there has been a recent spurt of interest in the topic of *learning progressions* (see Duschl, Schweingruber, & Shouse, 2007; Wilson & Bertenthal, 2005). A variety of definitions of the learning progression (learning trajectory) construct now exist in the literature, with substantial differences in focus and intent (see e.g., Confrey, 2008; Confrey et al., 2009; Confrey, Maloney, Nguyen, Wilson, & Mojica, 2008; Corcoran, Mosher, & Rogat, 2009; Duncan & Hmelo-Silver, 2009). Perhaps the most extensive discussion of the learning progression construct can be found in the 2009 Consortium for Policy Research in Education (CPRE) *report Learning Progressions in Science* (Corcoran et al., 2009). As described therein, learning progressions (*trajectories*) are empirically grounded and testable hypotheses about how students' understanding of, and ability to use, core concepts and explanations and related disciplinary practices grow and become more sophisticated over time, with appropriate instruction (Duschl et al., 2007). These hypotheses describe the pathways students are likely to follow to the mastery of core concepts. They are based on research about how students' learning actually progresses. The hypothesized learning trajectories are tested empirically to ensure their construct validity (*does the hypothesized sequence describe a path most students actually experience given appropriate instruction?*) and ultimately to assess their consequential validity (*does instruction based on the learning progression produce better results for most students?*). The reliance on empirical evidence differentiates learning trajectories from traditional topical scope and sequence specification. Topical scope and sequence descriptions are typically based only on logical analysis of current disciplinary knowledge and on personal experiences in teaching.

The CPRE report (Corcoran et al., 2009) argued that learning progressions should contain at least the following elements:

1. *Target performances* or *learning goals* that are the end points of a learning progression and are defined by societal expectations, analysis of the discipline, and/or requirements for entry into the next level of education;

2. *Progress variables* that are the dimensions of understanding, application, and practice that are being developed and tracked over time. These may be core concepts in the discipline or practices central to literary, scientific or mathematical work;

3. *Levels of achievement* that are intermediate steps in the developmental pathway(s) traced by a learning progression. These levels may reflect levels of integration or common stages that characterize the development of student thinking. There may be intermediate steps that are non-canonical but are stepping stones to canonical ideas;

4. *Learning performances* that are the kinds of tasks students at a particular level of achievement would be capable of performing. They provide specifications for the development of assessments by which students would demonstrate their knowledge and understanding; and,

5. *Assessments* that are the specific measures used to track student development along the hypothesized progression. Learning progressions include an approach to assessment, as assessments are integral to their development, validation, and use.

In addition, the panelists contributing to the CPRE report (Corcoran et al., 2009) argued that learning progressions have some other common characteristics that seem especially relevant to the current discussion of developing more effective and useful measures of student achievement and growth. One of these characteristics is that they have internal conceptual coherence along several dimensions. For example, the progress variables capture important dimensions of understanding and practice and the achievement levels represent the successively more sophisticated levels of understanding and practice characterizing the development of student thinking over time. A progression may describe progress on a single progress variable or a cluster of related (and not just parallel) progress variables. It is also important that they can be empirically tested. The presumption is that they are not developmentally inevitable, but they may be developmentally constrained. Furthermore, they are crucially dependent on the instructional practices provided for the students whose development is studied in the processes of development and validation. Targeted instruction and curriculum will typically be required for students to progress along a trajectory; there may be multiple possible paths and progress is not necessarily linear. It may be more like ecological succession. A learning progression proposes and clarifies one or more possible paths and does not represent a complete list of all possible paths. At any given time, an individual may display thinking and/or practices characteristic of different points on the path, due to features of both the assessment context and the individual's cognition.

Figure 4 is an attempt to capture some of these ideas in a simple representation where it is assumed that there are core conceptual elements that are part of understanding a larger concept such as ratio and proportion, size and scale, or the atomic structure of matter. At each level in the system there are initial understandings of these core elements and they may be largely separate from one another. Some may be present in a person's knowledge structure and some may be absent. As one develops knowledge there is progress to the next level that implies a more sophisticated understanding of the core element. In addition, the elements within a level may become interconnected. Progress across levels involves assumptions in which there are increasing interconnections among the core elements and thus a much greater depth of understanding. We will try to illustrate this subsequently with some data from studies of the development of understanding for key concepts in science.

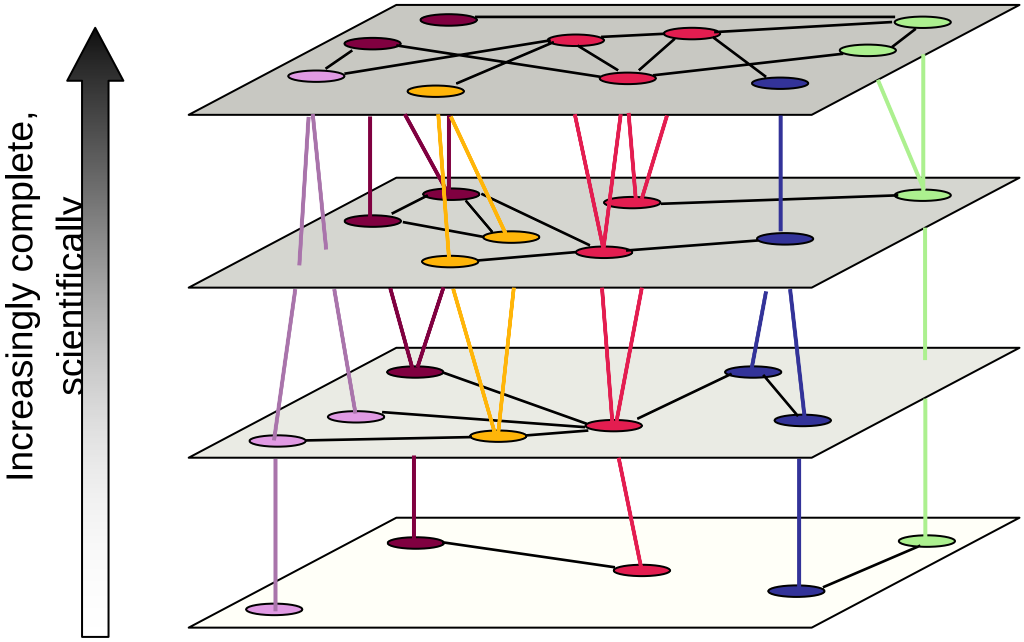

Figure 4. A representation of how conceptual elements may be linked within and across levels for a learning progression. From "Developing a Hypothetical Learning Progression for the Nature of Matter" by S. Y. Stevens, C. Delgado, and J. S. Krajcik, 2009, *Journal of Research in Science Teaching*, advance online publication. Copyright by S. Y. Stevens, C. Delgado, and J. S. Krajcik. Adapted with permission.

Below we illustrate some of what we currently know about the components of competence and the progression of learning in certain subdomains of mathematics, and science. We are not offering these descriptions as learning progressions that meet the criteria outlined above, but as illustrations of what is known about how knowledge and competence develops over time and with instruction for aspects of mathematics and science. It is far beyond the scope of this paper to try to present more than just a glimpse of how research and theory in the learning sciences is approaching the multiple conceptual and content elements within each academic domain as they play out across the K–12 grade span. Research on cognition and learning has produced a rich set of descriptions of domain-specific performance that can serve to guide assessment design, particularly for certain areas of reading, mathematics and science (e.g., American Association for the Advancement of Science [AAAS], 2001; Bransford et al., 2000; Donovan & Bransford, 2005; Duschl et al., 2007; Snow, Burns, & Griffin, 1998; Kilpatrick, Swafford, & Findell, 2001; Wilson & Berthenthal, 2005). It is also decidedly true that there is much left to do in mapping out learning progressions for multiple areas of the curriculum, most especially in ways that can effectively guide the design of instruction and assessment. Nevertheless, there is a good bit known about student cognition and learning that we can make use of right now to guide how we design systems of assessments, especially those that attempt to cover the progress of learning within and across grades.

Mathematics Learning

Investment in recent decades by federal agencies and private foundations has produced a wealth of knowledge about the development of mathematical understanding, and correspondingly has led to the development of curricula that incorporate such knowledge. Much of contemporary research and theory is synthesized in a report on elementary mathematics by Kilpatrick et al. (2001), and in the work of a RAND study group that produced a mathematics research agenda (RAND Mathematics Study Panel & Ball, 2002). Kilpatrick et al. drew on a solid research base in cognitive psychology and mathematics education to present a view of what elementary school children should know and be able to do in mathematics. This view includes mastery of procedures as a critical element of mathematics competence, but places far more emphasis on understanding when and how to apply those procedures than is common in many mathematics classrooms. The latter is rooted in a deeper understanding of mathematical concepts, and a facility with mathematical reasoning. Kilpatrick et al. summarized this view in five intertwining strands that constitute mathematical proficiency:

- *Conceptual understanding*–comprehension of mathematical concepts, operations, and relations;
- *Procedural fluency*–skill in carrying out procedures flexibly, accurately, efficiently, and appropriately;
- *Strategic competence*–ability to formulate, represent, and solve mathematical problems;
- *Adaptive reasoning*–capacity for logical thought, reflection, explanation, and justification;
- *Productive disposition*–habitual inclination to see mathematics as sensible, useful, and worthwhile, coupled with a belief in diligence and one’s own efficacy.

It is far beyond the scope of this paper to try to capture what is known empirically about the multiple aspects of mathematical proficiency, including their development as a consequence of instruction. The literature on mathematical cognition and its development covers a diversity of topics, ranging from geometry problem solving to fundamentals or arithmetic, to infant perception of numerosity (e.g., Greeno, 1978; Pellegrino, 2009; Starkey & Cooper, 1980). For our present purposes it is perhaps most useful to consider some of the ways in which the concept of learning progressions is being used to map out key aspects of K-8 mathematics.

As part of work that Jere Confrey and her colleagues have pursued on mapping out possible learning progressions for elementary and middle school mathematics, they have offered the following definition of a learning progression (trajectory): *A researcher-conjectured, empirically-supported description of the ordered network of constructs a student encounters through instruction (i.e. activities, tasks, tools, forms of interaction, and methods of evaluation), in order to move from informal ideas, through successive refinements of representation, articulation, and reflection, towards increasingly complex concepts over time* (Confrey et al., 2008). Figure 5 is an attempt to capture some of the major elements and interconnections for the domain of mathematics across the K–5 grade span (Confrey et al., 2009). This map can in turn be used to make sense of the literature on mathematics instruction and learning, and to develop sets of learning progressions for key aspects of this conceptual space.

One the major concepts that Confrey and her colleagues have been exploring is that of *equipartitioning* as a core concept that underlies the development of an increasingly sophisticated understanding of topics that comprise a major part of the typical K–5 curriculum, including multiplication and division, ratio and proportion, and fractions and decimals. As part of this work, Confrey et al. (2009) have

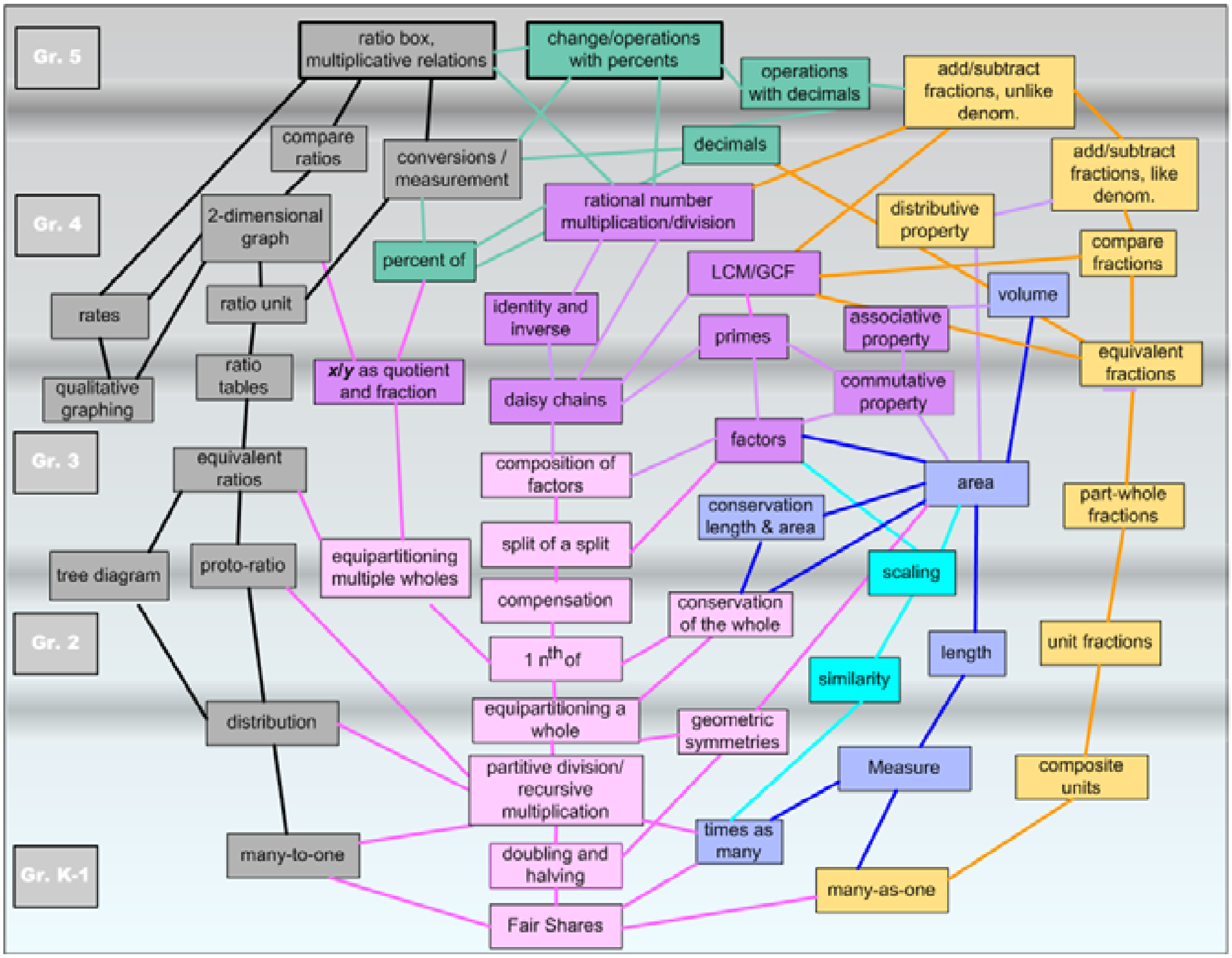

Figure 5. Partial map of the construct space underlying learning progressions for K–5 mathematics concepts. From "Equipartitioning/Splitting as a Foundation of Rational Number Reasoning Using Learning Trajectories" by J. Confrey, A. P. Maloney, K. H. Nguyen, G. Mojica, and M. Myers, 2009, paper presented at the 33rd Conference of the International Group for the Psychology of Mathematics Education, Thessaloniki, Greece. Copyright 2009 by J. Confrey. Adapted with permission.

proposed a learning trajectory for equipartioning that encompasses the following performance levels such that students can successively:

1. *Equipartition collections*
   by dealing single units or composite units

2. *Equipartition a single whole* (rectangles, circles)
   Criteria: correct number of parts, equal-sized parts, exhaust the whole

3. Justify equivalence of shares--by counting, stacking, arrays, patterns, etc.

4. *Name* the shares, in relation to referent unit:
   a. Of collections: sharing 12 among 2: *half* or *six*
   b. Of single wholes: sharing a whole among n: *1/n* or *1/n* of

5. *Reassemble* equal groups (of collections) or parts (of a whole)
   a. *n times as many*, or *n times as much*

6. Predict effect of changes in number of people sharing on size of shares *(qualitative compensation)*

7. Predict outcome of a *composition of splits* (splits of a split of a whole)
   a. Two or more splits, and identification of factor-based pairs

8. Demonstrate and justify the effect of factor-based changes in number of persons sharing on the *size of shares,* and vice versa*, for collections or single whole (quantitative compensation 1)*

9. Demonstrate and justify how extra shares can be redistributed for fewer people *(additive changes) sharing collections (quantitative compensation 2).*

10. Demonstrate *equivalence of noncongruent parts* across or within methods of non-prime equipartitioning *(transitivity)*
    a. *Decomposition/composition*:
    b. *Transitivity*: if $X = Y$, $x = 1/2X$, and $y = 1/2Y$, then $x = y$

11. Assert that a whole can be *equipartitioned* for all natural numbers greater than 1 (*continuity principle*)

12. *Equipartition multiple wholes* among multiple persons and name the resulting shares in relation to referent units.

13. Predict the outcome of a *composition of splits* on *multiple wholes.*

14. Make *factor or split-based* changes in number of objects, number of people sharing, and/or the size of fair shares, and predict the effects on the other variables *(direct, inverse, and covariation to quantify compensation).*

15. Apply *distributive property* to *multiple wholes,* demonstrating *equipartitioning over breaking or fracturing.*

16. Generalize that *a objects shared among b persons results in a/b objects per person,* based on both distributive property and ratio reasoning. (Confrey et al., 2009)

The above is a hypothesized sequence for a major conceptual strand of elementary grades mathematics and the details need to be validated empirically. The progression is stated in terms of a set of increasingly complex performances that students would be expected to achieve, which makes it especially useful for thinking about the nature of the instruction that might support development of these competencies as well as ways in which they might be assessed. Not stated, however, are details of the knowledge representations and conceptual understanding that students would develop along the way as their understanding grows, although aspects of that understanding are implicit in the statements of performance at each of the levels.

This is but one illustration of what a learning progression might look like for critical aspects of the knowledge and skill we want students to develop over the course of multiple grades of instruction in mathematics. The value of such descriptions to the design of instruction and assessment, and the implications for measurement of growth, are topics we will return to after illustrating similar conceptual work in science education. It is also worth noting that in addition to the work underway by Confrey and her colleagues, there is similar work being pursued by others in mapping out possible progressions for various aspects of the K–12 mathematics curriculum. Such work includes a CPRE-led effort similar to the one they led for science learning progressions (e.g., Corcoran et al., 2009).

Science Learning

As is true for the area of K–12 mathematics, investment in recent decades by federal agencies and private foundations has produced a wealth of knowledge about the development of students' science knowledge and understanding that, correspondingly, has led to the development of curricula that incorporate such knowledge. Much of contemporary research and theory is synthesized in a report on elementary science education by Duschl et al. (2007), *Taking Science to School*, which presented a view of what elementary school children should know and be able to do in science that drew on a solid research base in the learning sciences and science education. This view included mastery of facts as a critical element of science competence, but placed far more emphasis on understanding the process of science in explaining the natural world and the nature of scientific evidence and argument. It emphasized a deeper understanding of core science concepts and so-called big ideas and facility with the scientific reasoning process and the practices that scientists engage in when doing their work as part of a larger community that builds knowledge. Duschl et al. summarized this view in terms of four intertwining strands in which student proficiency in science is defined as being able to

- know, use, and interpret scientific explanations of the natural world;
- generate and evaluate scientific evidence and explanations;
- understand the nature and development of scientific knowledge; and
- participate productively in scientific practices and discourse.

These proficiencies share a great deal in common with those identified for mathematics in the sense that they emphasize not just factual and procedural competence but the development of conceptual understanding that is connected to the doing of science. As was true for the preceding discussion of mathematics, it is far beyond the scope of this paper to try to capture what is known empirically about the multiple aspects of science proficiency, including their development as a consequence of instruction. The literature on science cognition and its development covers a diversity of topics that span areas of science from the life sciences to physical sciences to the earth, space and environmental sciences (see Duschl et al., 2007). For our present purposes it is perhaps most useful to consider some of the ways in which the concept of learning progressions is being used to map out key aspects of science knowledge and understanding across the K–12+ spectrum.

The science education community has moved away from using a multitude of national, state and district standards to guide thinking about the science curriculum to a view that curriculum and instruction should be focused on helping students develop an understanding of core or *big ideas* ideas (Duschl et al., 2007). Big ideas of science involve concepts, principles, and models that help explain a broad range of phenomena and may encompass knowledge within a single or across multiple disciplines (Smith et al., 2006). Thus, big ideas are important for science literacy and should be considered the foundation for building a coherent science curriculum.

As part of the effort to identify big ideas that have implications for science and engineering education, four big ideas have been given special attention: structure of matter, forces and interactions, size and scale, and quantum effects. These four constitute foundational science content for nanoscale science and each of the big ideas informs and is informed by the others (see Figure 6; Stevens, Sutherland, & Krajcik, 2009). For example, the structure of matter and the way matter interacts are inextricably linked. The forces that generally dominate those interactions change with scale. Different physical models (i.e., classical mechanics, quantum mechanics, general relativity) are more appropriate to explain the behavior of matter at different scales. Therefore, building integrated understandings of these ideas requires connecting concepts across them.

Unfortunately, evidence shows that students have weak and underdeveloped understandings of these fundamental ideas and that includes the foundational concepts of size and scale (AAAS, 1993; Tretter, Jones, Andre, Negishi, & Minogue, 2006). Accordingly, efforts have been underway to develop an empirical learning progression for size and scale to better understand what students do know and understand and how we might design better instruction and assessment (Delgado, Stevens, Shin, & Krajcik, 2008). For present purposes we can represent possible understandings of size and scale in terms of four components that vary in sophistication. These four components have implications for the kinds of judgments and performances that individuals who posses each component understanding can demonstrate:

- *Qualitative relative*–A person can order objects by size:
  A > B > C > D > E > F > G > H > I > J.

- *Categorical*–A person can group objects of “similar” size and order groups by size:
  {A, B, C} > {D, E} > {F, G} > {H, I, J}.

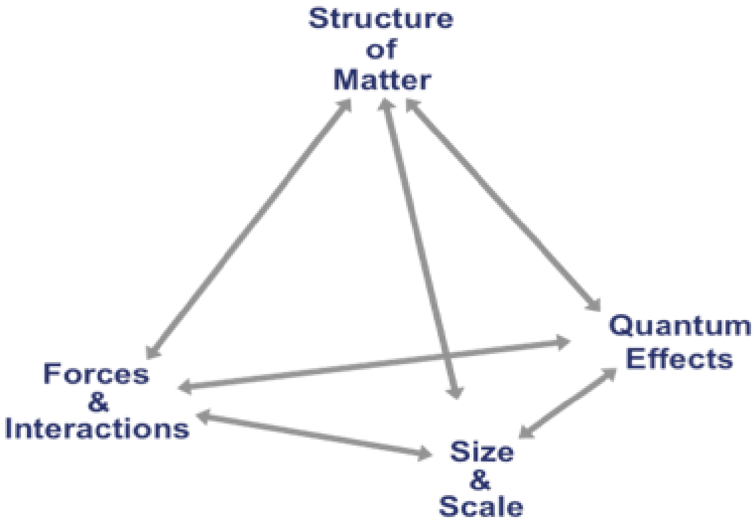

Figure 6. Four of the big ideas critical for understanding nanoscale science. From *The Big Ideas of Nanoscale Science and Engineering: A Guidebook for Secondary Teachers* by S. Y. Stevens, L. M. Sutherland, and J. S. Krajcik, 2009, Arlington, VA: NSTA Press. Copyright 2009 by S. Y. Stevens, L. M. Sutherland, and J. S. Krajcik. Adapted with permission.

- *Quantitative relative*–A person can designate that Object C is 1,000 times bigger than object E.
- *Absolute*–A person can specify absolute size such as Object E is 1 nm in length.

These components were used to generate a set of tasks suitable for probing the knowledge of students who spanned a range of grades from Grade 6 through college.

Figures 7 and 8 illustrate some of the results expressed in terms of the levels of sophistication of reasoning and understanding that students demonstrated across the full set of tasks. As shown in both figures, it was possible to order students along a continuum of levels of performance that reflect the components of size and scale that they understood as defined by the types of judgments they could make. As shown in Figure 7 (Delgado, Stevens, & Shin, 2008), some students were at the lowest level (Level 0) while others were at the highest level (Level 5) and there were students who fell into all of the levels in-between. In fact, it was possible to classify virtually all students into one of the six levels of knowledge and understanding. Figure 8 (Delgado, 2008) is an attempt to illustrate what this might mean using the visual representation of ordered levels shown earlier in Figure 4.

In general, while students in more advanced science courses have more sophisticated conceptions, there is wide variation between students in a given class. Most students cannot use the number of times bigger/smaller an unfamiliar object is compared to an object of known size to find the size of the unfamiliar object, and two thirds do not believe that the actual sizes and relative sizes are necessarily

Number of students at the level	Order–group	Order–relative	Order–absolute	Absolute–relative (conceptual)	Absolute–relative (procedural)	Level
6	✓	✓	✓	✓	✓	5
8	✓	✓	✓	✓		4
13	✓	✓	✓			3
3	✓	✓				2
5	✓					1
4						0
# students	35	30	27	14	6	

Figure 7. A graphical representation of results showing levels of student understanding for critical components of size and scale and their hypothetical interconnections within and across levels. From *Development of a Learning Progression for Students' Conceptions of Size and Scale*, by C. Delgado, S. Stevens, and N. Shin, 2008, paper presented at the National Association for Research in Science Teaching Annual Conference, Baltimore, MD.

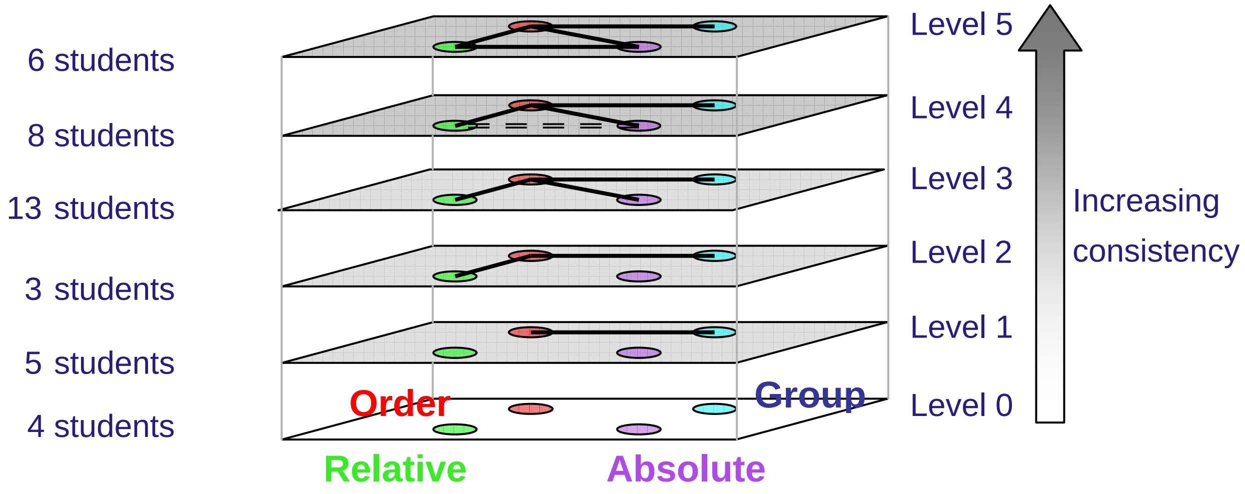

Figure 8. A graphical representation of results showing levels of student understanding for critical components of size and scale and their hypothetical interconnections within and across levels. From Development of a Research-Based Learning Progression for Middle School Through Undergraduate Students' Conceptual Understanding of Size and Scale, by C. Delgado, 2008, paper presented at the International Conference of the Learning Sciences. Copyright 2008 by C. Delgado. Reprinted with permission.

and logically related. Other connections, such as those between ordering and grouping, are much more widespread. As shown in Figures 7 and 8, it is possible to determine the order in which students tend toestablish connections across facets of size and scale, and studies are underway to understand how this progression is related to the accumulation of accurate content knowledge about the size of specific objects.

Learning progressions of the type just illustrated have been delineated for other physical science concepts such as the structure of matter, as well as biological concepts like genetics, and earth and environmental science concepts (see, e.g., Corcoran et al., 2009). Such progressions are based on various sources of evidence and much needs to be done to validate them. A part of the validation process includes using them productively for purposes of designing curriculum, instruction and assessment.

Some Concluding Comments About Learning Progressions

There is considerable current interest in learning progressions but the field of practice and policy must be cautious in assuming that everything being espoused has a sound base and is ready for prime time. There is a danger in leaping too readily to embrace the construct without questioning the evidentiary base behind any given progression that is proposed. That said, there is much to potentially recommend learning progressions as ways to think about the assessment of student learning. One benefit of carefully described learning progressions is that they can be used to guide the specification of learning performances—statements of what students would be expected to know and be able to do. This was illustrated in the examples provided above for math and science. The learning performances can in turn guide the development of tasks that allow one to observe and infer students' levels of competence for major constructs that are the target of instruction and assessment within and across grade levels.

The potential relevance of any learning progression may vary with the purpose of the assessment and intended use of the information. This will be a function of the scope and specificity of the learning progression. The more detailed it is and the finer the grain size, the more useful it may be at levels close to classroom instruction. They have potential roles to play in supporting and monitoring development and growth and they may be especially relevant for aspects of diagnosis and instructional support. Finally, learning progressions can help us understand why working from a micro- to the macro-level understanding of student cognition and learning to generate assessments is more likely to lead to valid inferences about student achievement than the reverse. When we have detailed maps of the progress of student learning, at grain sizes that support instructional design and diagnostic assessment, we have a conceptual base that can be collapsed to make coarser judgments about aspects of growth and change appropriate to a broader timescale of learning. In doing so we preserve the validity of the assessment because we have a clear sense of the construct being measured and the level at which we can describe and understand student performance. Some of these issues are considered further in the next section.

Implications for Assessment Design and Measurement

Assessment Purposes, Levels, and Timescales

Although assessments are currently used for many purposes in the educational system, a premise of the *Knowing What Students Know* report (Pellegrino et al., 2001) is that their effectiveness and utility must ultimately be judged by the extent to which they promote student learning. The aim of assessment should be "*to educate and improve* student performance, not merely to *audit* it" (Wiggins, 1998, p. 7). Because assessments are developed for specific purposes, the nature of their design is very much constrained by their intended use. The reciprocal relationship between function and design leads to concerns about the inappropriate and ineffective use of assessments for purposes beyond their original intent. To clarify some of these issues of assessment purpose, design, and use, it is worth considering two pervasive dichotomies in the literature that are often misunderstood and conflated.

The first dichotomy is between *internal* classroom assessments administered by teachers, and *external* tests administered by districts, states, or nations. Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) showed that these two very different types of assessments are better understood as two points on a continuum that is defined by their distance from the enactment of specific instructional activities. They defined five discrete points on the continuum of assessment distance: *immediate* (e.g., observations or artifacts from the enactment of a specific activity), *close* (e.g., embedded assessments and semiformal quizzes of learning from one or more activities)*, proximal* (e.g., formal classroom exams of learning from a specific curriculum), *distal* (e.g., criterion-referenced achievement tests such as required by the U.S. No Child Left Behind legislation), and *remote* (broader outcomes measured over time, including norm-referenced achievement tests and some national and international achievement measures. Different assessments should be understood as different points on this continuum if they are to be effectively aligned with each other and with curriculum and instruction.

A second pervasive dichotomy is the one between formative assessments used to advance learning and summative assessments used to provide evidence of prior learning. Often it is assumed that classroom assessment is synonymous with formative assessment, and that large-scale assessment is synonymous with summative assessment. What are now widely understood as different types of assessment practices are more productively understood as different functions of assessment practice, and summative *and* formative functions can be identified for most assessment activities, regardless of the level on which they function.

Drawing from the work of Lemke (2001), it is apparent that different assessment practices can be understood as operating at different *timescales*. The timescales for the five levels defined above can be characterized as *minutes, days, weeks, months,* and *years.* Timescale is important because the different competencies that various assessments aim to measure (and, therefore, the appropriate timing for being impacted by feedback) are *timescale-specific*. The cycles, or periodicity, of educational processes build from individual utterances into an individual's lifespan of educational development. What teachers and students say in class constitute verbal exchanges; these exchanges make up the lesson; a sequence of lessons make up the unit; units form a curriculum, and the curricula form an education. Each of these elements operates on different cycles or timescales: second to second, day to day, month to month, and year to year.

The level at which an assessment is intended to function, which involves varying distance in space and time" from the enactment of instruction and learning, has implications for how and how well it can fulfill various functions of assessment, be they formative, summative, or program evaluation (see National Research Council [NRC], 2003). As argued elsewhere (Hickey & Pellegrino, 2005; Pellegrino & Hickey, 2006), it is also the case that the different levels and functions of assessment can have varying degrees of match with theoretical stances about the nature of knowing and learning. With this in mind we now turn to the implications of cognitive theory and research for both classroom assessment practices and for large-scale assessment. These two contexts reflect some of the rich variation in assessment captured by the foregoing discussion of levels, functions, and timescales.

Implications for Design and Use of Classroom Assessment

Shepard (2000) discussed ways in which classroom assessment practices need to change to better support learning: the content and character of assessments need to be significantly improved to reflect contemporary understanding of learning; the gathering and use of assessment information and insights must become a part of the ongoing learning process; and assessment must become a central concern in methods courses in teacher preparation programs. Her messages are reflective of a growing belief among many educational assessment experts that if assessment, curriculum, and instruction were more integrally connected, as argued in Section 1, student learning would improve (e.g., Pellegrino, Baxter, et al., 1999; Stiggins, 1997).

Sadler (1989) provided a conceptual framework that places classroom assessment in the context of curriculum and instruction. According to this framework, three elements are required for assessment to promote learning:

1. A clear view of the learning goals (derived from the curriculum).
2. Information about the present state of the learner (derived from assessment).
3. Action to close the gap (taken through instruction).

Furthermore, there are ongoing, dynamic relationships among formative assessment, curriculum, and instruction. That is, there are important bidirectional interactions among the three elements, such that each informs the other. For instance, formulating assessment procedures for classroom use can spur a teacher to think more specifically about learning goals, thus leading to modification of curriculum and instruction. These modifications can, in turn, lead to refined assessment procedures, and so on. The mere existence of classroom assessment along the lines discussed here will not ensure effective learning. The clarity and appropriateness of the curriculum goals, the validity of the assessments in relationship to these goals, the interpretation of the assessment evidence, and the relevance and quality of the instruction that ensues are all critical determinants of the outcome. Starting with a model of cognition and learning in the domain can enhance each of these determinants.

For most teachers, the ultimate goals for learning are established by the curriculum, which is usually mandated externally (e.g., by state curriculum standards). However, teachers and others responsible for designing curriculum, instruction, and assessment must fashion intermediate goals that can serve as an effective route to achieving the ultimate goals, and to do so effectively, they must have an understanding of how students represent knowledge and develop competence in the domain. National and state curriculum standards set forth learning goals, but often not at a level of detail that is useful for operationalizing those goals in instruction and assessment. By dividing goal descriptions into sets appropriate for different age and grade ranges, current curriculum standards provide broad guidance about the nature of the progression to be expected in various subject domains. Whereas this kind of epistemological and conceptual analysis of the subject domain is an essential basis for guiding assessment, deeper cognitive analysis of how people learn the subject matter is also needed. Formative assessment should be based in cognitive theories about how people learn particular subject matter to ensure that instruction centers on what is most important for the next stage of learning, given a learner's current state of understanding.

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

Implications for Design and Use of Large-Scale Assessment

Large-scale assessments are further removed from instruction but can still benefit learning if well designed and properly used. Substantially more valid, useful, and fair information could be gained from large-scale assessments if the principles of design set forth above earlier in Section 2 were applied. However, fully capitalizing on contemporary theory and research will require more substantial changes in the way large-scale assessment is approached, and relaxation of some of the constraints that currently drive large-scale assessment practices. Below we discuss some of the needed changes.

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

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

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

A major problem is that only limited improvements in large-scale assessments are possible under current constraints and typical standardized testing scenarios. Large-scale assessments are designed to meet certain purposes under constraints that often include providing reliable and comparable scores for individuals as well as groups; sampling a broad set of curriculum standards within a limited testing time per student; and offering cost-efficiency in terms of development, scoring, and administration. To meet these kinds of demands, designers typically create assessments that are given at a specified time, with all students being given the same (or parallel) tests under strictly standardized conditions (often referred to as *on-demand* assessment). Tasks are generally of the kind that can be presented in paper-and-pencil format, that students can respond to quickly, and that can be scored reliably and efficiently. In general, competencies that lend themselves to being assessed in these ways are tapped, while aspects of learning that cannot be observed under such constrained conditions are not addressed. To design new kinds of situations for capturing the complexity of cognition and learning will require examining the assumptions and values that currently drive assessment design choices and breaking out of the current paradigm to explore alternative approaches to large-scale assessment, including innovative uses of technology (see e.g., Quellmalz & Pellegrino, 2009).

Implications for Models of Measurement

As argued earlier, assessment is a process of drawing reasonable inferences about what students know on the basis of evidence derived from observations of what they say, do, or make in selected situations. The field of psychometrics has focused on how best to gather, synthesize, and communicate evidence of student understanding in an explicit and formal way (the interpretation vertex of the assessment triangle). Psychometric models are based on a probabilistic approach to reasoning. From this perspective, a statistical model is developed to characterize the patterns believed most likely to emerge in the data for students at varying levels of cognitive competence. When there are large masses of evidence to be interpreted and/or when the interpretations are complex, the complexity of these models can increase accordingly.

Humans have remarkable abilities to evaluate and summarize information—but remarkable limitations as well. Formal probability-based models for assessment were developed to overcome some of these limitations. Such models allow one to draw meaning from quantities of data far more vast than a person can grasp at once and to express the degree of uncertainty associated with one's conclusions. In other words, a measurement model is a framework for communicating with others how the evidence in observations can be used to inform the inferences one wants to draw about learners.

Findings from cognitive research suggest that new kinds of inferences are needed about students and how they acquire knowledge and skill if assessments are to be used to track and guide student learning. For example, Wolf, Bixby, Glenn, and Gardner (1991) presented a strong argument for the following needs:

> If we are able to design tasks and modes of data collection that permit us to change the data we collect about student performance, we will have still another task in front of us. This is the redesign or invention of educational psychometrics capable of answering the much-changed questions of educational achievement. In place of ranks, we will want to establish a developmentally ordered series of accomplishments. First . . . we are opening up the possibility of multiple paths to excellence. . . . Second, if we indeed value clinical judgment and a diversity of opinions among appraisers (such as certainly occurs in professional settings and post secondary education), we will have to revise our notions of high-agreement reliability as a cardinal symptom of a useful and viable approach to scoring student performance. . . . Third, we will have to break step with the drive to arrive at a single, summary statistic for student performance. . . .After all, it is critical to know that a student can arrive at an idea but cannot organize her or his writing or cannot use the resources of language in any but the most conventional and boring ways . . . . Finally, we have to consider different units of analysis. . . [b]ecause so much of learning occurs either in social situations or in conjunction with tools or resources, we need to consider what performance looks like in those more complex units. (pp. 63-64)

Although a single statement could not be expected to outline all possible needs, the list provided by Wolf et al. (1991) is challenging and instructive. A major message of the *Knowing What Students Know* (Pellegrino et al., 2001) report is that the measurement models currently available can support the kinds of inferences that learning sciences suggest are important to pursue. Much of Wolf et al.'s agenda could be accomplished now, with the measurement tools already available. Other parts of this agenda are less easily satisfied.

In particular, it is now possible to characterize student achievement in terms of multiple aspects of proficiency, rather than a single score; chart students' progress over time, instead of simply measuring performance at a particular point in time; deal with multiple paths or alternative methods of valued performance; model, monitor, and improve judgments based on informed evaluations; and model performance not only at the level of students, but also at the levels of groups, classes, schools, and states.

Table 1 summarizes and gives examples of some of the major developments in methods of measurement and the assessment challenges they help address. (The measurement term *construct* is used in the table to refer to the aspects of cognition and learning that are the targets for assessment.) The table shows that work on measurement models has progressed from (a) developing models that are intended to measure general proficiency and/or to rank examinees (referred to here as *standard* models); to (b) adding enhancements to a standard psychometric model to make it more consistent with changing conceptions of learning, cognition, and curricular emphasis; to (c) incorporating cognitive elements, including a model of learning and curriculum, directly into psychometric models as parameters; to (d) creating a family of models that are adaptable to a broad range of contexts. Each model and adaptation has its particular uses, strengths, and limitations, which are discussed in greater detail in Pellegrino et al. (2001).

However, many of the newer models and methods are not widely used because they are not easily understood or packaged in accessible ways for those without a strong technical background. Technology offers the possibility of addressing this shortcoming. For instance, by building statistical models into technology-based learning environments for use in classrooms, teachers can employ more complex tasks, capture and replay students' performances, share exemplars of competent performance, and in the process gain critical information about student competence.

Much hard work remains to focus psychometric model building on the critical features of models of cognition and learning and on observations that reveal meaningful cognitive variation in a particular domain. If anything, the task has become more difficult because an additional step is now required--determining the inferences that must be drawn, the observations needed, the tasks that will provide them, and the statistical models that will express the necessary patterns most efficiently. Therefore, having a broad array of models available does not mean that the measurement model problem is solved. The longstanding tradition of leaving scientists, educators, task designers, and psychometricians each to their own realms represents perhaps the most serious barrier to progress.

The upshot of the preceding discussion is that measuring student performance and growth is not a simple matter. A variety of sophisticated statistical and analytic models are available but their choice and appropriateness will vary with three factors: (a) the time scale over which learning has occurred, (b) the grain size and details of what was to be learned, and (c) the intended purpose and use of the inferences we wish to make. A single metric of growth may not be feasible, sensible nor attainable for many cases of educational practice, and a multivariate measurement approach may be what we need.

Table 1. **An Array of Available Measurement Models**

CLASS OF MEASUREMENT MODELS	*USEFUL IF ONE WANTS TO...*	*SOME AVAILABLE METHODS*
Standard psychometric models	Represent achievement as a continuous variable—a single progression from less to more knowledge, skills, etc. (e.g., making inferences about students' general mathematics proficiency). These models are the basis for the dominant current approaches to educational assessment.	Classical test theory (Lord & Novick, 1969) Generalizability theory (Brennan, 1983) Item response modeling (IRM) (van der Linden & Hambleton, 1996)
	Represent the cognitive construct as a set of discrete classes, ordered or unordered. Student cognition is characterized as falling into one of a finite set of classes (e.g., classifying students according to which of several possible strategies they are using).	Latent class models (Haertel, 1990)
	Represent the cognitive construct as composed of multiple attributes (e.g., the ability to solve mathematics word problems might be conceived of as requiring both computation and verbal competencies).	Factor analysis Multidimensional item response model (Adams, Wilson, & Wang, 1997)
	Capture student change over time (e.g., growth in student understanding of inquiry over the course of a science curriculum). The modeling of change adds a hierarchical level to the construct—one level is a theory of measuring student cognition at a particular time, and a second level is the theory of how cognition tends to change over time	Multilevel models (Bryk & Raudenbush, 1992) Structural equation modeling (Willet & Sayer, 1994) Special item response models (Embretson, 1996) Combination of IRM and multilevel modeling (Adams, Wilson, & Wu, 1997)
Enhanced standard psychometric models that are more consistent with contemporary views of learning	Enhance the interpretability of measures by displaying student achievement graphically in the form of a "progress map" or "construct map" (e.g., report students' spelling achievement in relation to a typical developmental progression in the domain).	Developmental assessment (Masters, Adams & Wilson, 1990)
	Add diagnostics by examining the patterns of student responses. These approaches can be used for identifying individual differences (e.g., for identifying when an individual student responds to a set of problems in surprising ways that diverge from the predicted responses).	Kidmap (Mead, 1976) GradeMap (Wilson, Draney & Kennedy, 2001) Rule-space representations (Tatsuoka, 1995)
Models that incorporate cognitive elements as parameters	Investigate whether different groups of examinees of approximately the same ability appear to be using different cognitive processes to respond to the same set of items (e.g., differential response patterns among language groups can be detected statistically and then followed up with linguistic comparison of the test versions to uncover reasons for the differential responses.)	Differential item functioning (DIF) (Ercikan, 1998; Lane, Wang, & Magone, 1996)
	Model different classes of items rather than items themselves—item sets are considered to be a sample from their respective item latent class. This approach allows one to make diagnostic statements about students (e.g., whether the student possesses the "compare fractions skill").	Hierarchical IRT model (Janssen, Tuerlinckx, Meulders, & De Boeck, 2000; Junker, 1999)
	Model possibility of people using different strategies to solve problems—for example, developmental stages.	Mixture models (e.g., Mislevy & Wilson, 1996)
Generalized approaches to modeling of cognitive structures	Great effort may be required to develop and apply measurement models with features specific to particular to learning theories and assessment purposes. Hence one is drawn to the possibility of establishing families of models that can be applied more broadly. Several such models have been developed and, in varying degrees, implemented.	Unified model (DiBello, Jiang, & Stout, 1999) M^{2}RCML (Pirolli & Wilson, 1998) Bayes nets (Mislevy, 1996)

Balanced Assessment Systems

Given that one form of assessment does not serve all purposes, it is inevitable that multiple assessments (or assessments consisting of multiple components) will be required to serve the varying educational assessment needs of different audiences. A multitude of different assessments are already being used in schools. It is not surprising that users are often frustrated when such assessments have conflicting achievement goals and results. Sometimes such discrepancies can be meaningful and useful, such as when assessments are explicitly aimed at measuring different school outcomes. More often, however, conflicting assessment goals and feedback cause much confusion for educators, students, and parents. In this section we describe a vision for coordinated systems of multiple assessments that work together, along with curriculum and instruction, to promote learning, but first we consider issues of balance and allocation of resources across classroom and large-scale assessment (see also NRC, 2003).

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

Not only does large-scale assessment dominate over classroom assessment, but there is also ample evidence of accountability measures negatively impacting classroom instruction and assessment. For instance, as discussed earlier, teachers feel pressure to teach to the test, which results in a narrowing of instruction. They also model their own classroom tests after less-than-ideal standardized tests (Linn, 2000; Shepard, 2000). These kinds of problems suggest that beyond striking a better balance between classroom and large-scale assessment, what is needed are coordinated systems of assessments that collectively support a common set of learning goals, rather than work at cross-purposes.

To this end, an assessment system should exhibit three properties: comprehensiveness, coherence, and continuity. These notions of alignment are consistent with those set forth by various groups including the National Council of Teachers of Mathematics and the National Academy of Education.

By *comprehensiveness*, we mean that a range of measurement approaches should be used to provide a variety of evidence to support educational decision-making. Multiple measures take on particular importance when important, life-altering decisions (such as high school graduation) are being made about individuals (Bransford et al., 2000). No single test score can be considered a definitive measure of a student's competence. Multiple measures enhance the validity and fairness of the inferences drawn by giving students various ways and opportunities to demonstrate their competence. Multiple measures can also be used to provide evidence that improvements in test scores represent real gains in learning, as opposed to score inflation due to teaching narrowly to one particular test (Heubert & Hauser, 1999).

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

Finally, an ideal assessment system would be designed to be *continuous*. That is, assessments should measure student progress over time, akin more to a videotape record rather than to the snapshots provided by most current tests. To provide such pictures of progress, multiple sets of observations over time must be linked conceptually so that change can be observed and interpreted. Models of student progress in learning should underlie the assessment system, and tests should be designed to provide information that maps back to the progression. Thus, continuity calls for alignment along the third dimension of time.

Figure 9, developed by the Center for Assessment and Evaluation of Student Learning, provides a graphical illustration of what an assessment system might look and some of the factors that would serve to achieve balance and support the principles described above (Herman, Wilson, Shavelson, Timms, & Schneider, 2005). The system illustrated in Figure 9 would be; (a) coordinated across levels, (b) unified

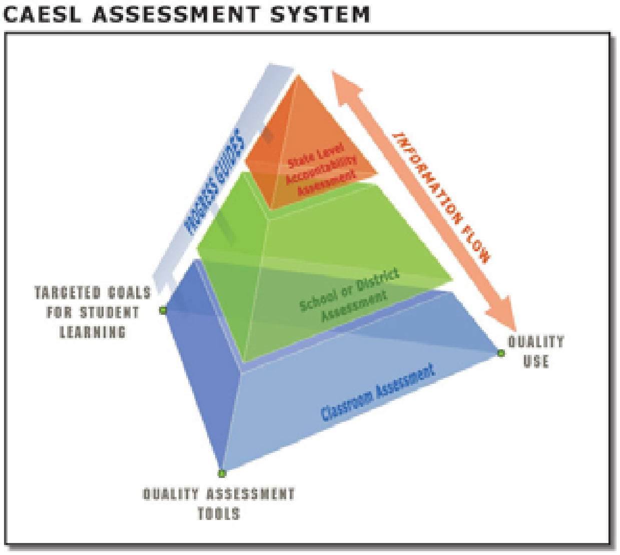

Figure 9. Center for Assessment and Evaluation of Student Learning (CAESL) representation of a coordinated, multilevel assessment system. From *The CAESL Assessment Model*, by J. L. Herman, M. R., Wilson, R. Shavelson, M. Timms, and S. Schneider, 2005, paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. Copyright 2005 by J. L. Herman, M. R. Wilson, R. Shavelson, M. Timms, and S. Schneider. Reprinted with permission.

by common learning goals, and (c) synchronized by unifying progress variables. No existing system of assessments that we know of has these design features and meets all three criteria of comprehensiveness, coherence, and continuity, but there are examples of assessments that represent steps toward these goals. For instance, Australia's Developmental Assessment program (Masters & Forster, 1996) and the BEAR assessment system (Wilson & Sloane, 2000; Wilson, Draney & Kennedy, 2001) show how progress maps can be used to achieve coherence between formative and summative assessment, as well as among curriculum, instruction, and assessments. Progress maps also enable the measurement of growth (continuity). The Australian Council for Educational Research (Forster & Masters, 2001) has produced an excellent set of resource materials for teachers to support their use of a wide range of assessment strategies—from written tests to portfolios to projects at the classroom level—that can all be designed to link back to the progress maps (comprehensiveness), however the underlying models of learning are not as strongly tied to research as they could be and should.

Some Final Thoughts

This is a heady time in education, given interest in developing common core standards for major areas of the curriculum combined with an interest in developing quality assessment programs that can be used to measure student achievement and growth against those standards. All of this work is directed towards a more coherent approach to educational improvement and accountability than has been the case for the last decade. The U.S. Department of Education's current Race to the Top initiative provides an opportunity to rethink how we design assessments to serve multiple purposes and how they can provide the types of information needed by the full range of actors and agents in the educational system from classroom teachers, to district leaders, to state superintendents, to federal policy makers. But with opportunity comes the peril of moving too quickly to implement systems without thinking through the complexity of the information needs of the various users and ways to insure compatibility and coherence across all levels of the system.

There are also serious issues regarding existing capacity to develop the full range of quality assessment materials and tools that are needed to get the job done, certainly if there is a short time horizon to accomplish everything. The policy community wants answers to design and implementation questions that the learning sciences and measurement sciences cannot fully answer. This is especially true for a complex topic like the most valid ways to measure that growth of knowledge and skill across time. It is not that the job can't be done but an investment must be made in research and development that runs the gamut from very basic science to highly applied design, and everything in between. One would hope that the Race to the Top initiative puts us on the right path even if it means that getting to the top takes a bit longer than is comfortable under typical political and legislative cycles. There no doubt will be tensions in working this all out. It is hoped that the conceptual framing of issues and the types of work on student learning described in this paper and can be helpful in steering the course. The education community needs to maintain focus on defining the progressions of knowledge and understanding that are both desirable and attainable and in so doing give substantive meaning to notions such as *fewer, clearer, higher*. And we need to make use of principled assessment design processes that start with careful deliberations about the evidence about student achievement we value and need rather than settling for the data we can most easily and cheaply collect.

References

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

Adams, R.J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47–76.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. Washington, DC: Author.

American Association for the Advancement of Science. (2001). *Atlas of science literacy*. Washington, DC: Author.

Anderson, J. R. (Ed.). (1981). *Cognitive skills and their acquisition.* Hillsdale, NJ: Erlbaum.

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

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

Bransford, J. D., Brown, A. L., Cocking, R. R., Donovan, M. S., & Pellegrino, J. W. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (expanded ed.). Washington, DC: National Academies Press.

Brennan, R. L. (1983). The elements of generalizability theory. Iowa City, IA: American College Testing Program.

Bryk, A. S., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, England: Sage Publications.

Carpenter, T., Fennema, E., & Franke, M. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *Elementary School Journal, 97*(1), 3–20.

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.

Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., & Perlwitz, M. (1991). Assessment of a problem-centered second-grade mathematics project. *Journal for Research in Mathematics Education, 22*(1), 3–29.

Confrey, J. (2008, July). *A synthesis of the research on rational number reasoning: A learning progressions trajectories approach to synthesis*. Paper presented at the 11th meeting of the International Congress of Mathematics Instruction, Monterrey, Mexico.

Confrey, J., Maloney, A. P., Nguyen, K. H., Mojica, G., & Myers, M. (2009, July). *Equipartitioning/splitting as a foundation of rational number reasoning using learning trajectories.* Paper presented at the

33rd Conference of the International Group for the Psychology of Mathematics Education, Thessaloniki, Greece.

Confrey, J., Maloney, A., Nguyen, K. H., Wilson, P. H., & Mojica, G. F. (2008, April). *Synthesizing research on rational number reasoning.* Paper presented at the National Council of Teachers of Mathematics Research presession, Salt Lake City, UT.

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform.* New York, NY: Columbia University, Teachers College, Consortium for Policy Research in Education, Center on Continuous Instructional Improvement.

Delgado, C. (2008, June). *Development of a research-based learning progression for middle school through undergraduate students' conceptual understanding of size and scale*. Paper presented at International Conference of the Learning Sciences.

Delgado, C., Stevens, S., & Shin, N. (2008, April). *Development of a learning progression for students' conceptions of size and scale*. Paper presented at the National Association for Research in Science Teaching Annual Conference, Baltimore, MD.

Delgado, C., Stevens, S., Shin, N., & Krajcik, J. (2008). Development of a learning progression for size and scale. In *Proceedings of 8th International Conference of the Learning Sciences–Vol. 3* (pp. 317–318). Utrecht, Netherlands: ISLS.

DiBello, L., Jiang, H., & Stout, W. F. (1999). A multidimentional IRT model for practical cognitive diagnosis. *Applied Psychological Methods, 3,* 23–32.

Donovan, M. S., & Bransford, J. W. (Eds.). (2005). *How students learn history, science and mathematics in the classroom*. Washington DC: The National Academies Press.

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (Eds.). (1999). *How people learn: Bridging research and practice*. Committee on Learning Research and Educational Practice. Washington, DC: National Academies Press.

Duncan, R. G., & Hmelo-Silver, C. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal for Research in Science Teaching, 46*(6), 606–609.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grade K–8.* Washington DC: The National Academies Press.

Embretson, S. E. (1996). Multicomponent response models. In W. J.van der Linden & R. K.Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*, 543–553.

Fay, A. & Klahr, D. (1996). Knowing about guessing and guessing about knowing: Preschoolers' understanding of indeterminacy. *Child Development, 67*, 689–716.

Forster, M., & Masters, G. (2001). *Progress maps.* Victoria, Australia: Australian Council for Educational Research.

Geary, D. (1995). Reflections of evolution and culture in children's cognition: Implications for mathematical development and instruction. *American Psychologist, 50*(1), 24–37.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate–inanimate distinction as examples. *Cognitive Science, 14*, 79–106.

Gelman, R., & Gallistel, C.R. (1978). *The child's understanding of number.* Cambridge, MA: Harvard University Press.

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: Erlbaum.

Greeno, J.G. (1978). A study of problem solving. In R. Glaser (Ed.) *Advances in instructional psychology* (Vol. 1, pp.13–75). Hillsdale, NJ: Erlbaum.

Griffin, S., & Case, R., (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education, 3*(1), 1–49.

Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika, 55,* 477–494.

Herman, J.L., Wilson, M.R., Shavelson, R., Timms, M., & Schneider, S. (2005, April). *The CAESL assessment model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academies Press.

Hickey, D., & Pellegrino, J.W. (2005). Theory, level, and function: Three dimensions for understanding transfer and student assessment. In J. P. Mestre (Ed.). *Transfer of learning from a modern multidisciplinary perspective* (pp. 251–293*)*. Greenwich, CO: Information Age Publishing.

Janssen, R., Tuerlinckx, F, Meulders, M., & De Boeck, P. (2000). An hierarchical IRT model for mastery classification. *Journal of Educational and Behavioral Statistics, 25*(3), 285–306.

Junker, B. (1999, April). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Retrieved from http://www.stat.cmu.edu/~brian/nrc/cfa/.

Kaiser, M. K., Proffitt, D. R., and McCloskey, M. (1985). The development of beliefs about falling objects. *Perception & Psychophysics, 38*(6), 533–539

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academies Press.

Lane, L, Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a Middle-School Mathematics Performance Assessment. *Educational Measurement: Issues and Practice, 15*(4), 21–28.

Lemke, J. J. (2000). Across the scale of time: Artifacts, activities, and meaning in ecosocial systems. *Mind, Culture, and Activity, 7*(4), 273–290.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Lord, R. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marion, S., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, *Winter 2006,* 47–57. doi: 10.1111/j.1745–3992.2006.00078.x

Massey, C. M., & Gelman, R. (1988). Preschoolers decide whether pictured unfamiliar objects can move themselves. *Developmental Psychology, 24*, 307–317.

Masters, G. N., Adams, R. A., & Wilson, M. (1990). Charting of student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary volume 2* (pp. 628–634). Oxford, England: Pergamon Press.

Masters, G., & Forster, M. (1996). *Progress maps. Assessment resource kit*. Victoria, Australia: Commonwealth of Australia.

Mead, R. (1976). *Assessment of fit of data to the Rasch model through analysis of residuals.* Unpublished doctoral dissertation, University of Chicago.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379–416

Mislevy, R. J., & Wilson, M. R. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41–71.

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

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Research Council. (2003). *Assessment in support of learning and instruction: Bridging the gap between large-scale and classroom assessment.* Washington, DC: National Academies Press.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 49–60). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W. (2009). The challenges of conceptualizing what low achievers know and how to assess their competence. In M. Perie (Ed.), *Considerations for the alternate assessment based on modified achievement standards (AA-MAS): Understanding the eligible population and applying that knowledge to their instruction and assessment.* New York, NY: New York Comprehensive Center.

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 Academies Press.

Pellegrino, J. W., & Hickey, D. (2006). Educational assessment: Towards better alignment between theory and practice. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.). *Instructional psychology: Past, present and future trends. Sixteen essays in honour of Erik De Corte* (pp. 169–189). Oxford, England: Elsevier.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academies Press.

Pirolli, P., & Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review, 105*(1), 588–592.

Quellmalz, E., & Pellegrino, J. W. (2009). Technology and testing. *Science*, *323*, 75–79.

RAND Mathematics Study Panel, & Ball, D. L. (2002). *Mathematical proficiency for all students: Toward a strategic research and development program in mathematics education* (DRU-2773-OERI). Santa Monica, CA: RAND.

Rosenbloom, P., & Newell, A. (1987). Learning by chunking: A production system model of practice. In D. Klahr & P. Langley (Eds.), *Production system models of learning and development* (pp. 221–286). Cambridge, MA: MIT Press.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39,* 369–393.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

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

Siegler, R. S. (1998). *Children's thinking* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Simon, H. A. (1980). Problem solving and education. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 81–96). Hillsdale, NJ: Erlbaum.

Smith, C., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of children's learning for assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement*, *14*(1&2), 1–98.

Snow, C. E., Burns, M., & Griffin, M. (Eds). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.

Starkey, P., & Cooper, R. G. (1980). Perception of numbers by human infants. *Science, 210,* 1033–1035.

Stevens, S. Y., Delgado, C., & Krajcik, J. S. (2009). Developing a hypothetical learning progression for the nature of matter. *Journal of Research in Science Teaching.* Advance online publication. doi: 10.1002/tea.20324.

Stevens, S. Y., Sutherland, L. M., & Krajcik, J. S. (2009). *The big ideas of nanoscale science and engineering: A guidebook for secondary teachers*. Arlington, VA: NSTA Press.

Stiggins, R. J. (1997). *Student-centered classroom assessment*. Upper Saddle River, NJ: Prentice-Hall.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, E .L. (1931). *Human learning*. New York, NY: Century.

Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., & Minogue, J. (2006). Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. *Journal of Research in Science Teaching, 43*(3), 282–319.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1996). *Handbook of modern item response theory.* New York, NY: Springer.

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.

Wiggins, G. (1998). *Educative assessment*: *Designing assessments to inform and improve student performance.* San Francisco, CA: Jossey-Bass.

Wiliam, D. (2007). Keeping learning on track: formative assessment and the regulation of learning. In F. K. Lester, Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich, CT: Information Age Publishing.

Willet, J., & Sayer, A. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*(2), 363–380.

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

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

Wilson, M. R., & Bertenthal, M. W. (Eds.). (2005). *Systems for state science assessments*. 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. 1, pp. 31–74). Washington, DC: American Educational Research Association.