Mapping Student Understanding in Chemistry: The Perspectives of Chemists

JENNIFER CLAESGENS
Graduate Group in Science and Mathematics Education (SESAME),
University of California, Berkeley, Berkeley, CA 94720, USA

KATHLEEN SCALISE
Educational Leadership—Applied Measurement, University of Oregon,
Eugene, OR 97403, USA

MARK WILSON
Graduate School of Education—Quantitative Measurement and Evaluation,
University of California, Berkeley, Berkeley, CA 94720, USA

ANGELICA STACY
Department of Chemistry, University of California, Berkeley, Berkeley, CA 94720, USA

Received 1 November 2006; revised 21 April 2008, 28 April 2008; accepted 1 May 2008

DOI 10.1002/sce.20292
Published online 5 November 2008 in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: Preliminary pilot studies and a field study show how a generalizable conceptual framework calibrated with item response modeling can be used to describe the development of student conceptual understanding in chemistry. ChemQuery is an assessment system that uses a framework of the key ideas in the discipline, called the Perspectives of Chemists, and criterion-referenced analysis using item response models (item response theory (IRT)) to map student progress. It includes assessment questions, a scoring rubric, item exemplars, and a framework to describe the paths of student understanding that emerge.

Correspondence to: Kathleen Scalise; e-mail: kscalise@uoregon.edu

Contract grant sponsor: National Science Foundation.
Contract grant number: DUE 0125651.

Integral to criterion-referenced measurement is a focus on what is being measured: the intention of the assessment, its purpose, and the context in which it is going to be used. The Perspectives framework allows us to begin to narrate the development of understanding that occurs as students "learn" over the course of instruction, helping to form a crosswalk among educational science standards and underscore the importance of scientific reasoning with domain knowledge. Here, we explain a framework we have investigated in chemistry and present evidence on measures of student understanding to describe the development of conceptual understanding at the high school and university levels. *Sci Ed* **93:**56–85, 2009

INTRODUCTION

Assessment is more than just tests or grades. Good assessment entails engaging yourself with your students' work. Over the past 5 years we have been developing the ChemQuery assessment system to describe the paths of student understanding that appear to develop as chemistry knowledge progresses, in the learning of high school and introductory college chemistry. ChemQuery includes assessment questions, a scoring rubric, item exemplars, a framework to describe the paths of student understanding that emerge, and criterion-referenced analysis using item response theory (IRT) to map student progress. Even though we are focused on chemistry, the lessons we are learning are quite general. In this paper, we share some of the design process and results for this kind of assessment, for use in both chemistry and as a view on measurement that might be helpful in other STEM fields.

In the United States, two widely used and cited educational standards documents describing what students should understand and be able to do in chemistry are the National Science Education Standards (National Research Council [NRC]/National Academy of Sciences, 1996) and the Benchmarks for Science Literacy (American Association for the Advancement of Science [AAAS], 1993). Our challenge has been to propose a model of how student thinking develops that integrates standards of chemistry domain knowledge and changes in student knowledge structures on one framework so that conceptual understanding can be described and measured in terms of changes in student knowledge of chemistry. To measure student understanding we use criterion-referenced measurement, which tracks student learning by making people-to-discipline measurements that assess links between student understanding and chemistry domain concepts such as educational standards, rather than the people-to-people comparisons as in norm-referenced assessments. Therefore, to accomplish this has entailed developing a model to organize the overarching ideas of the discipline of chemistry into a framework from novice levels of understanding to graduate and expert levels, which is the Perspectives of Chemists framework within the ChemQuery assessment system.

The Perspectives is a multidimensional framework that describes a hierarchy of student levels of understanding in chemistry. It is built on the theoretical idea that the field of chemistry, at least through high school and college general chemistry, can be largely grouped into three core conceptions, or scientific models: matter, change, and energy, as shown in Table 1. (A fourth model, quantum, is necessary in more advanced general chemistry.) The purpose in framing the "big ideas" of chemistry as perspectives is to organize the overarching ideas of the discipline while simultaneously constructing an instrument for measuring the values of these variables for individual students. In terms of measurement, these big ideas are considered progress variables, which help to indicate how far the student has progressed in their conceptual understanding of a topic. This is explained in more detail when the matter variable is illustrated later in this paper.

We acknowledge that there is no single "right answer" in choosing such variables. The three perspectives chosen are not a unique set, nor are these Perspectives completely

TABLE 1
Overview of the Three Proposed "Big Ideas"

	Perspectives of Chemists
Matter	Matter is composed of atoms arranged in various ways.
Change	Change is associated with the rearrangement of atoms.
Energy	Energy is associated with changes that occur.

independent from each other. While there are certainly other ways to divide the discipline of chemistry into overarching ideas, the usefulness of the approach is only realized when the details for a chosen set are worked out. In other words, the perspectives need to be clarified, and an instrument needs to be developed for measuring student learning gains along each variable. The process we have undertaken has been iterative. We chose a set of perspectives, constructed an instrument (a set of questions to assess student understanding), gathered student responses, and have been refining the perspectives and ChemQuery instrument over several iterations.

Furthermore, the variables of the Perspectives framework represent an instantiation both of understanding the "big" or enduring ideas of a discipline of science (Smith, Wiser, Anderson, & Krajcik, 2006) and of tracing out and expanding on such "key ideas" so that students grasp increasing levels of complexity over time for core concepts (NRC, 2007). Therefore in addition to this proposed organization of topics, the Perspectives approach further suggests that learning in each of these three areas or "strands" is a rich and varied progression. Forms of naive reasoning become more complete, and more consistent understanding of explanatory models of chemistry begins to develop. The ChemQuery assessment system uses this framework of the key ideas in the discipline along with criterion-referenced analysis with item response models to map and describe student progress. Examples of this kind of approach to assessment are described in a NRC report (Pellegrino, Chudowsky, & Glaser, 2001).

Identification of some small discrete number of "big idea" Perspectives, investigated in student learning data with qualitative and quantitative approaches, could be a useful point of comparison for this type of assessment across courses in other STEM disciplines. We have been in discussion with physicists, biologists, and mathematicians on what such perspectives might be in their fields, and similar measurement approaches have been launched in other disciplines (Draney, Wilson, & Pirolli, 1996; Wilson, 2005; Wilson & Sloane, 2000). It is also becoming clear that some big idea perspectives are shared between STEM fields. The matter and energy Perspectives discussed in this paper can be seen as shared to some degree between chemistry and physics. However, other aspects of the big ideas help to make the distinctions between the intellectual content of fields, such as the change perspective discussed in this paper can be seen as quintessentially chemistry, and is described in learning standards generally associated with chemistry objectives and standards. Thus, we discuss the development of this approach to criterion-referenced assessment here.

FOUNDATIONS FOR THE PERSPECTIVES FRAMEWORK

Developing the Perspectives has focused primarily on measuring student conceptual understanding in chemistry. Misconception and alternate conception research in chemistry and on chemistry-related topics such as matter and energy have been previously explored in the literature by numerous researchers over the last several decades. Research syntheses in 1998 (Krnel, Watson, & Glazar, 1998), 2001 (Liu, 2001), and 2003 (Liu, 2001) describe the development of student conceptions of matter. Other research reviews have also been helpful (Krajcik, 1991; Nakhleh, 1992; Wu & Shah, 2004) to summarize studies on where learning difficulties have been documented. Much of the research described in such syntheses is fragmented into different knowledge types. For example, student learning has been deconstructed into the following: problem solving, conceptual change, transfer, representations, and reasoning (Glaser, 1984). Often comparisons in understanding are made, such as problem solving versus conceptual understanding, domain knowledge versus reasoning, and experts versus novices. Research in chemistry education not only falls into these distinct categories but also tends to be topic specific to the domain of chemistry, such as describing student understanding of the mole (Furio & Guisasolo, 2002), thermodynamics (Boo, 1998; Greenbowe & Meltzer, 2003; Teichert and Stacy, 2002), or chemical change (Hesse & Anderson, 1992; Johnson, 2002; Yarroch, 1985).

Our challenge is that when the research looks at just one knowledge type or one specific topic, an overall understanding of how students learn chemistry is missing, resulting in snapshots of student learning rather than a coherent mapping and account of the understanding that develops over the course of instruction. Therefore, the aim of this research is to look at student learning more coherently by integrating what is known about how students learn with the organization of the domain of chemistry within the Perspectives framework.

Therefore, this paper attempts to extend the research on learning by suggesting a framework and a measurement approach by which discrete conceptual learning can cohesively be tied together around big ideas, and measured. The hope is to yield the ability to (i) measure students in reliable and valid ways that are criterion referenced, (ii) explicitly identify the large explanatory models to facilitate student conceptual understanding in relationship to the discrete standards and topics that instructors need to teach, (iii) ultimately make the goals of instruction clear enough that students too can participate in regulating their own understanding through criterion-referenced assessments, and (iv) yield information that is helpful in understanding how pacing, sequence, structure, and other aspects of learning activities might improve student learning outcomes, and whether this is the same or different among a range of students.

The approach of the Perspectives construct is to rely on literature, theory, data, and analysis to describe a hierarchy of chemistry content in relationship to student understanding that then *defines variables* to allow us to measure learning outcomes, *determines scales* for these variables, and *constructs instruments* for measuring the values of these variables for individual students.

Focusing on the progression of conceptual understanding of chemistry has required that we think innovatively about what we want to measure and how we are going to accomplish it. This has involved an extensive discourse between educators, teachers, students, chemists, and measurement professionals, and considerable reflection and revision based on the qualitative and quantitative data we gather using the assessment. Here we share some of that conversation, so that the development process may help others to think about the assessment of key "perspectives" in their fields.

The learning theory behind the development of the Perspectives framework has been primarily constructivist based. We acknowledge both a cognitive constructivist and a social constructivist viewpoint. In terms of cognitive foundations, knowledge in the framework is conceived as a creation of the mind, an active process of sensemaking. Sensemaking can include interacting with the world around us as well as engaging in constructive and instructional activities. Prior knowledge and beliefs are building blocks for sensemaking. Instructionally in terms of key ideas (NRC, 2007), students can construct and build conceptual understanding as they encounter and grasp increasing levels of complexity in concepts over time. Types of thinking that students seem to exhibit while constructing new knowledge are represented in the Perspectives framework. (See also the section on the structure of the learning outcome [SOLO] taxonomy in the Measurement Models and Scoring section.) At the same time, at a more general level, it is acknowledged that the Perspectives variables trace model-based scientific reasoning. Scientific models are constructions of science and of scientists—therefore constructed as models of real-world behaviors by the professional community—and as such they attempt to approximate reality but are not reality. Therefore, both cognitive constructivism and social constructivism are embedded in the framework.

To begin to answer the question of how students develop conceptual understanding of chemistry requires describing what is meant by conceptual understanding and more specifically what is meant by conceptual understanding in chemistry. Chemistry is defined as the study of matter and its transformations. The premise is that understanding the material world in terms of atoms, molecules, and bonds is a powerful model. As Feynman (1963) in his book *Six Easy Pieces* states:

> If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generations of creatures, what statement would contain the most information in the fewest words? I believe it is the atomic hypothesis. . . that all things are made of atoms—little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. In that one sentence, you will see, there is an enormous amount of information about the world, if just a little imagination and thinking are applied. (from the chapter Atoms in Motion, p. 4)

Chemists use the atomic theory to explain the properties of substances, changes in bonding, changes in state, and conservation. Johnson (2002) points out, on the basis of his research on students' understanding of substances and chemical change, "What kind of 'explanation' is there at a macroscopic level which is believable, consistent with the science view of 'a substance,' but ultimately does not owe its sense to particle ideas? (p. 1053).

The development of the particulate nature of matter through the history of science provides insights into the conceptual difficulties that students face in their chemistry class. Like students, early chemists only had observations of materials at the macroscopic level. Students seem to have very similar problems in explaining the natural world as the early scientists. Typically students describe matter "based on immediate perceptual clues (Krnel et al., 1998)." For many students matter is viewed as continuous instead of particulate (Driver, Squires, Rushworth, & Wood-Robinson, 1994; Krnel et al., 1998). Many students even after instruction from weeks to months retain a "concrete, continuous" view of atoms and molecules, in which each particle retains the macroscopic properties of a small piece of its parent substance (Ben-Zvi, Eylon, & Silberstein, 1986). Subjects often believe water molecules, for instance, contain components other than oxygen and hydrogen, such as water, air, chlorine, minerals, and 'impurities,' or may have shapes in different phases, for example, water molecules frozen into ice cubes are square (Griffiths & Preston, 1992). Individual molecules can be "hot" or "cold," and belief in atoms and molecules as alive is common (de Vos & Verdonk, 1985). The implication of this thinking is that a single molecule of water (H_2O) is viewed as a teeny tiny teardrop.

In addition, if that piece is no longer perceptible it has "disappeared." There is no accounting for conservation of mass or atoms. Furthermore, gaseous products or reactants are often ignored (Hesse & Anderson, 1992), reflecting the view that substances that float in air have no mass and thus are not substances that need to be conserved (Samarapungavan &

Robinson, 2001). Even after university-level instruction, most students do not understand chemical reactions as reorganization of atoms, with breaking and reformation of bonds. For instance, in one study only 6% of secondary and 14% of university chemistry students could, after instruction, describe chemical reactions as the breaking and reforming of bonds (Ahtee & Varjola, 1998). This lack of a particulate view of matter affects how student understanding of chemistry develops.

Students rarely define matter by volume and weight (Stavy, 1991). Students can manipulate equations to balance them without thinking about atoms or conservation of mass (Yarroch, 1985) or they have difficulties explaining combustion reaction because gases are involved that are not perceived as matter (Johnson, 2002). The implication is that conceptual understanding in chemistry requires knowledge of the atomic theory and that lack of understanding of the particulate nature of matter affects students' understanding of properties and chemical change of substances (Krnel et al., 1998; Johnson, 2002). However, if the ultimate goal is for students to reason like a chemist with a particulate view, this necessitates thinking beyond how the content is typically organized in the domain to thinking about how domain knowledge is organized in the student's head.

Driver and other researchers have emphasized the interplay among the various factors of personal experience, language, and socialization in the process of learning science in classrooms and argue that it is important to appreciate that scientific knowledge is both symbolic in nature and also socially negotiated (Driver, Asoko, Leach, Mortimer, & Scott, 1994; Driver & Scanlon, 1989; Driver, Squires, et al., 1994). By socially negotiated, these researchers mean that scientific entities and ideas are unlikely to be discovered by individual students through their own empirical enquiry, so learning science involves being initiated into the ideas and practices of the scientific community.

Hesse and Anderson (1992) have argued that it takes time to build sufficient understanding to be able to combine scientific models with prior knowledge and develop working understanding on which knowledge of chemistry can build. They argue that while the rules for writing and balancing chemical equations are fairly simple, the equations that result are meaningful only when they are embedded in a complex "conceptual ecology"—an array of facts, theories, and beliefs about the nature of matter and the functions of explanation that chemists have developed over time, and that is part of the discourse language in chemistry.

Beyond documenting that conceptual understanding is not easily achieved and often results in student misconceptions, how conceptual understanding develops, and what develops is not agreed upon. Conceptual understanding is described as understanding beyond rote memorization of facts and algorithms. The shift is from student answers limited to scientific terms or formulaic solutions to student explanations of their understanding in terms of their ability to integrate the new domain knowledge, relate between ideas, and use the ideas to explain and predict natural phenomenon (National Research Council/National Academy of Sciences, 1996). Glaser (1984) explains that historically knowledge has been described as problem solving, domain knowledge, and cognitive processes but argues that these different forms of knowledge should be integrated into theories of learning and instruction to move students from basic skills to higher order thinking. Thus, the argument could be made that conceptual understanding should potentially include aspects of conceptual change, reasoning, problem solving, knowledge of facts, and even transfer. Unfortunately, each of the different knowledge types comes from different research agendas that attend to different analytical frameworks. For example, transfer is described in terms of "near" or "far" transfer and looks at the analogical mapping from the source problem to the novel problem situation (Larkin, 1989). Reasoning is argued in terms of general strategies and processing skills with and without domain knowledge (Metz, 1995; Kuhn, 1989). Research on problem solving attends to strategic and procedural knowledge processes (Kotovsky, Hayes, & Simon, 1985) while the debate in conceptual change theory focuses on whether students have theories or uncoordinated knowledge pieces (Carey, 1991; diSessa, 1993; diSessa & Sherin, 1998; Vosniadou & Brewer, 1992). By framing learning in terms of conceptual understanding we seek to emphasize the construction of understanding beyond rote memorization and procedures to recognize the dynamic between what is desired versus what is achieved by the students.

Furthermore, research on various aspects of these interactions suggests that it is not an equal part of one type of knowledge or another knowledge component that results in the understanding that develops. There is a complex interaction of these components that form the student's knowledge base. For example, if students do not have domain knowledge to incorporate into their explanation then they will use either surface features of the problem (Chi, Feltovich, & Glaser, 1981) or rely on experience from their everyday lives (as evidenced in the misconceptions literature). Other research in science education has studied how students use and interpret empirical data and evidence. An example from our preliminary findings might help to illustrate the interaction between domain knowledge and reasoning that can occur when students are asked to explain an empirical observation. Responses from high school chemistry students were collected for the following question:

> Both of the solutions have the same molecular formulas $C_4H_8O_2$, but butyric acid smells bad and putrid while ethyl acetate smells good and sweet. Explain why these two solutions smell differently.

A few students (and some teachers) respond that this cannot be true, that the question must be written wrong, or more specifically there is a mistake in the molecular formulas. Lacking an understanding of molecular structure, they disregard the data and dispute it in favor of their own conceptions. It seems that many factors could be attributed to the student's understanding relating to prior knowledge and or lack of domain knowledge. Nevertheless, there is an interaction of domain knowledge, everyday notions, and reasoning that come to play as students construct their understanding and that these compose the resources that students access when forming an explanation. Therefore, the way that students make sense of the world is due to the coordination of these components into the understanding that develops.

Understanding the patterns by which concepts, experience, representations, domain knowledge, and reasoning among other conceptual change mechanisms, interact to generate a working knowledge structure is a daunting task.

What we want to see develop in chemistry education is that students rely less on information from their everyday lives and begin to access and use chemistry in their reasoning about matter and its transformations. However, the research suggests that it is not clear when students use data, experience, or domain knowledge in their explanations. Moreover, reasoning with domain knowledge is difficult for students. Research on student misconceptions substantiates the claim that most student explanations in chemistry are not correct. For example, students do not use the particulate nature of matter to explain the macroscopic properties of matter (Driver et al., 1994; Krnel et al., 1998), and they also do not relate mathematical symbols to coherent chemistry models of understanding (Lythcott, 1990; Nurrenbern & Pickering, 1987). Recognizing that conceptual understanding is not described by one single knowledge type and that many components seem to be at play as a student develops conceptual understanding over the course of instruction, the challenge then is to propose a model of how student thinking develops by constructing a framework that integrates chemistry domain knowledge and changes in student knowledge structures on one framework so that conceptual understanding can be described and measured in terms of changes in student knowledge of chemistry. To this end, the ChemQuery project has used observational studies and measures on longitudinal task assessments to refine the development of an organizing framework, called Perspectives of Chemists, that defines core chemistry concepts and emphasizes the integration and coordination of students' emerging chemistry knowledge. An important aspect of this research has been how IRT has been used to inform our understanding of how student conceptual understanding of chemistry develops. The next section describes the development of the Perspectives framework.

The Perspectives of Chemistry as a "Construct"

Describing what we want students to learn and what successful learning patterns should look like for particular objectives is always a challenge. One approach is to capture not only what mastery should look like but also the patterns of progress of student understanding toward mastery. This is the construct intended to express what we want to measure.

Often in educational and psychological research, the term "construct" refers to any variable that cannot be directly observed but rather is measured through indirect methods (Messick, 1995; Wilson, 2005), such as intelligence or motivation. We might think that student understanding of key ideas in a discipline such as chemistry could be directly measured and would not need to involve constructs. However, understanding how students approach the thinking to be discussed in this paper, such as their developing perspective of the atomic view of matter or the increasing complexity and sophistication of problem-solving strategies surrounding reactivity, may involve latent cognitive processes for which some observations can be taken and perhaps an underlying mental model, knowledge state or knowledge structure might be considered to be inferred.

This may seem contradictory to how the term construct is used in some other papers in the literature, but we believe it instead addresses the misconception that we can really tap student understanding of this type as a directly manifest variable. Using the language of construct also reminds us that when assessing student understanding, we are often attempting to measure thinking and reasoning patterns that are not directly manifest—we do not "open" the brain and "see" reasoning—but instead the observables such as answers to questions and tasks give us information on what that reasoning, or latent construct, may be. This is true whether we think of reasoning in terms of knowledge structures, knowledge states, mental models, normative and alternate conceptions, higher order reasoning, migratory atheoretical patterns, or any of a number of other possible cognitive representations—all are essentially latent. They can be observed not directly but through manifestations that must be interpreted. This thinking is similar to the way "construct" is used in measurement contexts to refer to "construct validity" in assessment (Messick, 1995). Messick explains that construct validity is based on integration of evidence that bears on the interpretation or meaning of test or assessment scores, such that content- and criterion-related validity evidence can be subsumed under the concept of construct validity. In construct validation, the test score is seen as attempting to measure the *construct* it attempts to tap and is one set of indicators of the construct (Cronbach & Meehl, 1955). Messick explains that construct representation refers to the relative dependence of task responses on processes, strategies, and knowledge to be measured that are implicated in task performance.

The Perspectives construct developed for this project attempted to specify some important aspects of conceptual understanding in chemistry. The construct to be measured is often established or adopted by the people who design and develop the learning goals to be assessed with the instruments (Wilson, 2005), informed by standards, research literature, empirical evidence, and other guidelines. Classroom-based assessment often begins in current U.S. educational practice with addressing a series of many discrete standards. The construct helps to theoretically structure and describe the relationships among the pieces of knowledge, as well as helping to clarify considerations when specifying items, or in other words, assessment questions and tasks; scoring; and what measurement models might be appropriate to use for criterion-referenced assessment.

As instructors, or test developers, one often addresses standards in science education and then starts with the questions and tasks one wants students to be able to perform, as a way to measure at what capacity they can perform. Information is being collected but the assessment started with assessment questions rather than focusing on describing a model of the development of understanding. In comparison, with the construct, student responses are analyzed and student understanding is valued or scored based on not just what we want, but how we are going to help the students get there. The construct allows us to intentionally acknowledge that what we teach, despite our best efforts, is not what students learn or how they learn. Moreover, the construct allows us to begin to narrate the development of understanding that occurs as students "learn" over the course of instruction by providing a frame of reference to talk about the degrees and kinds of learning that take place. This is important both for the student as they learn about the domain and for the teacher as they plan strategies to build upon the current level of student understanding.

MIXED METHODS

This approach to research on student understanding allows us to use both qualitative elements and quantitative measures to test and refine hypotheses about how learning develops. This approach to measurement is really a mixed or multiple methods approach, with qualitative and quantitative information iteratively informing each other. Qualitative information in our process begins the cycle with analysis of student work through such approaches as classroom observations, cognitive task analysis, and the emergence of patterns in student response data through a process of phenomenography. Scoring and data analysis form a quantitative phase that yields information on where additional complexity is seen in the student learning data. Subject matter experts and others then return to the identified areas of complexity and further qualitatively explore how learning patterns might be understood, with techniques such as interviewing, verbal protocol analysis, continued classroom observations, and analysis of video and audio recordings.

Therefore, integral to developing a criterion-referenced measurement tool are the iterations in the design process that has allowed for the testing and refinement of the theoretical framework being measured. In this paper, we focus primarily on the quantitative aspects of the framework explication and exploration. Further discussion of some of the qualitative aspects has been previously reported (Scalise et al., 2004; Watanabe, Nunes, Mebane, Scalise, & Claesgens, 2007).

Measurement Models and Scoring

In U.S. chemistry classrooms, we often know where particular students fall relative to one another, but we lack an understanding of where each student fits on a scale of higher order thinking in the discipline. Moreover, we tend to test knowledge in discrete pieces. That is, there is no model or frame to organize the questions we use to assess student understanding. This method of making unlinked measurements to monitor student learning is in contrast to measurement strategies we use as scientists. For example, when talking about the weather, we could say it is warmer than previously (a norm-type comparison). We could describe the clouds in the sky (discrete knowledge). Or, we could develop a *model* about what factors influence weather, and then use the model to decide on what *variables* to measure (e.g., temperature, pressure, wind speed, and so on). For each variable, we can begin to define a *scale* for measurement (e.g., a temperature scale based on the phase changes of water). An essential part in developing the scales for each variable in the model is to design *instruments* for tracking changes in weather (e.g., a thermometer, a barometer, an anemometer, and so on). This measurement strategy is referred to as criterion-referenced measurement.

Item response models are used to model the data for the analysis in this paper. In general, item response models are statistical models that express the probability of an occurrence, such as the correct response on an assessment question or task, in terms of estimates of a person's ability and of the difficulty of the question or task. These models include the one-parameter Rasch model for dichotomous data, and related Rasch family models for polytomous data such as the partial credit model used in this paper, as well as two-parameter, three-parameter, and other models. The additional parameters can allow additional complexities to be modeled. The choice among models (Wilson, 2005) can involve such considerations as the type of data available, the degree to which items discriminate in some similar ways, the use of raters, the potential for effective random guessing in the item format, and other concerns.

Rasch family models are item response models (1PL or one-parameter logistic), and are usually considered for use in relationship to 2PL and 3PL models. 2PL models draw on an item discrimination parameter. 3PL models draw on an additional guessing parameter. As defined in Wilson's BEAR approach, Rasch family models must be used in the case of construct mapping as used in this paper, rather than 2PL or 3PL item response models. This is because Rasch family models have scale properties that preserve item difficulty locations on construct maps across student proficiency distributions, necessary for the mapping technique. In other words, location on the maps generated from this technique shows the distance between respondents and responses, and indicates the probability of making any particular response for a particular student. This condition necessitates that the scaling of an item must remain the same for all respondents (Wilson, 2005). Assessment tasks and questions of course must be evaluated for good fit to such models (see Results section). BEAR assessment addresses item discrimination issues that could lead to the need for the more parameterized models by generating item fit statistics in the item analysis process. These indicate when assessment tasks and items well fit the construct, and when the tasks and items may need to be adjusted or reworked to appropriately meet evidentiary requirements. Random guessing issues can be addressed by examining plots of the item response data. However, random guessing is less an issue in this project as tasks generally demand constructed responses that do not lend themselves to guessing.

Item response models help provide a basis for evaluating how well an assessment instrument is performing by generating validity and reliability evidence, and also generate estimates of how precise a student score is likely to be and therefore the confidence with which it can be interpreted. The item response model used in this paper is the partial credit model, which can be used to score students on small bundles of items that share common stimulus material (Rosenbaum, 1988; Wilson & Adams, 1995). The shared context introduces dependencies within the group of items, as is often the case with diagnostic e-learning assessments. This may affect student scores so with the partial credit model one score is generated for the set of common items rather than treating the questions as if they were separate, independent scores.

The partial credit model is the more general of two polytomous Rasch models (Wright & Masters, 1982) commonly expressed according to Eq. (1):

$$P(X_{is} = x|\theta_s) = \frac{\exp \sum_{j=0}^{x} (\theta_s - \delta_{ij})}{\sum_{r=0}^{m_i} \exp \sum_{j=0}^{r} (\theta_s - \delta_{ij})} \tag{1}$$

for item i scored $x = 0, \ldots, m_i$, where X_{is} is the score of student s on item i, x represents a given score level, θ_s represents the parameter associated with the person performance ability of student s, r in the denominator represents a summation of terms over the steps, and δ_{ij} represents the parameter associated with the difficulty of the jth step of item i.

Since the partial credit model represents a summation of terms over the steps, the issue of accumulating credit in a sequential scoring model arises. With alternate or misconceived reasoning, in using the partial credit model it is assumed that students who are able to establish and explain the correct reasoning can be considered on average to have surpassed naive reasoning that is in conflict with the correct reasoning. Therefore, students have "achieved" the prior scoring step by surpassing it in understanding (assuming good support and fit for the construct). Students therefore are not expected to actually have procedurally compared incorrect and correct answers and worked backward on each item, explicitly rejecting the incorrect answer, but rather that we posit a theoretical concept of something like a knowledge state that when achieved can be considered to supersede the incorrect reasoning. This is consistent with a cognitive constructivist perspective of the learner building understanding. Of course, qualitative data along with fit statistics and other evidence help support this assumption.

Using IRT, the scores for a set of student responses and the questions are calibrated relative to one another on the same scale and their fit, validity, and reliability are estimated, and matched against the framework. In this paper, we present the final results of the framework for the matter variable.

Through rounds of qualitative and quantitative data collection, analysis, and interpretation of findings by the experts group, the framework and measurement approaches were established and refined. Small groups of students would complete exploratory items and tasks, and were involved in such processes as cognitive task analysis and think alouds. Extensive qualitative data were collected, including student interviews and observations. Student written responses were analyzed, rubrics and examples of student work developed, and issues identified in the process.

Refinements we found necessary to the framework often took the form of more carefully accounting for differences in student conceptual understanding, such as distinguishing between those students who could use simple rules or "name" something, but could not successfully apply this knowledge. Also we found that the framework needed to better account for attempts to use chemistry based on an incorrect atomic view. For example, many students use the terms atom, proton, and electrons without understanding their meaning. All students exhibiting these various behaviors could simply be scored as wrong or not successful on an item, but often we found that would be to ignore interesting information, which seemed to show a variety of different levels of student ability prior to achieving a traditional level of "success" on the item. Also scoring guides and exemplars needed better to capture the process skills associated with scientific inquiry, such as observing, reasoning, modeling, and explaining (AAAS Project 2061, 1993; National Research Council/National Academy of Sciences, 1996).

The iterative process proved very important to the ultimate form of the framework and the assessment system components. In general, the steps consisted of the following:

- Using a well-balanced experts group and attention to the research literature in the area to be measured, propose some "big ideas" and general outlines for an assessment framework.
- On the basis of the current framework ideas, examine the curriculum to determine appropriate topic areas and contexts to consider for measuring, in this case, developing conceptual understanding of matter.
- Explore item-type possibilities and scoring approaches with curriculum developers, content experts, instructors, and project participants, and develop sets of items and rubrics.
- Test items on a small informant population of students.
- Collect feedback on items during an item paneling and scoring review session, and attempt to develop a body of exemplars, or examples of student work, at each level of scoring.
- Qualitatively analyze the results of scoring.
- Refine scoring rubrics and engage in or adjust approaches to rater training for the larger body of data collection to come.
- Collect and analyze a larger set of diverse student data through both qualitative and quantitative approaches that can be analyzed with the use of measurement models.
- Use the results of this larger data collection to refine the framework, improve items, evaluate and update the scoring approaches, and assess the appropriate use of measurement models.
- Repeat the process using the updated approaches, and iteratively engage in further rounds of student data collection. Note that a "saturation" or discount engineering evaluation approach (Landauer, 1991; Nielsen, 1994; Nielsen & Mack, 1994) can help decide when enough information has been collected for the purposes at hand. Such approaches help us understand when enough cases have been considered, or sufficient rounds of improvement completed, in an evaluation process. This is done by looking at the extent to which substantial new knowledge is being gained at each new iteration of the process, and whether much of the key information has been collected previously. (Of course, with unlimited resources, unlimited rounds of iteration would no doubt be useful and would generate some new knowledge, but discount engineering helps to balance available resources with information discovery needs.)

In the qualitative data collection stages, patterns were identified and answers were grouped to reflect similarities in thinking approaches and strategies. Students also participated in observations and interviews. A similar approach as has been employed in physics was used to identify reasoning for chemistry conceptual understanding (Minstrell, 1989, 2001; Minstrell, Anderson, Minstrell, & Kraus, in press). Cognitive taxonomies such as Bloom's taxonomy were also explored for their contribution relative to the chemistry subject matter area, and a literature synthesis of conceptual change research in chemistry education was completed to further inform results (Scalise, 2001).

For the Perspectives framework, construction of performance levels for all three LBC variables followed a generalizable pattern loosely similar to concepts associated with the SOLO taxonomy (Biggs & Collis, 1982). Similar to developmental psychology models that allocate students to Piagetian stages, SOLO allocates student observable responses on assessment tasks to a hierarchy of stages. The SOLO hierarchy is based on five levels, with level three containing additional sublevels. An overview is shown below, with some modifications for our purposes:

1. Prestructural: Student answer is an irrelevant response to the assessment task.
2. Unistructural: Student response focuses on single aspect of information available.
3. Multistructural: Student response uses multiple aspects of information available.

   a. Simple multistructural
   b. Fragments of relational
   c. Relational but wrong

4. Relational: Student takes the structure of information described in 3 and relates it to aspects of external information in one or more other structures, schemas, or scripts.
5. Extended abstract: Response draws on and relates these structures to additional information and concepts either (i) not supplied in learning materials or (ii) generative.

Since it appeared that the SOLO taxonomy (Biggs & Collis, 1982) helped to capture some of the trends seen in the student data (discussed below), it was loosely consulted to help with revisions to the matter and change variables at this point. Perhaps the most influential idea from the SOLO taxonomy described the construction of levels of understanding from a unistructural level focusing on one aspect of information followed by a multistructural level where students relate multiple aspects of information available, to a relational level in which multiple structures of information are integrated.

This aspect of the SOLO taxonomy helped reintegrate valuing scientific habits of mind, while retaining the distinctions of domain knowledge use. Where higher levels of cognitive reasoning such as explaining or evaluating might appear to have been achieved in a student response, the answers operating with less domain knowledge could still be viewed as less correct answers in the structural components and relationships, and this could be specified in scoring rubrics.

The SOLO taxonomy also helped tease apart qualitative and quantitative understanding as possible to frame as different "multistructural units," either of which could be valued independently at one level of the framework. Then the relational aspect of using both qualitative and quantitative understandings, either partially or fully correctly, could be valued as a more advanced performance. SOLO also allowed for the vast body of misconception research to be well situated within the framework. Some misconceptions could be seen as unistructural, or involving only incorrect student responses on single aspects of information, while others were more apparently relational, involving correct ideas about each aspect with the misconception coming in an incorrect relationship among the ideas. Thus, misconceptions were not confined only to the simplest ideas in the answers of lowest achieving students, but depending on the type of misconception could be found in areas of varying complexity throughout the framework (and even among professional scientists, where the state of the knowledge may have embedded an incorrect relationship into working scientific models, yet to be discovered).

The SOLO taxonomic approach has been mapped by its authors to neo-Piagetian learning theory. Our learning theory as described above in the Foundations section draws more broadly on elements both of cognitive and social constructivism. But to what degree more specific learning theories such as Piagetian stages, information processing, or cognitive load theories may be contributing to progressions of student learning in this situation we believe is a subject for future work. To us it seems quite possible that a SOLO taxonomic view might arise out of several of these competing theories, or to be an intersection of theories.

ChemQuery Perspectives Framework

The current ChemQuery assessment system consists of the Perspectives framework shown in Table 2. A fuller description of one variable, matter, is shown in Table 3. A scale of student progression in understanding for use across the variables is shown in Table 4. Along the horizontal axis in Table 2 are the three dimensions of domain knowledge and along the vertical axis is the perceived progression of explanatory reasoning that develops as students gain understanding in chemistry. The emphasis is on understanding, thinking, and reasoning with chemistry that relates basic concepts (ideas, facts, and models) to analytical methods (reasoning). Simply stated, the aim of the organization framework is to capture how students learn to reason like chemists as they develop explanatory models of understanding in chemistry.

It should be noted that other big ideas besides matter, change, and energy would likely need to be included for some chemistry topics, such as quantum chemistry.

TABLE 2
Outline of the Three Current Perspectives Variables

Levels (Low to High)	Matter	Change	Energy
1. Notions	What do you know about matter? (initial ideas, logic, real-world knowledge)	What do you know about change? (initial ideas, logic, real-world knowledge)	What do you know about energy in relationship to chemistry?
2. Recognition	How do chemists describe matter?	How do chemists describe change?	How do chemists explain energy transfer for chemical and physical changes?
3. Formulation	How can we think about interactions between atoms?	How can we think about rearrangement of atoms?	How can chemists predict the extent of energy transfer and the resulting products of chemical and physical change?
4. Construction	How can we understand composition, structure, properties, and amounts?	How can we understand type, progression, and extent of change?	How do different models explain the chemical and physical changes that may occur?
5. Generation	What new experiments can we design to gain a deeper understanding of matter?	What new reactions can be designed to generate desired products?	How can we use these models to optimize this type of change?

TABLE 3
ChemQuery Assessment System: Perspectives of Chemists on Matter

Level of Success	Big Ideas	Descriptions of Level	Item Exemplars
Generation 5: What new experiments can we design to gain a deeper understanding of matter?	Bonding models are used as a foundation for the generation of new knowledge (e.g., about living systems, the environment, and materials).	Students are becoming experts as they gain proficiency in generating new understanding of complex systems through the development of new instruments and new experiments.	*Composition*: What is the composition of complex systems? (e.g., cells, composites, and computer microchips) *Structure*: What gives rise to the structure of complex systems? (e.g., skin, bones, plastics, fabrics, paints, and food) *Properties*: What is the nature of the interactions in complex systems that accounts for their properties? (e.g., between drug molecules and receptor sites, in ecosystems, and between device components) *Quantities*: How can we determine the composition of complex systems? (e.g., biomolecules and nanocomposites)
Construction 4: How can we understand composition, structure, properties, and amounts? (Using models)	The composition, structure, and properties of matter are explained by varying strengths of interactions between particles (electrons, nuclei, atoms, and ions, molecules) and by the motions of these particles.	Students are able to reason using normative models of chemistry, and use these models to explain and analyze the phase, composition, and properties of matter. They are using accurate and appropriate chemistry models in their explanations, and understand the assumptions used to construct the models.	*Composition*: How can we account for composition? *Structure*: How can we account for 3-D structure? (e.g., crystal structure and formation of drops) *Properties*: How can we account for variations in the properties of matter? (e.g., boiling point, viscosity, solubility, hardness, pH, and so on) *Amount*: What assumptions do we make when we measure the amount of matter? (e.g., nonideal gas law and average mass)

Continued

TABLE 3
Continued

Level of Success	Big Ideas	Descriptions of Level	Item Exemplars
Formulation 3: How can we think about interactions between atoms? (Multirelational)	The composition, structure, and properties of matter are related to how electrons are distributed among atoms.	Students are developing a more coherent understanding that matter is made of particles and the arrangements of these particles relate to the properties of matter. Their definitions are accurate, but understanding is not fully developed so that student reasoning is limited to causal instead of explanatory mechanisms. In their interpretations of new situations, students may overgeneralize as they try to relate multiple ideas and construct formulas.	*Composition*: Why is the periodic table a roadmap for chemists? (Why is it called a “periodic” table?) How can we think about the arrangements of electrons in atoms? (e.g., shells and orbitals.) How do the numbers of valence electrons relate to composition? (e.g., transfer or sharing of electrons) *Structure*: How can simple ideas about connections between atoms (bonds) and motions of atoms be used to explain the 3-D structure of matter? (e.g., diamond is rigid, water flows, and air is invisible) *Properties*: How can matter be classified according to the types of bonds? (e.g., ionic solids dissolve in water, covalent solids are hard, and molecules tend to exist as liquids and gases) *Amount*: How can one quantity of matter be related to another? (e.g., mass/mole/number, ideal gas law, and Beer’s law)

Continued

TABLE 3
Continued

Level of Success	Big Ideas	Descriptions of Level	Item Exemplars
Recognition 2: How do chemists describe matter? (Unirelational)	Matter is categorized and described by various types of subatomic particles, atoms, ions, and molecules.	Students begin to explore the language and specific symbols used by chemists to describe matter. They relate numbers of electrons, protons, and neutrons to elements and mass, and the arrangements and motions of atoms to composition and phase. The ways of thinking about and classifying matter are limited to relating one idea to another at a simplistic level of understanding.	*Composition*: How is the periodic table used to understand atoms and elements? How can elements, compounds, and mixtures be classified by the letters and symbols used by chemists? (e.g., $CuCl_2$ (s) is a blue solid and $CuCl_2$(aq) is a clear, blue solution) *Structure*: How do the arrangements and motions of atoms differ in solids, liquids, and gases? *Properties*: How can the periodic table be used to predict properties? *Amount*: How do chemists keep track of quantities of particles? (e.g., number, mass, volume, pressure, and mole)
Notions 1: What do you know about matter? (Initial ideas)	Matter has mass and takes up space. It can be classified according to how it occupies space.	Students articulate their ideas about matter, and use prior experiences, observations, logical reasoning, and knowledge to provide evidence for their ideas. The focus is largely on macroscopic (not particulate) descriptions of matter.	*Composition*: How is matter distinct from energy, thoughts, and feelings? *Structure*: How do solids, liquids, and gases differ from one another? *Properties*: How can you use properties to classify matter? *Amount*: How can you measure the amount of matter?

More detailed description of one Perspectives variable, matter. Variables in the Perspectives framework can be considered a type of "progress variable," to monitor student progress or mastery of conceptual understanding in particular key areas.

TABLE 4
Scoring for the Matter Variable

Level 0: Prestructural
Student response is irrelevant to question (blank, "I don't know," doodle with no distinguishable words or diagrams).

Level 1: Notions
Describes materials or activity observed with senses; compares and contrasts, or generates logical patterns but without employing chemical concepts; using properties of matter as evidence for misconceptions of chemical explanations.

Level 2: Recognition
Explores meaning of words, symbols, and definitions to represent properties of matter; represents matter through arrangements of atoms as discrete particles; translates information represented in the periodic table to an atomic model that includes protons, neutrons, and electrons; and interprets simple patterns of the periodic table.

Level 3: Formulation
Recognizes that matter has characteristic properties due to the arrangement of atoms into molecules and compounds; describes chemical bonds as interaction of valence electrons in atoms; combines individual atoms to make molecules in patterns of bonding based on characteristic atomic properties; and interprets how electrons are shared, shifted, or transferred depending on atoms and types of chemical bonds formed.

Level 4: Construction
Explains molecular behavior and properties in terms of stability and energies involved in intra- and inter-molecular bonding; recognizes that changes in energy can change the condition/properties of matter; predicts effects of transfer of energy; relates energy to the motion and interaction of molecules; and explains changes in matter based on the energy required to break bonds.

The assessment system also includes 30 open-ended items in the matter variable, 20 in change, and 20 in energy and about 12 open-ended and computer-based items (Scalise, 2004; Scalise, Claesgens, Wilson, & Stacy, 2006b), as well as a scoring rubric and item exemplars. IRT analysis with the program GradeMap (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2006), now called ConstructMap, includes generation of maps, item fits, student progress reports, individual and class "growth" maps, and reliability and validity statistics (Wilson & Scalise, 2003).

ChemQuery Scale: Levels of Student Understanding

Within each of the Perspectives, a scale to describe student understanding was proposed, as shown in Table 4. The levels within the proposed variables are constructed such that students give more complex and sophisticated responses as they develop from describing their initial ideas in Level 1 (notions), to relating the language of chemists to their view of the world in Level 2 (recognition), to formulating connections between several ideas in Level 3 (formulation), to fully developing models in Level 4 (construction), to asking and researching new scientific questions in Level 5 (generation). Note that generation is not yet described in the scoring levels, as we have not yet assessed student work samples at this level.

Advancement through the levels is designed to be cumulative. In other words, students measured at a score of 2 in the recognition level are expected to be able to describe matter

Sample Problem and Scoring Guide

"You are given two liquids. One of the solutions is butyric acid with a molecular formula of $C_4H_8O_2$. The other solution is ethyl acetate with the molecular formula $C_4H_8O_2$. Both of the solutions have the same molecular formulas, but butyric acid smells bad and putrid while ethyl acetate smells good and sweet. Explain why you think these two solutions smell differently."

	0	Response: "I have absolutely no idea." Analysis: Response contains no information relevant to item.

Notions	1-	Response: "Just because. That doesn't seem possible. How can they be different when they have the same molecular formula?" Analysis: Student makes one macroscopic observation by noting that the molecular formulas in the problem setup are the same.
	1	Response: "Using chemistry theories, I don't have the faintest idea, but using common knowledge I will say that the producers of the ethyl products add smell to them so that you can tell them apart." Response: "Just because they have the same molecular formula doesn't mean they are the same substance. Like different races of people: black people, white people. Maybe made of the same stuff but look different." Analysis: These students use ideas about phenomena they are familiar with from their experience combined with logic/comparative skills to generate a reasonable answer, but do not employ molecular chemistry concepts.

	1+	Response: "Maybe the structure is the same but when it breaks into different little pieces and changes from liquid into gas they have a different structure in the center and have a different reaction with the air. (Shows drawing:) butyric acid liquid breaking Ethyl acetate liquid breaking Analysis: This answer acknowledges that chemical principles or concepts can be used to explain phenomena. Attempts are made to employ chemical concepts based on a "perceived" but incorrect understanding of chemistry in the first example and solely employing chemical jargon in the second example.

Recognition	2-	Response: "I think these two solutions smell different is because one chemical is an acid and most acids smell bad and putrid while the ethyl acetate smells good and sweet because its solution name ends with "ate" and that usually has a good sweet smell." Analysis: This response correctly cites evidence for the difference in smells between the two chemicals, appropriately using smell combinatorial patterns taught in class and chemical naming conventions, but does not explain the root cause as the difference in molecular structure between the two chemicals.
	2	Response: "They smell differently b/c even though they have the same molecular formula, they have different structural formulas with different arrangements and patterns." Analysis This response appropriately cites the principle that molecules with the same formula can have different structures, or arrangements of atoms within the structure described by the formula. However it shows an incomplete attempt to use such principles to describe the simple molecules given in the problem setup.
	2+	Response: (Begins with problem setup below, showing molecular formula of labeled butyric acid and same formula labeled ethyl acetate.) $C_4H_8O_2$ - butyric acid $C_4H_8O_2$ - ethyl acetate "The two molecules smell differently because the have different molecular structures. The butyric acid contains a carboxylic acid structure (which smells bad) and the ethyl acetate contains an ester (which smells good). We can tell which molecule will smell bad and which will smell good by studying the molecular structure and by looking at the names. Any 'ACID' ending name will smell bad and any '-ATE' ending name will smell good." Analysis: Response cites and appropriately uses the principle that molecules with the same formula can have different structures. Student correctly cites rule learned in class pertaining to smell patterns in relation to functional groups identified by chemical name, and uses this information to begin to explore simple molecules. However, student stops short of a Level Three response, which could be made by examining structure-property relationships through, for instance, presenting possible structural formulas for the two chemicals and explaining the bonding involved.

Figure 1. Example item and scoring guide.

```
--------------------------------------------------------------------------------------------------------------------------------
Wright Map (MLE)    Variable: matter
================================================================================================================================
logit        students                  Thurstonian Thresholds by Item Number
                          | 4  7   11   12   14a  16   22a  22b  23   25   30a  33   35   37   39a  39c  40   43b  44   63
--------------------------------------------------------------------------------------------------------------------------------
4  |                      |   7.8                                                              39A.6
   |                      |                                                                                        44.6
   |                     X|
   |                      |                                                               37.5
3  |                      |
   |                     X|4.6&7        12.6                               30A.6
   |                     X|                  14A.6                         30A.5
   |                   XXX|                                                     33.8
2  |                    XX|                            22A.6                                                            63.8
   |                 XXXXX|   7.7                 16.5                                              39C.8     40.6&7
   |               XXXXXXX|   7.6  11.5                                         33.7                                    63.6&7
   |                   XXX|4.5          12.5                22B.5     25.7 30A.4               39A.539C.6&7 40.5   44.5
1  |         XXXXXXXXXXXXX|                            22A.5     23.5           33.6 35.8      39A.4          43B.5
   |       XXXXXXXXXXXXXXX|   7.5            14A.4&5             23.4 25.6      33.5 35.6&7         39C.540.4
   |           XXXXXXXXXXX|4.4     11.4 12.4                          25.5      33.4 35.5           39C.4
0  |  ----XXXXXXXXXXXXXXXX|        11.3      14A.3          22B.4     25.4      33.3 35.4 37.4                43B.4     63.5
   |     XXXXXXXXXXXXXXXXX|   7.4       12.3           22A.4               30A.3     35.3           39C.3          44.4 63.4
   |         XXXXXXXXXXXXX|                       16.4 22A.3     23.3                     37.3 39A.3          43B.3
   |            XXXXXXXXXX|4.3 7.3      12.2           22A.222B.3                                        40.3
-1 |          XXXXXXXXXXXX|   7.2                 16.3                                                             44.3 63.3
   |                   XXX|4.2 7.1                 16.2                25.3      33.2                39C.2
   |                   XXX|                  14A.1&2             23.2 25.1&2    33.1 35.2      39A.2          43B.1&2   63.2
   |                     X|                            22A.122B.1&2        30A.2          37.2                     44.2
-2 |                     X|4.1     11.2 12.1      16.1                                    37.1      39C.140.2           63.1
   |                     X|        11.1                          23.1      30A.1     35.1                          44.1
   |                      |                                                                              40.1
   |                     X|                                                                    39A.1
-3 |                      |
   |                     X|
   |                     X|
   |                      |
-4 |                     X|
================================================================================================================================
          Each X represents 4 students, each row is 0.255 logits
```

Figure 2. Wright map for the matter variable.

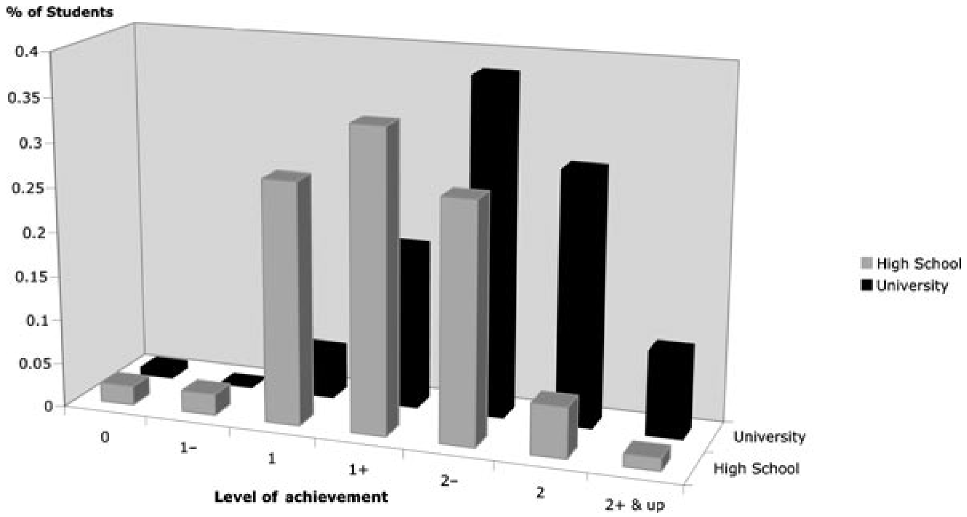

Figure 3. Distribution of high school and university students on the matter variable, using the IRT scaled scores divided into criterion zones according to the framework.

accurately and use chemical symbolism to represent matter. Of course, this does not mean that every response would be at the recognition level. Variations in context and other aspects of the item for a particular student will lead to variation in the level of specific response for any particular question or task. But we would expect recognition-level responses to be common in score vectors, or patterns of scores over a set of items, for students estimated by the statistical models to fall with a proficiency estimate in the recognition category. Recognition is essential before they can begin to relate descriptions of matter with an atomic scale view in Level 3. Our evidence to date shows that students typically enter high school chemistry in the lower areas of Level 1 (notions), and that they are in the upper region of Level 2 and lower region of Level 3 (between recognition and formulation) after 1 year of college-level chemistry in our data sets. Level 4 is sometimes seen at the end of second-year studies in university chemistry. Robust capability to pursue science at Level 5 across topics (generation) is not expected to be reached until, in most cases, graduate school.

Taking Level 1 (notions) as an example, early iterations showed that notions category responses seemed to fall into three general categories, which by preliminary IRT analysis were scaled from low to high within Level 1. The first category of answer observed, labeled as a 1−, is a simple macroscopic observation, usually regarding a piece of data provided in the task or on some component of the item stem. The second type of answer (labeled 1) uses logical patterning and comparative reasoning in addition to observations to generate an answer, but employs no attempts to use domain knowledge of chemistry. The third category (labeled 1+) sought to use chemistry domain knowledge, but employed one of a variety of fundamental misconceptions, thereby skewing the answer in an entirely incorrect direction. When somewhat correct answers begin to appear, these were coded at Level 2 (recognition), as described above, beginning with a score at 2−.

As a specific example of these leveled responses, consider the matter variable in Table 3. A sample question, the scoring guide, and examples of student work in Levels 1 and 2 are shown in Figure 1. In Level 1 (notions), students can articulate their ideas about matter, and use prior experiences, observations, logical reasoning, and knowledge to provide evidence for their ideas. The focus is largely on macroscopic (not particulate) descriptions of matter, since students at this level rarely have particulate models to share. In Level 2 (recognition), students begin to explore the language and specific symbols used by chemists to describe matter. The ways of thinking about and classifying matter are limited to relating one idea to another at a simplistic level of understanding, and include both particulate and macroscopic ideas. In Level 3 (formulation), students are developing a more coherent understanding that matter is made of particles and the arrangements of these particles relate to the properties of matter. Their definitions are accurate, but understanding is not fully developed so that student reasoning often is limited. In their interpretations of new situations, students may overgeneralize as they try to relate multiple ideas and construct formulas. In Level 4 (construction), students are able to reason using normative models of chemistry, and use these models to explain and analyze the phase, composition, and properties of matter. They are using accurate and appropriate chemistry models in their explanations, and understand the assumptions used to construct the models. In Level 5 (generation), students are becoming experts as they gain proficiency in generating new understanding of complex systems through the development of new instruments and new experiments.

RESULTS

With instruments based on assessment questions and tasks such as those shown in Figure 1, we used the framework in Tables 2 and 3 to analyze student data on the matter variable. This study is unusual in that the same assessment was used to map student performance in chemistry across high school and university levels. The same open-ended questions given on posttests at the end of 1 year of instruction to 418 high school students (academic year 2002–2003) were administered to 116 first-year university students at the University of California, Berkeley (June 2002), after they had completed college-level introductory chemistry and were beginning organic chemistry. Both groups of students were assessed on two of the variables described above, matter and change.

Interrater reliability studies were undertaken prior to the administration and scoring in this study. A paired-samples t-test compared the person performance estimates for pairs of raters. Raters were trained and scored on an overlap comparison sample of all open-ended items for 30 cases. Scores correlated strongly ($r = .84$, $p < .001$). However, the mean difference between the two sets of scores—0.16—was statistically significant ($t = 2.265$, $p = .03$). A further analysis was performed to estimate the range of difference in student scores depending on rater harshness. The effect size of interrater comparisons of scoring using a rater model with a single harshness/leniency parameter for each rater was found to be on average ±0.5 on a scale of 1–15, meaning that on average having one rater across all items rather than another would move a student's score about one-half score on the 1–15 Perspectives scale. No students were scored with single raters across all items, and one rater scored all students on a single item whenever this was possible.

The Wright map for the matter variable is shown in Figure 2. A logit, or log odds, scale shows at the far left, ranging from −4 to 4, with the mean of the item difficulty estimates at 0. On the left side, under students, are noted the location of the respondents at posttest, in the shape of an on-the-side histogram. Each X represents four students in the distribution. On the right are columns of items. The columns show the Thurstonian thresholds for each item, with the item number followed by a period and then the threshold number. The kth Thurstonian threshold for each item can be interpreted as the point at which the probability of the scores below k is equal to the probability of the scores k and above, and that probability is .5 (Wilson, 2005). For instance, the threshold of item 39a from a score of 0 to a score of 1 is the lowest threshold, based on this data set, and appears at the bottom of the map. The four students represented by the "X" to the left of 39a.1 on the Wright map have a 50% probability of exceeding a score at or below about −2.75 on the logit scale, as shown at the left of the map. They have a low probability of achieving much higher scores on the item (or on any other item as 39a.1 is the easiest to achieve in the item set). A general tendency across items can be seen for lower thresholds appearing toward the bottom of the map and higher thresholds toward the top. Note that in this case the thresholds to scores of 2+ to 3 will be considered one category throughout this analysis and are fairly mixed at the top of the map, as few items had scores in this range and few students were able to respond at these levels for accurate estimates to be determined among the levels.

Scoring zones, or "criterion" zones, were established using the threshold estimates. A criterion zone is a range of logit values that encompasses a level of the framework (Draney et al., 1996; Wilson & Draney, 2002). Cut scores between levels, called here "composite thresholds," were calculated as the mean of the Thurstonian threshold scores for a given threshold, that is, with all 0–1 thresholds averaged across items to determine the 0 to 1− cutpoint, all 1− to 1 thresholds averaged across items to determine the 1− to 1 cutpoint and so forth. Average standard error of the mean for the composite thresholds was 0.11 logits (0.05 *SD*) for the matter thresholds and 0.13 logits (0.04 *SD*) for the change thresholds.

Student fit across the instrument was also reasonably good, with only 17 of the 538 students, or only about 3.2% of students, with posttest scores outside a standard range.[1] For all these students, the fit was somewhat too random, or in other words, the pattern of correct and incorrect responses was not consistent with the estimated difficulty of the items.

The average standard error for the student proficiency estimates (maximum likelihood estimation) on the posttest was 0.49 logits (0.16 *SD*) on the matter variable and 0.50 logits (0.27 *SD*) on the change variable. The ConQuest software used for the analysis (Wu et al., 1998) provides a selection of methods for calculating person parameters and associated person separation reliabilities, including expected a posteriori estimation based upon plausible values (EAP/PV). EAP/PV reliability was .85 for the matter variable and .83 for the change variable, with person separation reliability estimated at .82 and .80, respectively.

The correlation on student proficiency estimates between the matter and change variables was .68, indicating a moderately strong tendency for the way in which students performed on one variable to be similar to how they performed on the other. However, the correlation was still substantially below 1, indicating that information was gained by measuring students on their understanding in both areas.

Item fit for the partial credit model was reasonably good, with only 2 of the 38 estimated item difficulty parameters, or about 5%, misfitting outside a standard range.[2] Of the two somewhat misfitting items, only one was more random than expected across student proficiencies (the other was slightly overfitting), and that not by much outside the above tolerance range (1.37 mean square weighted fit, weighted fit T of 2.4). All step parameters were within the above tolerance. These item fit statistics were generated in ConQuest (Wu et al., 1998).

This pilot study was intended to be primarily on the matter variable, as the change item bank was in preliminary development at the time. Subsequent work has been done in the change area, and also on the third portion of the Perspectives framework, energy, and that data are currently in analysis. The remainder of this paper discusses only the matter results, pending more information on change and energy.

[1] Of 3/4–4/3 mean square weighted fit (Wu, Adams, & Wilson, 1998), for parameters in which the weighted fit T was greater than 2.

[2] That is, of 3/4–4/3 mean square weighted fit (Wu et al., 1998), for parameters in which the weighted fit T was greater than 2.

Item response model student proficiency estimates are categorized into the five-point Perspectives scale, with cut scores between the categories specified as the mean of the item thresholds shown in Figure 2. High school students averaged 1.75 (0.18 *SD*) on matter and 1.70 (0.18 *SD*) on change. This can be compared to the average score of 2.3 (0.18 *SD*) out of 5 on both matter and change for the college freshman, who had completed college general chemistry as well as usually 1 year of high school chemistry.[3] The distribution of students on the matter variable can be seen in Figure 3.

In Figure 3, this is not a raw score scale, but rather the IRT student performance estimates have been grouped by "criterion zones," or in other words bands of scores on the Wright map that are specified ranges of performance. The band into which the student overall proficiency estimates falls tells us something about their probability of achieving various levels of reasoning on various items, as can be read on the Wright map.

Interpreting these scores qualitatively, after 1 year of high school and 1 year of college chemistry, university students in the sample population scored on the framework in the range of recognition of basic science models of matter, with some students beginning to move to a sound conceptual understanding of multirelational interactions between atoms, and the generation of accurate causal mechanisms. Many of these students still overgeneralize as they try to relate multiple ideas and engage in problem solving, so students at this level would not be expected to effectively construct many models predicting composition, structure, properties, and amounts of matter, but many are beginning to grasp the multiple relationships needed to reason with these models. This is quite interesting because prior to assessment, some instructors described this student population as expected to enter the course *at the beginning of general chemistry* in Level 3 of the Perspectives framework, a level that was actually reached in conceptual understanding only by a subset and not the majority of students *by the end of the general chemistry sequence* of instruction.

Speaking generally, high school students in the sample population were found to be moving from a "notions" conception—based often on their perceptions of real-world knowledge rather than including chemistry concepts—toward simple models of chemistry that begin to describe and explain matter at a particulate level. These students can articulate their ideas of matter, using prior experiences, observations, logical reasoning, and real-world knowledge to provide evidence for their ideas, but much of the evidence they bring to bear remains out of scope, off-topic or "invented" relative to normative models of chemistry. We call these "hybrid models" of reasoning, as they incorporate new learning combined with prior thinking, but do not transition the student to a correct next Perspective in the student's model-based reasoning. Many students are beginning to relate numbers of electrons, protons, and neutrons to elements and mass, and the arrangements and motions of atoms to composition and phase. However, lower achieving students consistently generate answers focusing almost exclusively on macroscopic rather than particulate descriptions of matter, while even for highest achieving high school students ways of thinking and classifying matter are limited to recognizing very basic and simple definitional concepts and representations. Most students are attempting to use chemistry domain knowledge but with little or no correctness.

This idea of limited correctness with domain knowledge is shown in an example in Table 5. Table 5 shows examples of answers to one question for another group of college students, who were assessed at the completion of one-quarter of college-level general

[3] In the United States, high school students rarely complete more than 1 year of chemistry. Some students may complete a year of general science prior to taking chemistry. This introductory science year may include some preparation in physical sciences including an overview of some concepts in both chemistry and physics.

TABLE 5
Examples of Patterns of Reasoning Used in Student Responses Regarding Why N_2H_6 Does Not Exist, a Valence Electron Question

If NH_3 Exists, Why Does Not N_2H_6?	*N*	%	Score
I don't know, "no idea," or nonresponse	13	21.7	0
N_2H_6 can be made, question is wrong	1	1.7	1−
N_2H_6 name is wrong (no explanation for why not)	1	1.7	1−
NH_3 and N_2H_6 have different names	1	1.7	1−
Gases can't be put in a container	2	3.3	1
Nitrogen and hydrogen can't be mixed	2	3.3	1
The container will be too full with more gas	2	3.3	1
NH_3 cannot be broken apart	6	10.0	1
NH_3 can't be "doubled" to make N_2H_6 (no explanation)	1	1.7	1
Not enough nitrogen available to make N_2H_6	5	8.3	1
N and H both have the same charge (+ or −)	2	3.3	1+
Conditions aren't right (acidity or non-aq.)	2	3.3	1+
Nitrogen only forms triple bonds	1	1.7	1+
Conservation of mass—not all particles conserved[a]	10	16.7	1+
N ion has a charge of 3, H ion has a charge of 1	2	3.3	1+
Charges won't balance	1	1.7	2−
Valence elec. explanation, octet rule or Lewis dot described but inaccuracies	5	8.3	2−
Valence elec. explanation, octet rule or Lewis dot fairly correctly	3	5.0	2

[a] These answers appear to be based on confusing this question with a prior question in which carbon was included as one of the reactants.

chemistry. Their range of answers shows most students in Level 1, with some at Level 1+ attempting to use domain knowledge but incorrectly. Level 2 answers, which only about 15% of the students exhibited, begin to show more correct use of domain knowledge for this question.

The ordering of categories of incorrect answers is also interesting to consider. The IRT analysis revealed that students who primarily held a 1+, or attempted chemistry, hybrid model approach or strategy to their answers had a significantly higher probability of answering questions on each instrument correctly than did students who offered answers showing primarily a 1− level reasoning (sound logical reasoning but no use of chemistry models) or a 1 level reasoning (observations only). This suggests the important instructional idea that hybrid models with attempted use of chemistry domain knowledge, *even when applications or interpretations of the chemistry ideas are entirely incorrect*, are valuable advances in student reasoning and should be encouraged in the classroom as students begin to make attempts at using the new language and ideas of chemistry.

However, we found that teachers were often much more likely to "prefer" the Level 1 answers and to give more credit for these answers in their grading of student work. Especially, chemistry high school teachers seemed to be uncomfortable with Level 1+ answers that clearly attempted to use domain knowledge but were not correct. Yet it may be, based on these results, that student attempts to use domain knowledge, perhaps while still fairly unformed in the student's mind, still should be encouraged over answers that do not incorporate attempts at model use at all.

CONCLUSION

These studies show some ways in which a generalizable conceptual framework calibrated with item response modeling can be used to understand student conceptual change in chemistry. Learning within the first Perspective, matter, was shown to be reliably measurable by this approach, with student learning conceived not simply as a matter of acquiring more knowledge and skills, but as progress toward higher levels of competence as new knowledge is linked to existing knowledge, and deeper understandings are developed from and take the place of earlier understandings. This is in contrast to a sometimes much more fragmented view of the discipline in which students often fail to integrate their knowledge, and are unable to "reason like a chemist" when problem solving.

This work connects to the NRC report "Taking Science to School" (NRC, 2007) by capturing one way in which key ideas of increasing complexity in science can be selected and reliably measured. The NRC report "Systems for State Science Assessment" (Wilson & Berenthal, 2005) describes the necessity for coherence in assessment, both horizontally and vertically. Horizontally, instruction and assessment need to be aligned to the same goals and together support student understanding. A criterion-referenced framework mapping both assessment items and student proficiency on the same scale can help with this horizontal alignment. Vertical coherence calls for a shared understanding across the educational system and clear, research-based information on what understanding is needed to for student learning to progress at each stage of the process. The Perspectives framework offers one such view for chemistry. The approach also connects new methodologies to such prior work as the analysis of structures of knowledge in molecular concepts (Ault, Novak, & Gowin, 1984), the tracing of reasoning facets in physics (Minstrell, 2001), and study on learning progressions in matter for younger children (Smith et al., 2006).

In the process of this project, we have learned not only about our students but about our teaching and our discipline. We are learning that it takes substantial time for students to achieve conceptual understanding of chemistry. But we are also learning that most students are able to significantly improve their thinking given the time and opportunity. We are also finding that in the notions level at the beginnings of that understanding, students often need to extensively explore and use the language of chemistry, and that types of "invented" chemistry reasoning are often apparent in their efforts. An example of "invented" chemistry includes student conceptions of solids and liquids. When students at the upper levels of notions, who show a great inclination to use invented chemistry, are told that a particular liquid substance and another solid substance have been measured out in amounts such that they weigh the same, students in notions often want to continue to insist that the solid is heavier. Students employing invented chemistry will often justify their instinctive reasoning with "invented" chemistry ideas drawn from something they have learned in class, such as that if the solid is iron, for instance, the sample should weigh more "because it has a higher mass on the periodic table."

The so-called "invented" ideas are somewhat of a hybrid model between the beginnings of more correct and complete chemistry thinking and the kinds of prior knowledge and real-world experiential reasoning that students bring to instruction. Hybrid models and hybrid reasoning in the notions level, though incorrectly answering the questions posed, do appear to bring value to the development of understanding. Students who reason with these models are significantly more likely to produce some correct answers on other questions and tasks than do students who attempt to employ only logical reasoning and do not introduce even incorrect attempts to incorporate chemistry models or domain knowledge. We believe this is an important finding about how teachers should encourage constant student attempts in chemistry thinking, even if the student attempts reveal considerable confusion. Repeated attempts to successfully piece together correct reasoning seem to be important in refining and improving models, and successfully using domain knowledge.

Students also show the need for opportunities to use the language of chemistry, as they are being introduced to simple beginnings of models. To "speak the language of chemistry," we have found that students need substantial opportunity to explore chemistry words and symbols, before they become fully able to reason in meaningful ways with the symbol systems. For instance, in exploring understanding about metals, students may know from real-world experience that silver, gold, and copper are all metals. But when asked to talk about what properties these three metals share, students may focus directly on the symbolic language rather than what they know about metals, citing for instance that Ag, Au, and Cu all include either the vowel "A" or "U." Discussion and exploration of the symbolic language with others including peers in their courses allow students to delve beneath the symbols and connect to concepts regarding metals, such as ideas of luster and hardness. Students can readily connect such concepts to metals when English-language words such as gold and silver are used rather than the chemical symbols.

Recognizing that the chemical symbolism is a new language is important both for students and for chemistry instructors. Time to work on decoding the symbols and looking for patterns in the symbols gives students time to use and "speak" with the new language.

We also found that there needed to be sufficient allowance in the framework for good reasoning without correct domain knowledge, which was highly evident in the student responses. There is a large body of work in chemistry education that discusses reasoning with and without domain knowledge, as has been previously mentioned, and it was found that a useful framework would need to take this more fully into account. Students were found to show what might be considered the cognitively more advanced ability to explain at every level of the framework, but it was their level of understanding chemistry that affected their explanations. Metz (1995) argues that the limits to the understanding that novice learners exhibit is due to a lack of domain knowledge rather than limits in their general reasoning ability. ChemQuery findings thus far concur. The issue is not that novices cannot reason, but just that they do not reason like chemists, or with the domain knowledge of chemists (Metz, 1995; Samarapungavan & Robinson, 2001). This is especially significant in chemistry where students develop fewer particulate model ideas of understanding from experience and are more likely to rely on instruction.

Our evidence to date shows that students enter high school chemistry in the lower areas of Level 1 (notions), and that they are on average approaching Level 2 (recognition) after a year of general chemistry in high school. On average in our data sets, students were solidly exhibiting Level 2 reasoning following 1 year of university-level general chemistry, with some students able to reason in the lower regions of Level 3 (formulation). Although not presented here, our most recent work is showing that some students at the end of organic chemistry are reaching Level 4 (construction) reasoning. There may be a bimodal distribution appearing in organic chemistry, with some students showing distinctly more conceptual preparedness than others (Scalise, Claesgens, Wilson, & Stacy, 2006a). Level 5 (generation) is not expected to be reached in most cases until graduate school.

It should be noted, however, that these averages do not reflect the substantial variance in models that students are reasoning with across classrooms at any given time, as shown by the distributions in Figure 3. Use of criterion-referenced assessment such as that described in this paper can help us understand the reasoning patterns of individual students and may help suggest appropriate interventions for their needs.

In what we have learned about our teaching as chemistry instructors by using criterion referencing to really explore ideas of how students learn, we are reminded that the way experts communicate information is not the way these experts learned it in the first place.

We know that the ideas and communications of novices are not the same as experts, and there is extensive literature on expert–novice understandings in science education. However, in measuring and tracking paths of development in understanding, we see that for students building toward expertise, ordering, and scaffolding of instruction are important. It is important to revisit new concepts and provide places for students to integrate ideas. Often in spiral instructional designs, this is a common approach, but we find that revisiting is more powerful when students think through the concepts again in a new context. This encourages students to rethink through the process of justification for the approach, which they may not have to do for a "remembered" context that had been presented previously.

In summary, often students are found to hold onto prior beliefs in chemistry and to develop hybrid models of reasoning. Understanding better the process and progress of learning for conceptual change in chemistry, and in other science disciplines, may help us to know what the hybrid models are that students develop, and what is helpful for bridging to new and more powerful understandings of science. This paper shows some ways in which criterion-referenced assessments can help us to think about what students actually know and how to help them learn. Finding ways to more robustly measure trends in developing understanding of the "big ideas" of science, such as with "perspectives" within and across disciplines, could be helpful in other STEM fields as well.

The authors thank Rebecca Krystyniak, Sheryl Mebane, Nathaniel Brown, and Karen Draney for their assistance with instrument and framework development, data collection, and scoring and assistance in discussions of student-learning patterns.

REFERENCES

Ahtee, M., & Varjola, I. (1998). Students' understanding of chemical reaction. International Journal of Science Education, 20(3), 305–316.

American Association for the Advancement of Science. (1993). Benchmarks for science literacy. New York: Oxford University Press.

American Association for the Advancement of Science Project 2061. (1993). Benchmarks for scientific literacy. New York: Oxford University Press.

Ault, C., Novak, J. D., & Gowin, D. B. (1984). Constructing vee maps for clinical interviews on molecule concepts. Science Education, 68(4), 441–462.

Ben-Zvi, R., Eylon, B., & Silberstein, J. (1986). Revision of course materials on the basis of research on conceptual difficulties. Studies in Educational Evaluation, 12, 213–223.

Biggs, J. B., & Collis, K. F. (1982). Evaluating the quality of learning: The solo taxonomy. New York: Academic Press.

Boo, H. K. (1998). Students' understandings of chemical bonds and the energetics of chemical reactions. Journal of Research in Science Teaching, 35(5), 569–581.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change? In S. Carey & R. Gelman (Eds.), The epigenesis of mind. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302.

de Vos, W., & Verdonk, A. H. (1985). A new road to reactions. Journal of Chemical Education, 64(8), 692–694.

diSessa, A. A. (1993). Toward an epistemology of physics. Cognition and Instruction, 10(2–3), 105–225.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? International Journal of Science Education, 20(10), 1135–1191.

Draney, K., Wilson, M., & Pirolli, P. (1996). Measuring learning in lisp: An application of the random coefficients multinomial logit model. Objective Measurement, Theory Into Practice, 3, 195–218.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. Educational Researcher, 23(7), 5–12.

Driver, R., & Scanlon, E. (1989). Conceptual change in science. Journal of Computer Assisted Learning, 5(1), 25–36.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). Making sense of secondary science: research into children's ideas. London: Routledge.

Feynman, R. P. (1963). Atoms in motion. In six easy pieces. Cambridge, MA: Perseus Books.

Furio, C., Azcona, R., & Guisasolo, J. (2002). The learning and teaching of the concepts "amount of substance" and "mole": A review of the literature. Chemistry Education: Research and Practice in Europe, 3(3): 277–292.

Glaser, R. (1984). The role of knowledge. American Psychologist, 39(2), 93–104.

Greenbowe, T. J., & Meltzer, D. E. (2003). Student learning of thermochemical concepts in the context of solution calorimetry. International Journal of Science Education, 25(7), 779–800.

Griffiths, A., & Preston, K. (1992). Grade-12 students' misconceptions relating to fundamental characteristics of atoms and molecules. Journal of Research in Science Teaching, 29(6), 629–632.

Hesse, J., & Anderson, C. W. (1992). Students' conceptions of chemical change. Journal of Research in Science Teaching, 29(3), 277–299.

Johnson, P. (2002). Children's understanding of substances, Part 2: Explaining chemical change. International Journal of Science, 24(10), 1037–1054.

Kennedy, C. A., Wilson, M. R., Draney, K., Tutunciyan, S., & Vorp, R. (2006). GradeMap v4.2 user guide home page, Berkeley Evaluation and Assessment Research Center. Retrieved July 1, 2006, from http://bearcenter.berkeley.edu/GradeMap/.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. Cognitive Psychology, 17, 248–294.

Krajcik, J. S. (1991). Developing students' understanding of chemical concepts. In S. M. Glynn, R. H. Yeany, & B. K. Britton (Eds.), The psychology of learning science: International perspective on the psychological foundations of technology-based learning environments (pp. 117–145). Hillsdale, NJ: Erlbaum.

Krnel, D., Watson, R., & Glazar, S. A. (1998). Survey of research related to the development of the concept of "matter." International Journal of Science Education, 20(3), 257–289.

Kuhn, D. (1989). Children and adults as intuitive scientists. Psychological Review, 96(4), 674–689.

Landauer, T. K. (1991). Let's get real: A position paper on the role of cognitive psychology in the design of humanly useful and usable systems. In J. Caroll (Eds.), Designing instruction (pp. 60–73). Boston: Cambridge University Press.

Larkin, J. H. (1989). What kind of knowledge transfers? In L. B. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 283–306). Hillsdale, NJ: Erlbaum.

Liu, X. (2001). Synthesizing research on students' conceptions in science. International Journal of Science Education, 23(1), 55–81.

Lythcott, J. J. (1990). Problem solving and requisite knowledge of chemistry. Journal of Chemical Education, 67, 248–252.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749.

Metz, K. (1995). Reassessment of developmental constraints on children's science instruction. Review of Educational Research, 65(2), 93–127.

Minstrell, J. (1989). Teaching science for understanding. In L. B. Resnick & L. E. Klopfer (Eds.), Toward the thinking curriculum: Current cognitive research (Chapter 7, pp. 129–149). Yearbook of the Association for Supervision and Curriculum Development.

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

Minstrell, J., Anderson, R., Minstrell, J., & Kraus, P. (in press). Bridging from practice to research and back.

Nakhleh, M. B. (1992). Why some students don't learn chemistry. Journal of Chemical Education, 69(3), 191–196.

National Research Council. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: National Academies Press.

National Research Council/National Academy of Sciences. (1996). National Science Education Standards. Washington, DC: National Academy Press.

Nielsen, J. (1994). Usability engineering. San Francisco: Morgan Kaufmann.

Nielsen, J., & Mack, R. L. (1994). Usability inspection methods. New York: Wiley.

Nurrenbern, S., & Pickering, M. (1987). Concept learning versus problem solving: Is there a difference? Journal of Chemical Education, 64(6), 508–510.

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

Rosenbaum, P. R. (1988). Item bundles. Psychometrika, 53, 349–359.

Samarapungavan, A., & Robinson, W. (2001). Implications of cognitive science research for models of the science learner. Journal of Chemical Education, 78(8), 1107.

Scalise, K. (2001). A mini-survey of conceptual change literature in chemistry. Berkeley, CA: Education 205 Instruction and Development, University of California, Berkeley.

Scalise, K. (2004). BEAR CAT: Toward a theoretical basis for dynamically driven content in computer-mediated environments. Dissertation. Berkeley: University of California.

Scalise, K., Claesgens, J., Krystyniak, R., Mebane, S., Wilson, M., & Stacy, A. (2004). Perspectives of chemists: Tracking conceptual understanding of student learning in chemistry at the secondary and university levels. Paper presented at the Enhancing the Visibility and Credibility of Educational Research, American Educational Research Association Annual Meeting, San Diego, CA.

Scalise, K., Claesgens, J., Wilson, M., & Stacy, A. (2006a). ChemQuery: An assessment system for mapping student progress in learning general chemistry. Paper presented at the NSF Conference for Assessment of Student Achievement, Washington, DC.

Scalise, K., Claesgens, J., Wilson, M., & Stacy, A. (2006b). Contrasting the expectations for student understanding of chemistry with levels achieved: A brief case-study of student nurses. Chemistry Education Research and Practice, The Royal Society of Chemistry, 7(3), 170–184.

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

Stavy, R. (1991). Children's ideas about matter. School Science and Mathematics, 91(6), 240–244.

Teichert, M., & Stacy, A. M. (2002). Promoting understanding of chemical bonding and spontaneity through student explanation and integration of ideas. Journal of Research in Science Teaching, 39(6), 464–496.

Vosniadou, S., & Brewer, W. (1992). Mental models of the Earth: A study of conceptual change in childhood. Cognitive Psychology, 24, 535–585.

Watanabe, M., Nunes, N., Mebane, S., Scalise, K., & Claesgens, J. (2007). "Chemistry for all, instead of chemistry just for the elite:" Lessons learned from detracked chemistry classrooms. Science Education, 91(5), 683–709.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Erlbaum.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. Psychometrika, 60(2), 181–198.

Wilson, M., & Berenthal, M. (2005). Systems for state science assessment. Washington, DC: National Academy.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), Measurement and multi-variate analysis (pp. 325–332). Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12–14, 2000. Tokyo: Springer-Verlag.

Wilson, M., & Scalise, K. (2003). Reporting progress to parents and others: Beyond grades. In J. M. Atkin & J. E. Coffey (Eds.), Everyday assessment in the science classroom (pp. 89–108). Arlington, VA: NSTA Press.

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

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Wu, H.-K., & Shah, P. (2004). Exploring visuospatial thinking in chemistry learning. Science Education, 88, 465–492.

Wu, M., Adams, R. J., & Wilson, M. (1998). The generalised Rasch model. In ACER ConQuest. Hawthorn, Australia: ACER.

Yarroch, W. (1985). Student understanding of chemical equation balancing. Journal of Research in Teaching, 22(5), 449–459.