Cognitive Diagnosis Using Item Response Models

Mark Wilson

University of California, Berkeley, CA, USA

Abstract. In this paper, I will describe a particular approach to cognitive diagnosis that is centered on the idea of developmental assessment, and illustrate how data from this approach can be modeled using explanatory item response models. The developmental assessment approach starts with the idea of a progression of learning embodied in what are called *progress variables*. In a progress variable, student understanding is conceptualized as a continuum with successive levels of development. Effectively, these are seen as a series of student conceptions – this is the first layer of diagnosis. Then, student misconceptions are seen as particular diagnoses within the student conceptions, forming a second layer of diagnosis. Explanatory measurement is introduced as a way to formally model the psychometrics of this situation, using the Berkeley Evaluation and Assessment Research (BEAR) assessment system as a specific example. The discussion is illustrated with examples from student learning about selected topics in science: Earth in the Solar System, and Conceptions of Matter. The paper concludes with a discussion of further steps that match complexities in the diagnostic situation with more complex explanatory models.

Keywords: developmental assessment, progress variables, item response models, BEAR assessment system (BAS)

The conceptualization of cognitive diagnosis presented in this paper is based on a blending of two traditions:

1. The first is an approach to psychometric modeling, which is sometimes called *developmental assessment* (Masters & Forster, 1996), sometimes called *construct modeling* (Wilson, 2005), which centers on the idea of finding a useful (partial) ordering of how students progress through a certain area of understanding.
2. The second is that part of cognitive modeling called *misconception analysis* (Confrey, 1990), which focuses attention on certain typical errors that students tend to make as they learn about particular content areas.

Separately, each of these two approaches have made valuable contributions to cognitive diagnosis (e.g., see Wilson & Carstensen, 2005; Wilson & Scalise, 2006; Wilson & Sloane, 2000; for the former, and Confrey, 1990; Eylon & Linn, 1988; for the latter). The general idea behind the combination of these is that each can offer some help to the other:

1. *Misconception analysis* offers a cognitively based interpretation system to construct modeling, which includes a substantive method to posit diagnoses, gather evidence for the diagnoses, and use them for professional decision-making–in other words, the scores of the psychometric method can be given meaningful psychological interpretations.
2. *Construct modeling* offers the possibility of using the misconception judgments as an explicit and central part of a system for following change through an expected series of developmental levels.

Note that, following the logic of Smith, diSessa, and Roschelle (1993), the misconceptions are seen in this paper in a positive way, as indicators of conceptions, in the sense of defining useful levels of the underlying psychological construct(s). A recent re-examination of the area of misconceptions has resulted in the development of the idea of "threshold concepts," which are posited to be (a) transformative, (b) probably irreversible, (c) integrative, (d) possibly bounded, and (e) troublesome (Meyer & Land, 2003), both in terms of the individual student and in terms of the educators involved, especially the teachers charged with advancing student learning across these threshold concepts.

As an example of this combination of misconception analysis and construct modeling, consider the area of knowledge in elementary science that focuses on the "Earth in the Solar System" (to be expanded on later). Some standard misconceptions that have been identified in this area of knowledge are shown in Figure 1.

While each of these misunderstandings is potentially interesting in terms of diagnosis, it would be a struggle to give a general picture or appreciation of a student's understanding

a) It gets dark at night because the Sun goes around the Earth once a day.
b) All motion in the sky is due to the Earth spinning on its axis.
c) The phases of the Moon are caused by clouds covering the Moon.
d) The Sun goes below the Earth at night.

Figure 1. Four student statements about "Earth in the Solar System."

Zeitschrift für Psychologie / Journal of Psychology 2008; Vol. 216(2):74–88
DOI 10.1027/0044-3409.216.2.74

Level B: Systematic thinking is growing (though still erroneous)

a) It gets dark at night because the Sun goes around the Earth once a day.

b) All motion in the sky is due to the Earth spinning on its axis.

Level A: A (scientifically) unsystematic level of understanding

c) The phases of the Moon are caused by clouds covering the Moon.

d) The Sun goes below the Earth at night.

Figure 2. The same four student statements organized into two levels.

in this area if all that you had was a list of such statements. However, consider if one knew that the bottom two belong to a lower level of understanding, and the top two to the next higher level. This is the sort of information that the construct map associated with this area of knowledge adds to this situation. Explicitly, the bottom two (c and d) represent misconceptions typical of a (scientifically) unsystematic understanding of such phenomena (see Figure 2), while the top two (a and b) represent misconceptions that are from a level where systematic thinking is growing (though still erroneous). This distinction carries the possibility of summarizing the multifarious misconception information as more parsimonious, and still useful, "conception" information. Note that this efficiency carries a cost also–the detail of which misconception was evidenced may or may not be recorded and used in diagnostic interpretation and/or statistical modeling, although it need not be lost if there is a comprehensive form of data recording available.

Following this introduction to the general idea of combining misconception analysis and construct modeling, this paper introduces the application of the BEAR Assessment System for measurement and assessment in cognitive diagnosis, using an example drawn from earth science education, and then gives a second example of the application of this approach in the area of chemistry education. The paper finishes with a discussion of further developments along these lines.

The BEAR Assessment System and the Assessment Triangle

The Berkeley Evaluation and Assessment Research (BEAR) Center has, for the last 15 years, been developing the BEAR Assessment System (BAS). The system consists of four principles, each associated with a practical *building block* (Wilson, 2005), and in addition, a capstone integrative activity that can take on different aspects under different circumstances (e.g., assessment moderation or standard setting). Its original deployment was as a curriculum-embedded system in science (Wilson & Sloane, 2006), but it has clear and logical extensions to other contexts such as in higher education (Wilson & Scalise, 2006), in large-scale assessment (Wilson & Draney, 2004), and across other disciplinary areas, such as mathematics (Wilson & Carstensen, 2007).

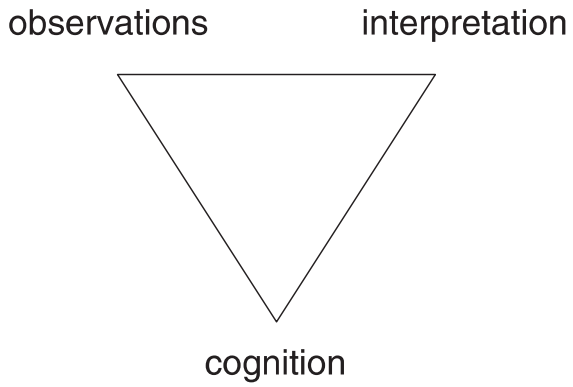


Figure 3. The *Knowing What Students Know* assessment triangle (NRC, 2001).

In this segment, the four principles of the BAS are discussed, and their application to cognitive diagnosis is described using an example based on an assessment in earth science education (which was partly introduced above).

Three broad elements on which every assessment should rest are described by the *assessment triangle* from the *Knowing What Students Know* report (NRC, 2001, p. 296), shown in Figure 3.

According to the Committee Report, an effective assessment design requires: (a) *a model of student cognition and learning in the field of study*; (b) well-designed and tested assessment questions and tasks, often called *items*; and (c) ways to make *inferences about student competence* for the *particular context of use*.

These elements are, of course, inextricably linked in any given application.

The BAS (see Figure 4) is based on the idea that good assessment addresses these considerations through four principles: (a) a developmental perspective (i.e., the cognition vertex), (b) a match between instruction and assessment (i.e., the observations vertex), (c) generating evidence of high-quality assessment, and (d) management by teachers to allow appropriate feedback, feed forward, and follow-up (i.e., together these last two constitute the interpretation vertex).

Principle 1: Developmental Perspective

A developmental perspective regarding student learning means assessing the development of student understanding of particular concepts and skills over time, as opposed to, for instance, making a single measurement at some final or supposedly significant time point. Criteria for developmental perspectives have been challenging goals for educators for many years. What to assess and how to assess it, whether to focus on generalized learning goals or domain-specific knowledge, and the implications of a variety of teaching

and learning theories all impact what approaches might best inform developmental assessment. One issue is that learning situations vary, and their goals and philosophical underpinnings take different forms; hence a "one-size-fits-all" development assessment approach rarely satisfies course needs. Much of the strength of the BAS comes in providing tools to model many different kinds of learning theories and learning domains. What is to be measured and how it is to be valued in each BEAR assessment application is drawn from the expertise and learning theories of the teachers, the curriculum developers, and/or the researchers involved in the developmental process.

Building Block 1: Progress Variables

Progress variables (Masters, Adams, & Wilson 1990; Wilson 1990) embody the first of the four principles: that of a developmental perspective on assessment of student achievement and growth. The four building blocks and their relationship to the assessment triangle are shown in Figure 4. The term *progress variable* is derived from the measurement concept of focusing on one characteristic to be measured at a time. A progress variable is a well-thought-out and researched ordering of qualitatively different levels of performance. Thus, a progress variable defines what is to be measured or assessed in terms general enough to be interpretable across a curriculum but specific enough to guide the development of the other components. When students' conceptions are linked to the progress variable, then it also defines what is to be learned. Progress variables are one model of how assessments can be integrated with diagnosis and instruction.

Assessing the growth of students' understanding of particular concepts and skills requires a model of how student learning develops over a set period of (instructional) time. A developmental growth perspective helps one to move away from "one shot" testing situations, and away from cross sectional approaches to defining student performance, toward an approach that focuses on the process of learning and on an individual's progress through that process.

Progress variables are derived in part from research into the underlying cognitive structure of the domain and in part from professional opinion about what constitutes higher and lower levels of performance or competence, but are also informed by empirical research into how students respond to instruction or perform in practice (NRC, 2001). The progress variable itself is usually expressed through a visual metaphor called a *construct map* (Wilson, 2005); an example is described below. To more clearly understand what a progress variable is, let us consider an example.

The first example explored in this paper is a test of science knowledge, focusing on earth science knowledge in the area of Earth and the Solar System (ESS). The items in this test are distinctive, as they consist of ordered multiple choice (OMC) items, which attempt to make use of the cognitive differences built into the options to make for more valid and reliable measurement (Briggs, Alonzo, Schwab, & Wilson, 2006). Two forms with OMC items were administered as a field test in the spring of 2002. The first form was taken by a sample of 140 students at the end of their 5th, 6th and 7th grade school years. I will refer to these as the Grade 5 sample. The second form was taken by a sample of 156 students at the end of their 8th- and 9th-grade school years. I will refer to these as the Grade 8 sample. The samples were taken from the classrooms of 10 middle school teachers from California who had agreed to participate in a research project. The Grade 5 and Grade 8 student samples responded to 18 and 23 OMC items, respectively. For each sample, eight of the OMC items were based on the ESS construct map, and these are the focus of the analysis that follows.

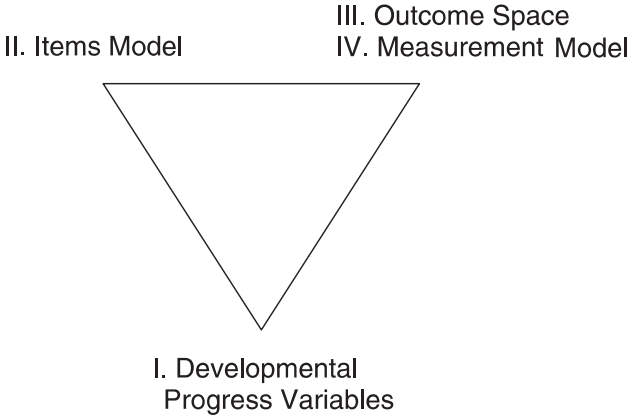


Figure 4. The building blocks of the BEAR Assessment System.

The standards and benchmarks for ESS appear in Appendix A of the Briggs et al. article (2006). According to these standards and the underlying research literature, by the 8th grade, students are expected to understand three different phenomena within the ESS domain: (a) the day/night cycle, (b) the phases of the moon, and (c) the seasons – in terms of the motion of objects in the solar system.

A complete scientific understanding of these three phenomena is the top level of our construct map. In order to define the lower levels of our construct map, the literature on student misconceptions with respect to ESS was reviewed by Briggs et al. Documented explanations of student misconceptions with respect to the day/night cycle, the phases of the moon, and the seasons are displayed in Appendix A of the Briggs et al. article (2006).

The goal was to create a single continuum that could be used to describe typical students' understanding of three phenomena within the ESS domain. In contrast, much of the existing literature documents students' understandings about a particular ESS phenomenon without connecting each understanding to their understandings about other related ESS phenomena (e.g., Baxter, 1995; Sadler, 1987, 1998; Schneps & Sadler, 1988; Stahly, Krockover, & Shep-

ardson, 1999; Targan, 1987; Trumper, 2001). Often, this research is discussed in the context of the debate about student misconceptions. Several researchers have investigated the consistency of student understanding in the ESS domain (Kikas, 1998; Klein, 1982; Roald & Mikalsen, 2001; Vosniadou & Brewer, 1994). Vosniadou & Brewer (1994) carried out the most elaborate of these investigations. They looked at the consistency of students' answers to a series of questions about the day/night cycle and found that the majority of students in their study consistently used just one model to explain their answers. Although this work does not link day/night to the other two phenomena included in the ESS construct map, it suggested that students may well employ the same model to explain multiple phenomena within the domain of ESS.

Several of these studies record the percentage of students in different age groups who held a certain conception about a particular phenomenon (Baxter, 1995; Kikas, 1998; Roald & Mikalsen, 2001; Sadler, 1998; Vosniadou, 1991). However, these researchers did not explore how students move from one level of conceptual understanding to another. Other researchers have proposed developmental progressions of students' ideas but the descriptions of the levels are very general. For example, Vosniadou & Brewer (1994) place students' ideas into three categories–initial, synthetic, and scientific. In their approach, there is only one step between students' naïve ideas, such as night being caused by clouds covering the sun, and complete scientific understanding. Unfortunately, this middle category describes the understanding of most students with respect to ESS phenomena (Roald & Mikalsen, 2001), so it is this step that must be more fully laid out to help teachers. Baxter (1995) describes four *conceptual phases* that are slightly more elaborated: Her phase one is analogous to the "initial level" of Vosniadou and Brewer. In phase two, students recognize that astral bodies move to cause observed phenomena, but characterize this movement as up/down or right/left. In phase three, the movement is characterized by an earth-centered orbit. Finally, in phase four, students have a heliocentric view. However, even students with a heliocentric view of the motion of objects in the solar system may not fully understand more complicated phenomena, such as the phases of the moon and the seasons.

By examining student conceptions across the three phenomena and building on the progressions described by Vosniadou and Brewer (1994) and Baxter (1995), Briggs et al. (2006) initially established a general outline of the construct map for student understanding of ESS. This general description helped them impose at least a partial order on the variety of student ideas represented in the literature. However, the levels were not fully defined until typical student thinking at each level could be specified. This typical student understanding is represented in the ESS construct map, shown in Figure 5, as "common errors." Common errors used to define Level 1 include explanations for day/night and the phases of the moon involving something covering the sun or moon, respectively.

Level	Description
5 8th grade	Student is able to put the motions of the Earth and Moon into a complete description of motion in the Solar System which explains: – the day/night cycle – the phases of the Moon (including the illumination of the Moon by the Sun) – the seasons
4 5th grade	Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that – the Earth is both orbiting the Sun and rotating on its axis – the Earth orbits the Sun once per year – the Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the sky – the Moon orbits the Earth once every 28 days, producing the phases of the Moon COMMON ERROR: Seasons are caused by the changing distance between the Earth and Sun. COMMON ERROR: The phases of the Moon are caused by a shadow of the planets, the Sun, or the Earth falling on the Moon.
3	Student knows that: – the Earth orbits the Sun – the Moon orbits the Earth – the Earth rotates on its axis However, student has not put this knowledge together with an understanding of apparent motion to form explanations and may not recognize that the Earth is both rotating and orbiting simultaneously. COMMON ERROR: It gets dark at night because the Earth goes around the Sun once a day.
2	Student recognizes that: – the Sun appears to move across the sky every day – the observable shape of the Moon changes every 28 days Student may believe that the Sun moves around the Earth. COMMON ERROR: All motion in the sky is due to the Earth spinning on its axis. COMMON ERROR: The Sun travels around the Earth. COMMON ERROR: It gets dark at night because the Sun goes around the Earth once a day. COMMON ERROR: The Earth is the center of the universe.
1	Student does not recognize the systematic nature of the appearance of objects in the sky. Students may not recognize that the Earth is spherical. COMMON ERROR: It gets dark at night because something (e.g., clouds, the atmosphere, "darkness") covers the Sun. COMMON ERROR: The phases of the Moon are caused by clouds covering the Moon. COMMON ERROR: The Sun goes below the Earth at night.
0	No evidence or off-track

Figure 5. Construct map for student understanding of "Earth in the Solar System."

In addition to defining student understanding at each level of the continuum, the notion of common errors helps to clarify the difference between levels. Misconceptions, represented as common errors in one level, are resolved in the next level of the construct map. For example, students at Level 3 think that it gets dark at night because the earth

goes around the sun once a day–a common error for Level 3–while students at Level 4 no longer believe that the earth orbits the sun daily but rather understand that this occurs on an annual basis.

The top level of the ESS construct map represents the understanding expected of 8th graders in national standards documents. Because students' understanding of ESS develops throughout their schooling, it was important that the same continuum be used to describe the understandings of both 5th- and 8th-grade students. However, the top level is not expected of 5th graders; equally, we do not expect many 8th-grade students to fall into the lowest levels of the continuum.

Principle 2: Match Between Instruction and Assessment

The match between instruction and assessment in the BAS is established and maintained through two major parts of the system: progress variables (as expressed by construct maps), which have been described in the section above, and specific types of assessment items or activities, which are described in this section. The main motivation for the progress variables so far developed is that they serve as a framework for the assessments and a method of making measurement possible, and diagnoses interpretable. However, this second principle makes clear that the framework for the measurement and the framework for the cognitive diagnoses must be one and the same. This is not to imply that the needs of measurement must drive the diagnosis, nor that the diagnosis will entirely determine the assessment, but rather that the two, measurement and diagnosis, must be in step–they must both be designed to accomplish the same thing, i.e., student learning.

Using construct maps to structure both measurement and diagnosis is one way to make sure that the two are in alignment, at least at the planning level. In order to make this alignment concrete, however, the match must also exist at the level of classroom interaction and that is where the nature of the assessment items becomes crucial. Of course, from the construct modeling point of view, assessment items should be designed principally to prompt students to supply responses that are evidence for different levels of the construct map. However, in addition, assessment items need to reflect the range and styles of the instructional and diagnostic practices in the classroom. They must have a place in the "rhythm" and "tone" of the classroom. The responses that they generate from students need to relate not only to the levels of the construct map, but also to the misconceptions within those levels.

Building Block 2: The Items Design

The items design governs the match between the measurement and the various cognitive diagnoses. The critical element to ensure this in the BAS is that each assessment item is designed to generate diagnostic student responses for at least one level of the construct map, preferably more.

A variety of different item types may be used in an assessment system, based on the requirements of the particular situation. There is a common tension in assessment situations between the use of multiple-choice items, which are perceived to contribute to more reliable assessment, and other, alternative forms of assessment, which are perceived to contribute to the validity of a testing situation. The BAS includes designs that allow both kinds of assessment items, so that each can be deployed to its strengths.

When using this assessment system within a curriculum, a particularly effective mode of assessment is what we call *embedded assessment*. By this we mean that opportunities to assess student progress and monitor their understandings are integrated into the instructional materials and are virtually indistinguishable from day-to-day classroom activities. We found it useful to think of the metaphor of a stream of instructional activity and student learning, with the teacher dipping into the stream of learning from time to time to evaluate student progress and performance. In this model or metaphor, assessment then becomes *part* of the teaching and learning process, and we can think of it being assessment *for* learning (AfL: Black, Harrison, Lee, Marshall, & Wiliam, 2003). If assessment is also a learning event, then it does not take unnecessary time away from instruction, and the number of assessment items can be more efficiently increased in order to improve the reliability of the results (Linn & Baker, 1996). In embedded assessment in classrooms, there will be a variety of different types of assessment items: These may include individual and group "challenges," data processing questions, questions following student readings, and even instruction/assessment events such as "town meetings." Such items may be open-ended, requiring students to fully explain their responses in order to achieve a high score, or they may be multiple choice, freeing teachers from having to laboriously hand-score all of the student work (Briggs et al., 2006).

There are many variations in the way that progress variables can be realized in practice, from using different assessment modes (multiple choice, performance assessment, mixed modes, etc.), to variations in the frequency of assessing students (once a week, once a month, etc.), to variations in the use of embedding of assessments (all assessments embedded, some assessments in a more traditional testing format, etc.). One traditional format for assessments is the end-of-unit test that is used for a variety of traditional classroom purposes. These are easily blended into the BAS approach, and are also important for certain technical purposes such as equating different parts of the assessment system. Typically we refer to these as *link items* in acknowledgment of their role in equating. A second type of assessment practice where the BAS can play an important role is in teachers' classroom verbal interactions (i.e., these range from classroom discussion to teacher discussions with groups of students and/or with individual stu-

dents). In these contexts, the BAS information can play an important qualitative role in supporting teachers' successful interactions to promote student learning.

Returning to the ESS example, the OMC items were written as a function of the underlying construct map, which is central to both the design and interpretation of the OMC items. Item prompts were determined by both the domain as defined in the construct map and canonical questions (i.e., those which are cited in standards documents and commonly used in research and assessment contexts). The ESS construct map focuses on students' understanding of the motion of objects in the solar system and explanations for observable phenomena (e.g., the day/night cycle, the phases of the moon, and the seasons) in terms of this motion. Therefore, the ESS OMC item prompts focused on students' understanding of the motion of objects in the solar system and the associated observable phenomena. Distractors were written to represent: (a) different levels of the construct map, based upon the description of both understandings and common errors expected of a student at a given level and (b) student responses that were observed from an open-ended version of the item.

Two sample OMC items, showing the correspondence between response options and levels of the construct map, are shown in Figure 6. Each item-response option is linked to a specific level of the construct map. Thus, instead of gathering information solely related to student understanding of the specific context described in the question, OMC items allow us to link student answers to the larger ESS domain represented in the construct map. Taken together, a student's responses to a set of OMC items permit an estimate of the student's level of understanding, as well as providing diagnostic information about that specific misconception.

The ESS construct map and OMC items were revised extensively and repeatedly by the research team, which included those with expertise in both science content and item design. Further revision occurred after the team received feedback from regional directors of a statewide science reform program and developers of a nationally recognized elementary school science curriculum. A pilot test of the OMC items was conducted in the fall of 2001. For each OMC item, a paired open-ended version was created. Both the OMC items and their open-ended counterparts were included in the pilot. A total of 138 5th graders and 112 8th graders participated in the pilot. The vast majority (93%) of student responses to the open-ended items could be readily scored using the ESS progress variable. In addition, student responses to the open-ended versions of the items were used to refine the answer choices to better reflect students' wording of the ideas expressed in the construct map.

Item appropriate for fifth graders:	
Which is the best explanation for why it gets dark at night?	
A. The Moon blocks the Sun at night.	Level 1
B. The Earth rotates on its axis once a day.	Level 4
C. The Sun moves around the Earth once a day.	Level 2
D. The Earth moves around the Sun once a day.	Level 3
E. The Sun and Moon switch places to create night.	Level 2
© WestEd, 2002	
Item appropriate for eight graders:	
Which is the best explanation for why we experience different seasons (winter, summer, etc.) on Earth?	
A. The Earth's orbit around the Sun makes us closer to the Sun in summer and farther away in winter.	Level 4
B. The Earth's orbit around the Sun makes us face the Sun in the summer and away from the Sun in the winter.	Level 3
C. The Earth's tilt causes the Sun to shine more directly in summer than in winter.	Level 5
D. The Earth's tilt makes us closer to the Sun in summer than in winter.	Level 4
© WestEd, 2002	

Figure 6. Sample OMC items based on "Earth in the Solar System" construct map.

Principle 3: Management by Teachers

In order for information from the assessment items and the BEAR analysis to be useful to instructors and students, it must be couched in terms that are directly related to the teachers' instructional goals that are behind the progress variables. Open-ended items, if used, must be quickly, readily, and reliably scorable. The categories into which they are scored must be readily interpreted in an educational setting, whether it is within a classroom, by a parent, or for policy analysis. The requirement for transparency in how item outcomes relate to how students actually respond to an item leads to the third building block.

Building Block 3: The Outcome Space

The outcome space is the set of categorical outcomes into which student performances are categorized for all the items associated with a particular progress variable. In practice, the outcome space is presented as *scoring guides* for student responses to assessment items. The scoring guide is divided into a series of increasingly sophisticated levels that student responses are categorized into. (Note that sometimes some of the categories may be at the same level–this is called an ordered partition.) The outcome space is the primary means by which the essential element of teacher professional judgment is implemented in the BAS. These are supplemented by "exemplars," which are examples of student work at each level for every item, and "blueprints," which provide the teachers with a layout showing opportune times in the curriculum to assess the students on the different progress variables. Note that it is possible to confuse the outcome space with the construct map. The fundamental distinction lies in the generality of the construct map compared to the specificity of the scoring guides. As there can be multiple types of assessments

related to a single construct map, the construct map is necessarily defined at a more general level than the scoring guides.

In the case of OMC-type items, the outcome space is simply the set of levels of the construct map that are used to generate the distractors for the items. The distractors must be chosen not only to cover a reasonable range of the construct map, but should also be attractive and meaningful to students of the appropriate age. For example, Figure 6 indicates the relevant level of the ESS construct for each of the distractors for the two items. As can be seen, these do not necessarily cover all levels of the construct map–they will match up to the levels that most readily correspond to the topic of the question. Nor are they necessarily in a one-to-one relationship with the levels of the construct map.

Principle 4: Evidence of High-Quality Assessment

Technical issues of reliability and validity, fairness, consistency, and bias can quickly sink any attempt to measure along a progress variable as described above, or even to develop a reasonable framework that can be supported by evidence. To ensure comparability of results across time and context, procedures are needed to (a) examine the coherence of information gathered using different formats, (b) map student performance onto the progress variables, (c) describe the structural elements of the accountability system – items and raters – in terms of the progress variables, and (d) establish uniform levels of system functioning in terms of quality control indices such as validity and reliability.

While this type of discussion can become very technical to consider, it is sufficient to keep in mind that the traditional elements of assessment standardization, such as validity and reliability studies and bias and equity studies, must be carried out to satisfy quality control and ensure that evidence can be relied upon. The core validity tool used in the BAS is the Wright map, which is an empirically-derived version of the construct map.

Building Block 4: Wright Maps

Wright maps are graphical and empirical representations of a progress variable, showing how it unfolds or evolves in terms of increasingly sophisticated student responses/performances (note that they are named after Ben Wright of the University of Chicago, who pioneered their use in measurement). The locations of items and students on the Wright map are derived from empirical analyses of student data on sets of assessment items. Wright maps are based on an ordering of these assessment items from relatively easy items to more difficult and complex ones. A key feature of these maps is that both students and items can be located on the same scale, giving student proficiency the possibility of substantive interpretation, in terms of what the student knows and can do and where the student is having difficulty. Once the scale of the Wright map has been established by estimation (also called "calibration"), the maps can be used in the classroom to interpret the pattern of achievement of groups of students, the progress of one particular student, or even the differential success of an individual on specific items. Wright maps can also be very useful in large-scale assessments, providing information that is not readily available through numerical score averages and other traditional summary information–they are used extensively, for example, in reporting on the PISA assessments (e.g., PISA, 2005).

Before developing a Wright map, we have to clear up one potential problem: In the usual situation in item response theory (IRT), each category of response has its own score. However, in the OMC case, it is quite possible for

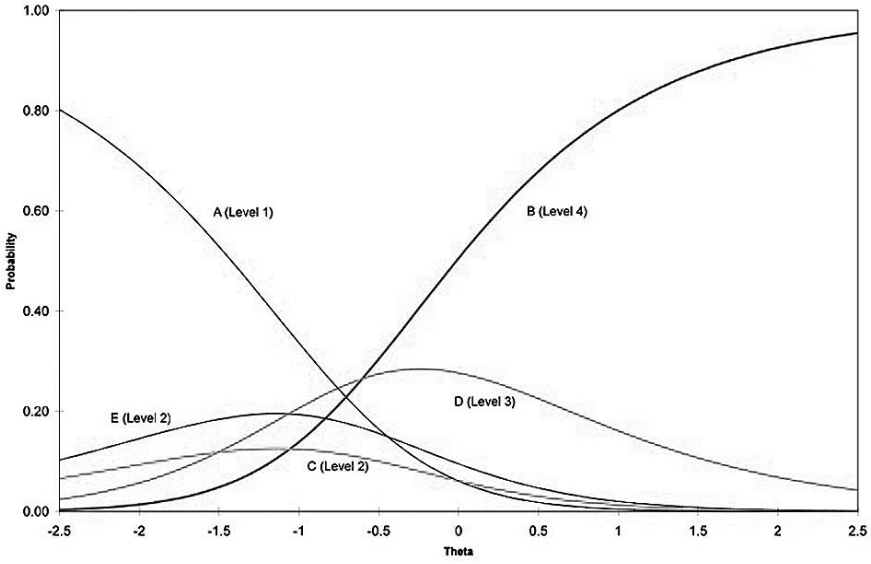


Figure 7. OPM-based plot of IOCCs for grade 5, ESS Item 1.

different responses to share the same score. Fortunately, this situation has already been investigated, and a solution is available, the Ordered Partition Model (OPM; Wilson 1992). One advantage of using the OPM is that it becomes possible to plot a curve for *each* response possibility in an OMC item. We call this an item options characteristic curve (IOCC). For an OMC item with five response options, the IOCC plots five different curves, as in Figure 7.

To express the OPM more properly, let X_{ni} denote the response of student n, with ability θ_n, to item i. Each item has K_i possible response categories ($k = 1, \ldots, K_i$). The probability of obtaining a response in category k can be written as

$$P(X_{ni} = k \mid \theta_n) = \frac{\exp[\theta_n B_i(k) - \delta_{ik}]}{\sum_{h=1}^{K_i} \exp[\theta_n B_i(h) - \delta_{ih}]} \qquad (1)$$

In Equation 1, d_{ik} is a difficulty parameter for category k in item i, where (by convention) $d_{i0} \equiv 0$, and $B_i(k)$ is a known scoring function that maps category k to score level m. (Note that if $B_i(k) = k = m$, the OPM reduces to the partial credit model.) Estimation of the item parameters d_{ik} is carried out using marginal maximum likelihood, with person parameters (θ_n) estimated in a second stage using empirical Bayes posterior distributions. For details, see Wilson and Adams (1993). The OPM can be estimated using the item response modeling software ConQuest (Wu, Adams, & Wilson, 1998).

To demonstrate the use of the OPM to investigate OMC items, consider Item 1 from Figure 6. There are five response alternatives to this item, so options A through E represent five categories, hence $K_i = 5$. Each of the categories is mapped to a hypothesized level on the ESS construct map, and this can be used to score responses in each category with the scoring function $B_1(k) = m$[1].The partial ordering (rather than complete ordering) of the options is because there is more than one way to choose a response option associated with Level 2 for this item. Once this scoring function has been utilized, the ordering of the options relative to the ESS construct map is A < (C = E) < D < B. If answer options C and E were collapsed into a single category, the OPM would be equivalent to the partial credit model (Masters, 1982), but this would mean that some of the information about the original response options would no longer be available for diagnostic use. To the extent that these answer options may tell different stories about student misconceptions, it is preferable to use the model that preserves as much of the original detail of the student response as possible.

A schematic Wright map illustrating the interpretation of item and person estimates is shown in Figure 8. For simplicity, suppose that the item has been scored dichotomously as "0" or "1" (i.e., "wrong"/"right"), that is $X_{ni} = 0$ or 1. The logic of the Wright map representation is that the student has a certain amount of the construct, indicated by θ, and that an item also has a certain amount, indicated by δ_i

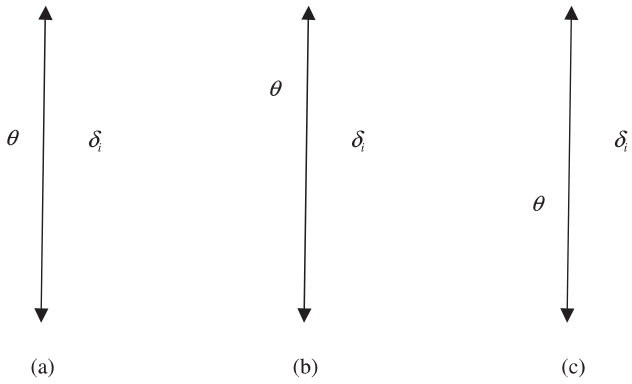


Figure 8. Representation of the relationships between respondent location (θ) and the location of an item (δ_i.) on a schematic Wright map.

(where only one subscript is needed, as there is only one parameter per item in a dichotomous context). However, the values work in opposite directions – hence the *difference* between them is what counts. We can consider three situations corresponding to the three panels of Figure 8:

1. When those amounts are the same, the probability of the response "1" is 0.5 (and, hence, the probability of "0" is the same, 0.5–see Figure 8, Panel a);
2. When the student has more ability than the item has difficulty (i.e., $\theta > \delta_i$,) the probability of a "1" is greater than 0.5 (see Figure 8, Panel b); and
3. When the item has more difficulty than the respondent has ability (i.e., $\theta < \delta_i$,), then the probability of a "1" is less than 0.5 (see Figure 8, Panel c).

The statistical basis for the measurement models used here is a class of models called explanatory item response models (EIRMs; De Boeck & Wilson, 2004), which lays out a large class of models that are suitable for estimating parameters of the psychological space. These are based on generalized linear mixed models and nonlinear mixed models (GLMMs and NLMMs respectively; Verbeke & Molenberghs, 2000), and are estimable with generalized software such as SAS NLMIXED (SAS Institute, 1999) and GLLAMM (Rabe-Hesketh, Skrondal, & Pickles, 2004). In this paper, the set of models to be estimated will be kept rather restricted, for pedagogic purposes, so more restricted (and, hence, more efficient) software can be used (ConQuest; Wu et al., 1988). We typically use a multi-dimensional Rasch modeling approach to calibrate the maps for use in the BAS (see Adams, Wilson, & Wang, 1997, for the specifics of this model).

These Wright maps have at least two advantages over the classical ways used in education to record and report student performance as a total score or percentage correct: First, the Wright map encourages teachers to interpret a

[1] That is: B_1(A) = 1, B_1(B) = 4, B_1(C) = 2, B_1(D) = 3 and B_1(E) = 2.

student's proficiency in terms of average or typical performance on assessment items near that student on the Wright map (it shares this feature with other approaches based on Rasch models); and second, it takes into consideration the relative difficulties of the items involved in assessing student proficiency (it shares this feature with other item response theory approaches).

Once calibrated, maps can be used to record and track student progress and to pick out the skills that students have mastered and those that they are currently working on. By placing students' performance on the continuum illustrated by the Wright map, teachers, administrators and the public can interpret student progress with respect to the standards that are inherent in the progress variables. Wright maps can come in many forms, and have many uses in classroom and other educational contexts. In order to make the maps flexible and convenient enough for use by teachers and administrators, we have also developed software for teachers to generate the maps. This software, which we call ConstructMap (Kennedy, Wilson, & Draney, 2008), allows teachers to enter the scores given to the students on assessments, and then map the performance of groups of students, either at a particular time or over a period of time.

Example 2: The Iota Model

We will illustrate the idea of cognitive diagnosis in the context of a progress variable by using a recent example from educational assessment–an assessment system built for a high school chemistry curriculum, *Living by Chemistry: Inquiry-Based Modules for High School* (Claesgens, Scalise, Draney, Wilson, & Stacey, 2002). The Living by Chemistry (LBC) project at the Lawrence Hall of Science was awarded a grant from the National Science Foundation in 1999 to create a year-long course based on real-world contexts that would be familiar and interesting to students. The goal is to make chemistry accessible to a larger and more diverse pool of students while improving preparation of students who traditionally take chemistry as a prerequisite for scientific study. The focus is on the domain knowledge the students have acquired during instructional interactions in terms of how they are able to think and reason with chemistry concepts.

The set of constructs on which both the LBC curriculum and its assessment system (an application of the BAS; Wilson & Sloane, 2000) are built is called "Perspectives of Chemists." Three progress variables, or strands, have been designed to describe chemistry views regarding three "big ideas" in the discipline: matter, change, and energy. The *matter* strand is concerned with describing atomic and molecular views of matter. *Change* involves kinetic views of change and the conservation of matter during chemical change. *Energy* considers the network of relationships involved with conservation of energy. The levels of the matter progress variable are shown in Figure 9. It describes how a student's view of matter progresses from a continuous, macro view, to a particulate view accounting for the existence of atoms and molecules, and then builds in sophistication.

5. Generation: Students use the models to generate new knowledge and to extend models. (~graduate school)

4. Construction: Students integrate scientific understanding into full working models of the domain. (~upper division)

3. Formulation: Students combine unirelational ideas, building more complex knowledge structures in the domain. (~lower division)

2. Recognition: Students begin to recognize normative scientific ideas, attaching meaning to unirelational concepts. (~high school)

1. Notions: Students bring real-world ideas, observation, logic and reasoning to explore scientific problem-solving. (~middle-school)

Figure 9. The levels of the LBC progress variables.

Assessments carried out in studies of this progress variable show that a student's atomic views of matter begin with having no atomic view at all, but simply the ability to describe some characteristics of matter. For example, this could include (a) differentiating between a gas and a solid on the basis of real-world knowledge of boiling solutions such as might be encountered in food preparation, or (b) bringing every-day logic and patterning skills to bear on a question of why a salt dissolves.

This is the lowest level of the matter variable. At this novice level of sophistication, students do not employ an accurate molecular model of chemistry. However, within this level, a progression in sophistication can be seen from (a) those unable or unwilling to make any relevant observation at all during an assessment item on matter, to (b) those who can make an observation and then follow it with logical reasoning, to (c) those who can extend this reasoning in an attempt to employ actual chemistry knowledge, although this will typically be done incorrectly in first attempts. All these behaviors fall into Level 1, called the "notions" level (see Figure 9), and are assigned incremental 1– and 1+ scores, which for simplicity of presentation are not shown in this version of the framework.

When students begin to make the transition to accurately using simple molecular chemistry concepts, Level 2 begins – this is called the "recognition" level. At Level 2 of the matter progress variable, we see students using very one-dimensional models of chemistry: a simple representation or a single definition will be used broadly to account for and interpret chemical phenomena. At this level students rarely combine these ideas together to form more complex scientific ideas. They do, however, begin extending experience and logical reasoning to include accurate chemistry-specific domain knowledge. In the conceptual framework, this is when students begin to employ definitions, terms, and principles with which they will later reason and negotiate meaning. At this level, students are concerned with learning the language and representations of the domain of chemistry and are introduced to the ontological categories and epistemological beliefs that fall within the domain of chemistry. Students will

tend to focus on a single aspect of correct information in their explanations but may not have developed more complete explanatory models to relate to the terms and language.

When students do begin to combine and relate patterns to account for, for example, the contribution of valence electrons and molecular geometry to the process of dissolving, they are considered to have moved to Level 3, "formulation." Coordinating and relating developing knowledge in chemistry becomes critical to move into this level. Niaz and Lawson (1985) argue that without generalizable models of understanding, students choose to memorize rules instead, limiting their understanding to the earlier recognition level of the perspectives. Students need a base of domain knowledge before integration and coordination of the knowledge develops into understanding (Metz, 1995). So as they move toward the formulation level, students should be developing a foundation of domain knowledge so that they can begin to reason like chemists by relating terms to conceptual models of understanding in chemistry, rather than simply memorizing algorithms and terms.

The LBC "matter" strand is an example of a relatively mature construct, although as yet untested at its upper end: the levels that cover college and graduate study[2]. When a construct map is first postulated, it will often be much less well-formed than this. The construct map will be refined through several processes as the instrument is being developed. These processes include (a) explaining the construct to others with the help of the construct map, (b) creating items that you believe will lead respondents to give responses that inform levels of the construct map, (c) trying out those items with a sample of respondents, and (d) analyzing the resulting data to check if the results are consistent with your intentions, as these intentions are expressed through the construct map.

An example of an LBC assessment prompt and actual student answers at Levels 1 and 2 is shown in Figure 10, along with interpretation. To match instruction and assessment, this LBC assessment question followed a laboratory project in which students explored chemicals that had different smells. Note that this example item is a partial-credit item, and spans multiple levels of the construct map with the awarding of varying degrees of credit. BEAR assessment items can take many formats and can be designed to span multiple levels (polytomous) or can act as "quick check" items that measure at one cut score (dichotomous), depending on the desires of the course instructors and developers.

The remaining levels of the framework, construction and generation, represent further extensions and refinements and are not expected to be mastered at the high school or introductory undergraduate levels.

A scoring guide, showing the outcome space for Levels 1 and 2 for LBC is shown in Figure 11. Much greater detail is shown in the exemplars in Figure 12. At this level of detail, the diagnostic value of the levels has been made clearer, just as the grain size of the levels has become finer, and the illustrations have become more concrete.

Question:

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.

Student Answers at Level 1 of Visualizing Matter Progress Variable

Response: "I think there could be a lot of different reasons as to why the two solutions smell differently. One could be that they're different ages, and one has gone bad or is older which changed the smell. Another reason could be that one is cold and one is hot."

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.

Student Answers at Level 2 of Visualizing Matter Progress Variable

Response: "They smell differently b/c even though they have the same molecular formula, they have different structural formulas with different arrangements and patterns."

Response: "Butyric acid smell bad. It's an acid and even though they have the same molecular formula but they structure differently."

Analysis: Both responses appropriately cite the principle that molecules with the same formula can have different structures, or arrangements of atoms within the structure described by the formula. However the first answer shows no attempt and the second answer shows an incomplete attempt to use such principles to describe the simple molecules given in the problem setup, which would have advanced response to the next level.

Figure 10. Example of an LBC assessment prompt and actual student answers at Levels 1 and 2.

A Wright map illustrating the estimates for the LBC matter progress-variable model is shown in Figure 13. The units of the map are called *logits*, or the log of the odds, and are shown along the far left-hand side. On this map, an X on the left hand side represents a single student (note that the actual number of students who were used for calibration of the items was larger–the sample shown is from a single data collection). The symbols on the right-hand side, such as 3.2– represent the locations of the thresholds for the items, in this case, it is the 2– threshold for Item 3–that is,

[2] Those interested in the upper levels should contact the LBC project at http://www.cchem.berkeley.edu/amsgrp/ed_pages/eduindex.html

0. Irrelevant or no response.
Response contains no information relevant to item.

Notions: Describe the properties of matter
The student relies on macroscopic observation and logic skills rather than employing an atomic model. Students use common sense and experience to express their initial ideas without employing correct chemistry concepts.

1– Makes one or more macroscopic observation and/or lists chemical terms without meaning.

1 Uses macroscopic observations/descriptions and restatement AND comparative/logic skills to generate classification, BUT shows no indication of employing chemistry concepts.

1+ Makes accurate simple macroscopic observations (often employing chemical jargon) and presents supporting examples and/or perceived rules of chemistry to logically explain observations, BUT chemical principles/definitions/rules cited incorrectly.

Recognition: Represent changes in matter with chemical symbols
The students are "learning" the definitions of chemistry to begin to describe, label, and represent matter in terms of its chemical composition. The students are beginning to use the correct chemical symbols (i.e., chemical formulas, atomic model) and terminology (i.e., dissolving, chemical change vs. physical change, solid liquid gas).

2– Cites definitions/rules/principles pertaining to matter somewhat correctly.

2 Correctly cites definitions/rules/principles pertaining to chemical composition.

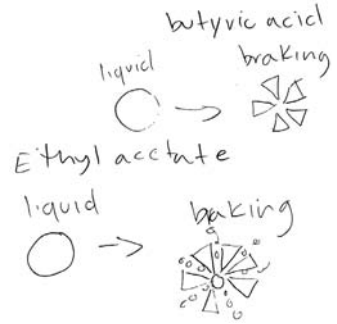
2+ Cites and appropriately uses definitions/rules/principles pertaining to the chemical composition of matter and its transformations.

Figure 11. LBC scoring guide showing the outcome space for Notions and Recognition levels.

Notions 1– *Response:* If they have the same formula, how can they be different?
Analysis: Student makes one macroscopic observation by noting that the molecular formulas in the problem setup are the same.

1 *Response:* I think there could be a lot of different reasons as to why the two solutions smell differently. One could be that they're different ages, and one has gone bad or is older

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:)



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.

Figure 12. Exemplars of Level 1 responses to the LBC item in Figure 7.

the location where we would predict that students would get Item 3 correct, at the 2– level or below, 50% of the time. The right-hand side shows the first five items in separate columns, and then the remainder of the items in a single column. The bands across the map show approximate segments where certain levels of response (i.e., 1+, 2–, etc.) predominate. Some items do not conform to this pattern of banding, and these are noted in the Figure.

The framework of the BAS can serve as the foundation for a diagnostic application that utilizes the structures and knowledge embedded in the construct maps, the items, their scoring guides, and exemplars. Kathleen Scalise of the University of Oregon has created just such a system to serve as the basis for a "homework helper" application running on Windows computers. It uses what she calls the iota model, which is an adaptation of the OPM described above to model complicated item bundles that take a student through a complex problem-solving path. The key is that this solution path can, nevertheless, be unscrambled by using the construct map described above. Figure 14 shows just one example of a branching design for an item bundle in the topic area of ions and atoms. In her work (Scalise, 2004), she shows that, even though the bundles were designed using this complicated branching structure, the value of the resulting data for diagnostic purposes is adequately summed up using just the resultant levels of the construct map. This result, which she achieved by directly modeling the more complicated structure (the iota model) and comparing it to the simpler model with just the levels (essentially a partial-credit model), is of considerable importance in making the results useful for teachers. It means that the complicated way that the results are achieved can be left aside in determining the "next step" in the educational strategy employed by the computer application, and, when used in a classroom situation, it makes it much easier for a teacher to keep track of a classroom of students.

Further Developments

The models that have been used in the two examples discussed above have been confined to relatively simple extensions of the standard item response models–in both cases the ordered partition model sufficed to model the cognitive complexity of the postulated structure, and, in the latter case, a somewhat simpler model was found to be sufficient. Note that, sufficiency in this situation is based on

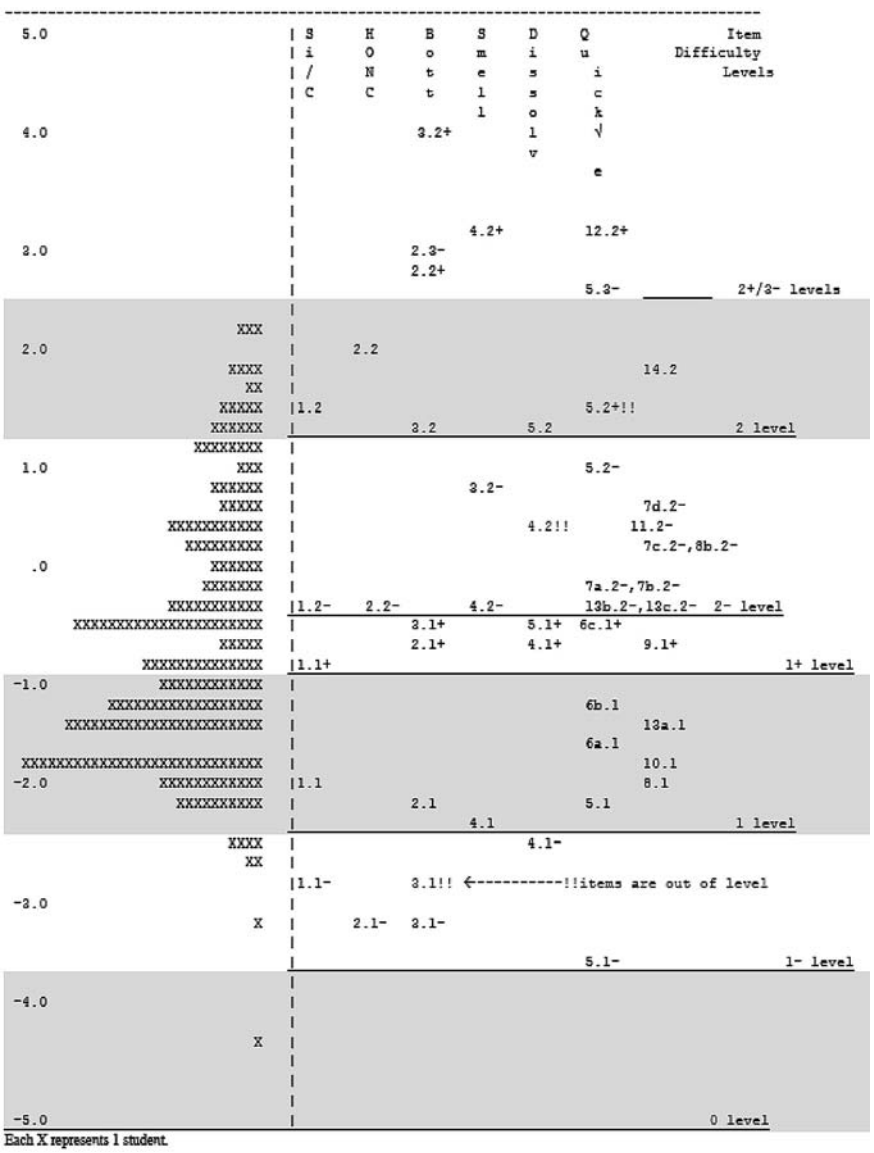
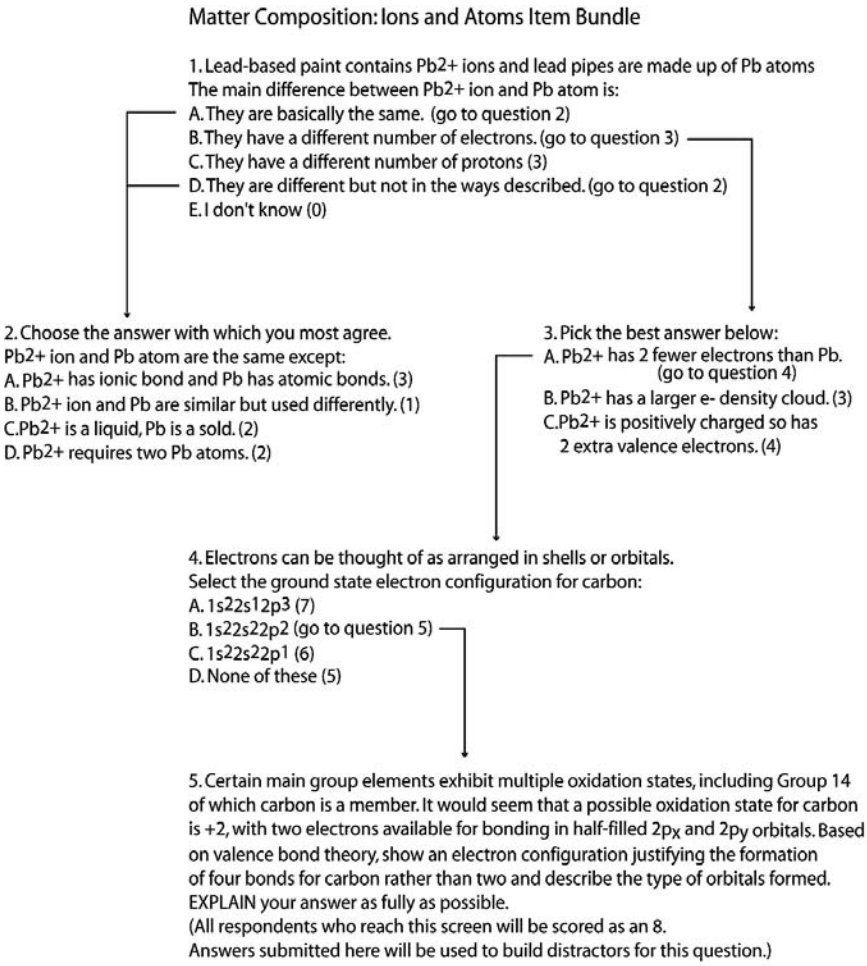


Figure 13. Wright map for the LBC matter variable. The shaded regions indicate the levels of the outcome space for these items, as indicated on the right hand side: from 0, through 1–, 1 and 1+, through to 2+. But also, the notation "6c.1+" indicates the location of the 1+ threshold for iItem 6c.

(a) measures of fit – did the simpler model fit as well as the more complex one? – and (b) judgments about effect size – were the estimates of the parameters representing the extra features large enough to be interpretable? There are a number of contexts in which such models will not suffice because of the inherent complexities of the situation. Several of these are discussed below, along with a sketch of the models one might use to address them. The models are drawn from those described in De Boeck and Wilson (2004), where a scheme for conceptually organizing the models, as well as a generalized statistical approach, is also given. I will not attempt to review that here, as it is beyond the scope of this paper.

One complexity that might enter into the situation of Example 1, above, is that the students could be a part of a cognitive treatment study, where they are trained to use certain strategies. Suppose there are just two groups, for simplicity. In this case, the OPM would need to be supplemented by both a person-group parameter (to indicate overall difference between the two groups), as well as a parameter to model interactions between the person groups and the categories. This second type of parameter is equivalent to a differential-item-functioning model, so the way to model it would be to add a group-level parameter, plus, for each category in each item, a parameter that estimated the difference between the two groups – δ_{gik}, where g indexes the groups. This will allow one to model both overall "effects" of the treatments, and detailed differences among the items (using the δ_{gik} estimates). Similar models and their interpretation are discussed in Meulders and Xie (2004). A more complicated model would result if one did not have observable differences between the two groups, but had to use the results to distinguish them. In this case, either a mixture Rasch model (von Davier & Yamamoto, 2007), or, if one had specific hypotheses about how the



Possible scores and parameters for this item bundle under iota model (assume constraint on cases):

score	0	1	2	3	4	5	6	7	8
# paths	1	1	2	3	1	1	1	1	1
parameters	δ_{71}	δ_{72}	δ_{73}, ι_{731}	$\delta_{74}, \iota_{741}, \iota_{742}$	δ_{75}	δ_{76}	δ_{77}	δ_{78}	δ_{79}

Figure 14. Storyboard showing the design for an item bundle on ions and atoms.

two strategy groups reacted differently to the items, a saltus model (Draney & Wilson, 2007), could be used. Both of these would give information about the size of the two (unobserved) groups, their membership, and item (and item-by-category) differences between them. Both these possibilities have been discussed in a paper by Fieuws, Spiesens, & Draney (2004).

Conclusions

This paper has described an approach to modeling cognitive diagnosis through a combination of misconception analysis and construct modeling. Both techniques have previously been applied to modeling student learning across a range of areas. Their combination has some advantages that are worth noting. The principal advantages of the combination over misconception analysis alone is that (a) it emphasizes how the misconceptions can be seen as positive manifestations of the cognitive level of sophistication of the student rather than as mere "errors," and (b) it makes available the strengths and possibilities of statistical modeling to the misconception analysis. The principal advantage for construct modeling is that it gives a context-based diagnostic interpretability to the estimates of student ability.

Acknowledgments

An early version of this paper was presented at the 4th Spearman Seminar at the University of Pennsylvania, October 2004, sponsored by the Educational Testing Service.

References

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

Baxter, J. (1995). Children's understanding of astronomy and the earth sciences. In S.M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 155–177). Mahwah, NJ: Erlbaum.

Black, P., Harrison, C., Lee, C., Marshall, B., & William, D. (2003). *Assessment for learning*. London: Open University Press.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.

Claesgens, J., Scalise, K., Draney, K., Wilson, M., & Stacy, A. (2002, April). *Perspectives of chemists: A framework to promote conceptual understanding of chemistry.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Confrey, J. (1990). A review of the research on student conceptions in mathematics, science, and programming. In C. Cazden (Ed.), *Review of research in education* (Vol. 16, pp. 3–56). Washington: American Educational Research Association.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer-Verlag.

Draney, K, & Wilson, M. (2007). Application of the Saltus model to stage-like data: Some applications and current developments. In M. von Davier & C.H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 119–130). New York: Springer-Verlag.

Eylon, B., & Linn, M.C. (1988). Learning and instruction: An examination of four research perspectives in science education. *Review of Educational Research, 58*, 251–301.

Fieuws, S., Spiesens, B., & Draney, K. (2004). Mixture models. In P.D. Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213–240). New York: Springer-Verlag.

Kennedy, C.A., Wilson, M., & Draney, K. (2008). *ConstructMap* (computer program). BEAR Center: UC Berkeley, CA.

Kikas, E. (1998). The impact of teaching on students' definitions and explanations of astronomical phenomena. *Learning and Instruction*, *8*, 439–454.

Klein, C.A. (1982). Children's concepts of the earth and the sun: A cross-cultural study. *Science Education, 65,* 95–107.

Linn, R., & Baker, E. (1996). Can performance-based student assessments be psychometrically sound? In J.B. Baron & D.P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. 95th yearbook of the National Society for the Study of Education* (pp. 84–103). Chicago: University of Chicago Press.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *49*, 359–381.

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

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

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

Meyer, J., & Land, R. (2003). *Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practicing within the disciplines* (ETL Occasional Report 4). Edinburgh: School of Education, University of Edinburgh.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P.D. Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213–240). New York: Springer-Verlag.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Niaz, M., & Lawson, A. (1985). Balancing chemical equations: The role of developmental level and mental capacity. *Journal of Research in Science Teaching, 22*, 41–51.

PISA (2005). *Learning for tomorrow's world: First results from PISA 2003.* Paris: OECD.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM Manual* (Working Paper Series, Working Paper 160). Berkeley, CA: U.C. Berkeley Division of Biostatistics.

Roald, I., & Mikalsen, O. (2001). Configuration and dynamics of the earth-sun-moon system: An investigation into conceptions of deaf and hearing pupils. *International Journal of Science Education*, *23*, 423–440.

Sadler, P.M. (1987). Misconceptions in astronomy. In J. Novak (Ed.), *Proceedings of the Second International Seminar on Misconceptions and Educational Strategies in Science and Mathematics* (Vol. III, pp. 422–425). Ithaca, NY: Cornell University Press.

Sadler, P.M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, *35*, 265–296.

SAS Institute (1999). *SAS Online Doc (Version 8)* (software manual on CD-ROM). Cary, NC: SAS Institute.

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

Schneps, M.H., & Sadler, P.M. (1988). *A private universe*. Santa Monica, CA: Pyramid Films.

Smith, J.P., diSessa, A.A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3*, 115–163.

Stahly, L.L., Krockover, G.H., & Shepardson, D.P. (1999). Third grade student' ideas about the lunar phases. *Journal of Research in Science Teaching*, *36*, 159–177.

Targan, D.S. (1987). A study of conceptual change in the content domain of the lunary phase. In J. Novak (Ed.), *Proceedings of the Second International Seminar on Misconceptions and Educational Strategies in Science and Mathematics* (Vol. II, pp. 499–511). Ithaca, NY: Cornell University Press.

Trumper, R. (2001). A cross-age study of science and nonscience students' conceptions of basic astronomy concepts in preservice training for high school teachers. *Journal of Science Education and Technology*, *10*, 189–195.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and hybrid Rasch models. In M. von Davier & C.H. Carstensen

(Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–118). New York: Springer-Verlag.

Vosniadou, S. (1991). Conceptual development in astronomy. In S.M. Glynn, R.H. Yeany, & B.K. Britton (Eds.), *The psychology of learning science* (pp. 149–177). Hillsdale, NJ: Erlbaum.

Vosniadou, S., & Brewer, W.F. (1994). Mental models of the day/night cycle. *Cognitive Science, 18,* 123–183.

Wilson, M. (1990). Measurement of developmental levels. In T. Husen & T.N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies* (Supplementary volume 2, pp. 628–634). Oxford: Pergamon.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, *16*, 309–325.

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

Wilson, M., & Adams, R. (1993). Marginal maximum likelihood estimation for the ordered partition model. *Journal of Educational Statistics*, *18*, 69–90.

Wilson, M., & Carstensen, C. (2007). Assessment to improve learning in mathematics: The BEAR Assessment system. In A. Schoenfeld (Ed.), *Assessing mathematical proficiency* (pp. 311–332). London: Cambridge University Press.

Wilson M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson (Ed.), *Toward coherence between classroom assessment and accountability* (103rd Yearbook of the National Society for the Study of Education, Part II, pp. 126–152). Chicago: University of Chicago Press.

Wilson, M., & Scalise, K. (2006). Assessment to improve learning in higher education: The BEAR Assessment System. *Higher Education, 52,* 635–663.

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

Wu, M., Adams, R., & Wilson, M. (1998) *ACER ConQuest: Generalized item response modeling software*. Camberwell, Australia: ACER Press.

Mark Wilson

4415 Tolman Hall
University of California, Berkeley, Graduate School of Education
Berkeley, CA 94720-1670
USA
Tel. +1 510 642-7966
E-mail MarkW@berkeley.edu

