THE PSYCHOMETRIC MODELING OF ORDERED MULTIPLE-CHOICE ITEM RESPONSES FOR DIAGNOSTIC ASSESSMENT WITH A LEARNING PROGRESSION

Derek C. Briggs, University of Colorado
Alicia C. Alonzo, Michigan State University

Pre-print from the book *Learning Progressions in Science*

Acknowledgements:
Initial development of the Earth in the Solar System learning progression and associated items was conducted at the University of California, Berkeley, in collaboration with Cheryl Schwab and Mark Wilson. This work was supported with funding from the National Science Foundation (#REC-0087848, Research in Standards-based Science Assessment), as part of a collaboration between the BEAR Center and WestEd. Additional data collection was funded by the University of Iowa Department of Teaching & Learning and the Iowa Mathematics & Science Education Partnership. Any opinions, findings, conclusions, or recommendations expressed in this chapter are those of the authors. They do not necessarily represent the official views, opinions, or policy of the National Science Foundation or other funding agencies.

Introduction

One of the more appealing features of learning progressions is their potential to facilitate diagnostic assessment of student understanding. In this context, diagnostic assessment hinges upon the development of items (i.e., tasks, problems) to efficiently elicit student conceptions that can be related back to a hypothesized learning progression. Briggs, Alonzo, Schwab & Wilson (2006) introduced Ordered Multiple-Choice (OMC) items as a means to this end. OMC items represent an attempt to combine the efficiency of traditional multiple-choice items with the qualitative richness of responses to open-ended questions. The potential efficiency comes because OMC items feature a constrained set of response options that can be scored objectively; the potential qualitative richness comes because OMC response options are both designed to correspond to what students might answer in response to an open-ended question and explicitly linked to a discrete level of an underlying learning progression. The OMC item format belongs to a broader class of constrained assessment items in which the interest is not solely in whether a student has chosen the "scientifically correct" answer, but on diagnosing the reasons behind a student's choice of a *less* scientifically correct answer (c.f., Minstrell, n.d., 1992, 2000). An appealing aspect of such items is that they are consistent with the spirit behind learning progressions, which at root represent an attempt to classify the gray area of cognition that muddies the notion that students either "get something" or they don't.

This chapter illustrates some of the challenges inherent to the psychometric modeling of a learning progression, using the context of a specific learning progression and an associated set of OMC items. There are two reasons why we view formal psychometric modeling as central to the burgeoning interest in learning progressions. First, a psychometric model can be used to draw probabilistic inferences about unobserved (i.e., latent) states of student understanding. This makes it possible to quantify the extent to which a student has mastered the content of a given learning progression. Second, the process of specifying a model and evaluating its fit can provide a systematic means of validating and refining a hypothesized learning progression. When the model that has been proposed for the purpose of making diagnostic inferences falls short because, for example, it does not support reliable diagnoses, or because the diagnoses do not correspond to other evidence of what a student appears to understand, this can raise important questions for any learning progression development team to address. Is the problem with the way student thinking is being elicited (i.e., the assessment instrument)? Or do the levels of the hypothesized learning progression need to be revised? Or perhaps has the wrong psychometric model been specified?

There are five sections that follow. In the first section we provide a brief background about the previous development of a relatively simple learning progression and associated set of OMC items focused on conceptual understanding of Earth and the Solar System. In the second section we present descriptive statistics from a recent administration of these OMC items to a convenience sample of 1,088 high school students in Iowa, along with some of the limitations when making inferences about student understanding on the basis of classical item statistics. In the third section we discuss the inherent challenges involved in choosing an approach to model student responses as a means of making diagnostic inferences about their levels of understanding, and we distinguish between approaches based on item response theory models and those based on diagnostic classification models. In the fourth section we introduce the

Attribute Hierarchy Method (AHM; Leighton, Gierl & Hunka, 2004) as a relatively novel approach for modeling OMC items. The AHM is a diagnostic classification model that builds upon the seminal work of Tatsuoka who developed what is known as the Rule Space Method for cognitive assessment (Tatsuoka, 1983, 2009). To our knowledge, the AHM has never been applied in the context of a learning progression in the domain of science. We illustrate the steps that would be necessary to apply the AHM to OMC item responses as a means of producing diagnostic student classifications. Finally, in the fifth section we speculate about strengths and weaknesses of the AHM approach.

Background

In previous studies we have developed learning progression hypotheses in the science content domains of earth science, life science and physical science (Alonzo & Steedle, 2009; Briggs et al., 2006). In this chapter we use one of these previously developed learning progressions, which focuses on conceptual understanding of Earth in the Solar System (ESS), as the context for the modeling issues that follow. The ESS learning progression describes students' developing understanding of target ideas in earth science, which, according to national science education standards documents, they should master by the end of 8th grade (American Association for the Advancement of Science [AAAS], 1993; National Research Council [NRC], 1996). However, there is substantial evidence that typical instruction has not been successful in helping students to achieve these levels of understanding. In fact, many college students retain misconceptions about these target ideas (e.g., Schneps & Sadler, 1987).

Initial development of the ESS learning progression followed the same process as that used for all other learning progressions in our work and is described in more detail in Briggs et al. (2006). We began by defining the top level of our learning progression, relying upon national science education documents (AAAS, 1993; NRC, 1996). With respect to ESS, by the end of 8th grade, students are expected to use an understanding of the relative motion of the Earth and other objects in the Solar System to explain phenomena such as the day/night cycle, the phases of the Moon, and the seasons. Lower levels of the learning progression (i.e., novice understanding, intermediate understanding, etc.) were defined using research literature on students' understanding of the targeted concepts (e.g., Atwood & Atwood, 1996; Baxter, 1995; Bisard, Aron, Francek, & Nelson, 1994; Dickinson, Flick, & Lederman, 2000; Furuness & Cohen, 1989; Jones, Lynch, & Reesink, 1987; Kikas, 1998; Klein, 1982; Newman, Morrison, & Torzs, 1993; Roald & Mikalsen, 2001; Sadler, 1987, 1998; Samarapungavan, Vosniadou, & Brewer, 1996; Stahly, Krockover, & Shepardson, 1999; Summers & Mant, 1995; Targan, 1987; Trumper, 2001; Vosniadou, 1991; Vosniadou & Brewer, 1992; Zeilik, Schau, & Mattern, 1998). In defining the levels, we relied upon information about both "misconceptions" and productive – but naïve – ideas that could provide a basis for further learning. While the target level of understanding at the top of this learning progression is linked to the AAAS and NSES expectations for an 8th grade student, the levels below are intended to represent understandings that are expected to develop from kindergarten through the middle grades.

At this point, it is important to note two key limitations of the available research base for the construction of this (and most other) learning progressions. Although learning progressions aim to describe how understanding develops in a given domain, the available research evidence is

primarily cross-sectional. So while we have important information about the prevalence of particular ideas at different ages, there has been little documentation of individual students actually progressing through these ideas over time as a result of instruction. In addition, much of the work in the area of ESS occurred in the context of an interest in students' misconceptions in the 1980s; therefore, research has tended to focus upon isolated (incorrect) ideas, rather than exploring the relationship between students' ideas (both correct and incorrect) in a given domain. Since the ESS learning progression encompasses multiple phenomena – the Earth orbiting the Sun, the Earth rotating on its axis, and the Moon orbiting the Earth – the definition of levels included grouping ideas about these phenomena on the basis of both experience and logical reasoning from experts. Thus, the learning progression represents a hypothesis – both about the ways in which students actually progress through identified ideas and about the ways in which ideas about different phenomena "hang together" as students move towards the targeted level of understanding. This hypothesis must be tested with further evidence, such that the development of a learning progression and its associated assessment items is an iterative process – the learning progression informs the development of assessment items; the items are used to collect data about student thinking; and this data is linked back the initial progression through the use of a psychometric model, which leads to revisions in both the items and the learning progression itself.

The current version of the ESS learning progression is depicted in Figure 1. Within the science education community, there is great interest in learning progressions which not only specify different levels of student knowledge, but also include the way(s) in which students can be expected to demonstrate that knowledge (for example, through assessment tasks). Smith, Wiser, Anderson, & Krajcik (2006) have called for learning progressions to include "learning performances" (p. 9). In the ESS learning progression, such learning performances are implied: students are expected to use the targeted knowledge to explain or predict phenomena such as the day/night cycle, the phases of the Moon, and the seasons.

Insert Figure 1 about here

Two examples of OMC items that were developed to assess the location of students on the ESS learning progression are shown in Figure 2.

Insert Figure 2 about here

At first glance the items resemble the typical multiple-choice format found in most standardized exams. What makes these items different is that each response option is intended to represent a qualitatively distinct level of understanding about ESS. While each item contains a single response option that is considered the "most" correct, students are given partial credit if they select a response that represents developing understanding of the phenomenon in the item stem. (For more detail on how these items were developed, see Briggs et al., 2006.) At this point we note three features of these OMC items that present impediments to any attempts at diagnostic inference:

1. For some items, it is not possible to write (and consequently, for students to select) a response at the *highest* levels of the ESS learning progression. This constitutes a

"ceiling effect" at the item level. For example, for item 3 in Figure 2 the highest possible response level is a 4.

2. For some items, it is not possible to write (and consequently, for students to select) a response at the *lowest* levels of the ESS learning progression. This constitutes a "floor effect" at the item level. For example, for item 2 in Figure 2 the lowest possible response level is a 3.
3. Many items feature response options that are linked to the same level. This increases the likelihood that students might select an option indicating a particular level of understanding by chance if they are guessing. For example, for both items 2 and 3 in Figure 2, two out the five possible response options correspond to a level 3.

The three possible features of OMC items described above could occur for practical reasons or by design. From a practical standpoint, it may not be possible to find response options that span the full range of a learning progression because the highest level of the learning progression may not be required to fully explain a given context or because more complex contexts may not be accessible for students at the lower levels of the learning progression. In addition, a highest level response will often involve jargon that makes it stand out too easily as the most correct response through contrast to lower level responses, and vice-versa. Such features could also be a matter of design preference. A ceiling effect might exist by design for a subset of OMC items if it is known that students have yet to receive the instruction that should be needed before they could respond at the highest level. A floor effect might be imposed on a subset of OMC items if it seems reasonable to assume that all students have already mastered to skills and concepts at lower levels. And finally, an item may include multiple responses at the same level because the responses are qualitatively distinct, but cannot be ordered. Even though these options do not add additional information with respect to a student's level on the LP, they may provide *qualitative* information about nuances in students' thinking and are important to include if there are multiple typical ways of thinking that are consistent with a particular level of the learning progression.

Data and Classical Item Statistics

During the 2008-09 school year, a science test was administered to a sample of 1,088 high school students (grades 9-12) attending six different high schools in rural and suburban Iowa. Any student enrolled in a high school science course at these schools was eligible to participate in the study, although not all science teachers granted permission for data collection in their classes. Participating students were drawn from 68 different high school science classes, representing a range of different science courses – including those enrolling primarily freshman, as well as upper-level courses. The science test consisted of 28 OMC items, 12 of which were associated with a hypothesized learning progression for ESS and 16 of which were associated with a hypothesized learning progression on the topic of force and motion. We focus here on the results from the ESS OMC items.

Students who agreed to answer the OMC items did so in their regular science classes. The average participation rate across all classes was 83%. The sample was fairly evenly divided between male and female students (52% male; 48% female). High school students were chosen because the majority of these students could be expected to have been exposed to ideas relevant to the two learning progressions; this was thought to minimize guessing. On the other hand, a

drawback to this sample is that we are less likely to find students choosing options consistent with the lower end of the learning progressions. After completing the ESS OMC items, students answered a question which asked "Was the content of [these] questions covered in a science class you've taken?" While 46% of our sample responded "yes," another 25% answered "no", 28% answered "I am not sure" and 2% did not respond at all. Part of this can be explained by the fact that ESS is not consistently covered as part of high school science curricula.

Table 1. Observed Distribution of ESS OMC Item Responses

	Easier ← OMC Items → Harder											
Level	11	6	3	1	4	12	10	5	9	7	2	8
5									43		28	20
4	74	72	66	64	63	62	61	59	41	29	34+15	35
3	5	16+4	15+5	14+6	21	12	18	14	7	47+14	10+14	18
2	7+14	7+2	12	12	11+6	14	7+5	7	9	11		15+11
1			3	5		12	9	14+5				
pt-bis	0.59	0.69	0.60	0.64	0.54	0.69	0.68	0.68	0.33	0.12	0.36	0.21

Note: Values in cells represent the percentage of students choosing a response option linked to a level of the underlying learning progression. Some columns may not sum to 100 due to rounding error

Table 1 shows the distribution of student OMC item responses mapped back to the levels of the ESS progression. The items in the columns of Table 1 are arranged from easiest to hardest, where "easiness" is defined as the proportion of students selecting a response option at the highest possible level. So, for example, 74% of students selected the highest possible response level for item 11 ("Which picture best represents the motion of the Earth (E) and Sun (S)?**"),** making it the easiest OMC item, while only 20% of students selected the highest possible response option for item 8 ("Which is the best explanation for why we see a full moon sometimes and a crescent moon other times?"), making it the hardest OMC item. The shaded cells represent levels for which there was no corresponding response option for the OMC item. There were only three items for which a response at the highest level of the ESS learning progression was possible, and on 5 out of 12 items, OMC responses were linked to only 3 out the 5 possible levels. Cells with two numbers expressed as a sum represent items for which two options were associated with the same score level. For example, on item 11, roughly 7% of students chose response option A and 14% selected response option B, but both options are linked to level 2 of the ESS progression. Finally, the last row of Table 1 provides the item to total score correlations (i.e., point-biserial) associated with the highest level response option for each OMC item.

Table 1 conveys information about classical item statistics that some would use as a basis for evaluating item quality. For example, from item to item, a majority of students select responses linked to the highest two available levels of the ESS progression (i.e., levels 3 or 4 or levels 4 or 5). Interestingly, all three items for which a response at level 5 was available had point-biserials less than 0.4, a value that is typically considered a cutoff for a "good" item in traditional testing contexts. In other words, the students selecting this option were not necessarily those who performed the best on the remaining items. Item 7 seems to stand out as a problematic item because there is a very low correlation between a choice of level 4 and the total score on the

remaining items (point-biserial = .12).[1] Finally, one can use true score theory to estimate the reliability of the total scores deriving from these items. An estimate based on Cronbach's alpha coefficient suggests a reliability of 0.67. This implies that about 1/3 of the variance in OMC scores across students could be attributable measurement error. If the results from these items were to be used to support high-stakes inferences about individual students, this would be a cause for concern. On the other hand, if the scores were being used for formative purposes or to compare group means, a reliability of 0.67 might be less worrisome.

There is certainly nothing "wrong" with the analysis above, which seems to suggest that some of the items from the test may need to be rewritten, or that the links between the options and ESS levels need to be reconsidered. However, these interpretations are somewhat arbitrary because, as is well-known, they are highly dependent on the particular sample of students taking the exam. If, for example, we were to find that the highest proportion of students were responding to options associated with levels 2 and 3, would this indicate a problem with the ordering of the item options, or could it reflect the fact that the students have not been exposed to this content in their curriculum? Furthermore, the analysis above could provide equivocal diagnostic information when it is disaggregated at the student level. Consider the following two randomly selected student response vectors for "Liz" and "Andrew" shown in Figure 3.

Figure 3. Observed ESS Level Classifications for Two Students' Item-Level Responses

"Liz"

Level	11	6	3	1	4	12	10	5	9	7	2	8
5	■	■	■	■	■	■	■	■		■		
4	X	X		X	X	X	X	X	X	X	X	
3			X									X
2											■	
1	■	■	■		■				■	■	■	■

"Andrew"

Level	11	6	3	1	4	12	10	5	9	7	2	8
5	■	■	■	■	■	■	■	■	X	■		X
4	X		X		X	X	X			X		
3		X						X			X	
2											■	
1	■	■	■	X	■				■	■	■	■

If the set of responses for each student were to be summarized by a report of central tendency, one would find that the median score level for both students is 4, while the arithmetic average for Liz (3.85) is just slightly higher than that for Andrew (3.69). Yet clearly Andrew's responses show greater variability than Liz's responses, and this implies greater uncertainty about the diagnostic utility of the information. While many teachers are able to make these sorts of

[1] The stem for this item was "A solar eclipse is possible because" and the level 4 response option was "The Sun is much bigger than the Moon and much further away from the Earth." The level 3 response chosen more frequently was "The Moon is always closer to the Earth than the Sun is."

informal inferences qualitatively on a student-by-student basis, this can be a very subjective and time-consuming process. Thus, there are some advantages to be realized if the diagnostic process above could be engineered through the use of a psychometric model. Such a model would serve both to test the validity of the hypothesized learning progression and to provide for formal probablistic inferences about student understanding. In the next section we discuss some challenges inherent to this endeavor before presenting one possible solution—the AHM—in detail.

Challenges to Modeling OMC Items to Support Diagnostic Inferences

We now draw attention to two important and related challenges that arise after a decision has been made to model OMC item responses probabilistically for the purpose of making diagnostic inferences. The first challenge is to decide upon the functional form of the model; the second is to decide about assumptions that can plausibly be made about whether the latent variable (or variables) being "measured" can be viewed as discrete or continuous.

To speak of "modeling" item responses is to make a formal statement about the factors involved when a student interacts with an assessment item. In item response theory (IRT; De Boeck & Wilson, 2004; van der Linden & Hambleton, 1996), this formal statement is made in terms of an *item response function* (IRF). Let the variable X_{pi} represent possible responses to assessment item *i* that could be given by student *p*. An IRF provides a mathematical expression for the probability of observing a response in score category k as a function of one or more parameters (i.e., dimensions) specific to respondents ($\boldsymbol{\theta}_p$), and one or more parameters specific to items ($\boldsymbol{\xi}_i$):

$$P(X_{pi} = k) = f(\boldsymbol{\theta}_p, \boldsymbol{\xi}_i). \quad (1)$$

The very general expression above accommodates IRFs that range from very simple (a single parameter for each student and a single parameter for each item) to very complex (multiple parameters per individual student and item). One well-known example of an IRF that is often applied when modeling student responses to the traditional multiple-choice items found on most large-scale assessments is the three-parameter logistic model (3PL; Birnbaum, 1968):

$$P\left(X_{pi} = 1 | \theta_p\right) = c_i + \left(1 - c_i\right)\frac{e^{a_i\left(\theta_p - b_i\right)}}{1 + e^{a_i\left(\theta_p - b_i\right)}}. \quad (2)$$

The 3PL model gets its name from the fact that three distinct parameters are specified for every item to which a student responds (a_i, b_i, c_i). Values of the parameter a_i affect the *slope* of the IRF. The larger the value of a_i, the steeper the curve. This means that on items with relatively large values of a_i, a small change in (unidimensional) θ_p will produce a large change in the probability of a correct response. Because such items appear to be better at discriminating between respondents with different underlying values of θ, it is sometimes referred to as the item discrimination parameter. In contrast, values of the parameter b_i affect the *location* of the IRF.

The larger the value of b_i for an item, the larger the value of θ that would be needed for a respondent to have a high probability of answering the item correctly. This is also often referred to as the item difficulty parameter. Finally, values of c_i, which can in theory range between 0 and 1, establish a lower asymptote for the item response function. The larger the value of c_i for an item, the higher the "floor" on the probability for any respondent to answer the item correctly. Because it is intended to capture the possibility that respondents have answered an item correctly by guessing, c_i is often referred to as a guessing parameter. The inclusion of a single student-specific variable θ in the expression above brings the total number of parameters to four.

An example of an IRF with a simpler functional form is the Rasch Model (Rasch, 1960):

$$P\left(X_{pi}=1|\,\theta_p\right)=\frac{e^{\left(\theta_p-b_i\right)}}{1+e^{\left(\theta_p-b_i\right)}}. \qquad (3)$$

Upon inspection, the mathematical difference between the IRFs in equations (2) and (3) is that in the latter, the item parameters a_i and c_i have been constrained to be 1 and 0 respectively. Yet there are also important differences between the two IRFs that are not so much mathematical as they are philosophical. In the Rasch tradition, items are developed to fit the model because when this particular IRF fits the data it allows for *invariant comparisons* between respondents. That is, comparisons of students will not vary as a function of the specific items chosen for a test instrument, just as comparisons of items will not depend upon the specific sample of test-takers that responded to them. The alternative tradition is to view the data as fixed and choose an IRF that bests fits the data—whether this leads to the Rasch Model or something much more complex. (A full discussion of these two positions is well outside the scope of the present chapter, but for details see Andrich, 2004; Bock, 1997; Thissen & Wainer, 2001; Wilson, 2005; Wright, 1997). We raise this issue to make the broader point that that even when one has developed assessment items with traditional score formats, the choice to be made when it comes to selecting an IRF using an IRT-based approach is not straightforward. The problem is that there are two different criteria for optimality. On one side is the technical need to model observed item responses as faithfully as possible, on the other the practical need for models that are parsimonious and readily interpretable. The BEAR Assessment System, which has been previously applied to model learning progressions in science education is an example of an IRT-based approach that tends to prioritize the latter (Wilson, 2009).

Because learning progressions attempt to distinguish between multiple levels of understanding, the sorts of items that would be developed to accomplish this will often (if not usually) need to be scored in more than two ordinal categories (i.e., polytomously). The OMC format described in this chapter is a case in point. At a minimum, the specification of an IRF for OMC items would need to take this added complexity in scoring into account. Beyond this, decisions would need to be made with regard to a parameterization that addresses the obstacles to score interpretations noted previously: floor and ceiling effects, as well as multiple options linked to the same score level. One paradigm for this can be found in the book *Explanatory Item Response Models* (De Boeck & Wilson, 2004). In this edited volume, many examples are given in which the conventional IRFs of IRT can be expanded through the addition of new variables and parameters

that serve to explain variability in observed item responses. As one example, to capture the notion that students may guess the correct OMC response, one solution might be to posit what is known as a "mixture model" (Mislevy & Verhelst, 1990; Wilson, 1989) in which there are effectively two populations of students: those that guess when they don't know the most sophisticated response option and those that do not guess in the same circumstance. In such a case it would be possible to specify two distinct IRFs, one for each hypothetical population of students. As a different example, to capture the fact that certain items have multiple options at the same level while others do not, one solution might be to posit what is known as the "logistic latent trait model" (Fischer, 1977; Fischer, 1983) in which item difficulty is modeled as a function of both the levels of possible response options and the number of these options.

In recent years, several authors have argued that even these elaborated IRFs may not be ideal if the purpose of the model is to make diagnostic classifications of students (Junker & Sijtsma, 2001; Leighton & Gierl, 2007; Rupp, Templin & Hensen, 2010). This is because IRT models posit that the latent variable (or variables) underlying a student's item responses are continuous. As a result, the estimation of student-specific values for these variables do not lead to direct classification of students into discrete categories. Rather, a second step is taken in which "cut-points" are established along the student-specific latent variable θ. This will typically require some degree of subjective judgment, as is the case when criterion-referenced standards are established for student performance on large-scale assessments.

This brings to the fore that what is meant when θ is defined as a "latent variable" or student "ability" is often rather murky. In the context of learning progressions, one might say that θ represents at least one of the attributes of a person that becomes more sophisticated as he or she receives instruction. In this sense θ is some unknown (i.e., latent) variable that takes on values spanning multiple levels of a hypothesized learning progression. But what is the mapping between the values of the latent variable and the levels of the learning progression? If θ is assumed to be continuous while the levels of the learning progression are discrete, then to some extent there will be a mismatch between the granularity of the hypothesis that underlies the design of assessment items and the granularity of the latent variable that underlies the design of the psychometric model being used to interpret and diagnose subsequent item responses. Such a mismatch seems inherent when θ is defined as a continuous latent variable in an IRT-based approach.

An alternative is the specification of what Rupp et al. (2010) have described as *diagnostic classification models* (DCMs). DCMs can be distinguished as models in which the latent variables of interest are discrete rather than continuous, and the objective of the model is to provide a profile of the knowledge and skills of individuals based on statistically derived classifications (Rupp & Templin, 2008).While providing a taxonomy of models that would fit this definition is well outside the scope of this chapter (see Rupp et al., 2010 for these details), in the next section we illustrate the basic principles involved when taking a DCM-based approach by showing how the Attribute Hierarchy Method (AHM) could be used to model OMC item responses according to the hierarchy implied by the ESS learning progression.

Before proceeding, we wish to make clear that we are not arguing that IRT-based approaches are an invalid means of making diagnostic inferences. Even when the assumptions of an IRT model

are wrong and that of an DCM are right (and unfortunately, the truth is never known a priori), the former may well serve as a first-order approximation of the latter or vice-versa. From a pragmatic standpoint, this is an empirical question that we are not addressing in this chapter. A latent variable is, after all, by definition unobservable, so assumptions are unavoidable. But in our view, the assumption that a latent variable has a continuous structure (implicit to IRT) is much stronger and less plausible than the assumption that the variable has an ordinal structure (implicit to a DCM)[2]. This provides some motivation for the approach described in what follows. We speculate about other pros and cons of taking an DCM-based approach relative to an IRT-based approach in the final section of the chapter.

Applying the Attribute Hierarchy Method to OMC Items

Background on the AHM

There has been an explosion in the development of DCMs for cognitive diagnostic assessment over the past decade.[3] Much of the interest in such models comes from the pioneering work of Kikumi Tatsuoka, beginning in the early 1980s. Tatsuoka's premise is fairly simple: that the score derived from a set of items (i.e., θ in IRT) often obscures important diagnostic information about more fine-grained "attributes" that students use to solve problems within a given domain. To address this issue, Tatsuoka developed the idea of a Q matrix that allows for the formal specification of a hypothesized linking between attributes and items. Specification of a Q matrix makes it possible to generate expected item response patterns associated with specific knowledge states, where the latter is defined by the attributes that a test-taker does or does not have. Given these expected response patterns and the actual response patterns produced by test-takers, Tatsuoka developed the Rule Space Method as a pattern-matching technique for probabilistic diagnostic classification.

More recently, Leighton et al. (2004) introduced an extension of Tatsuoka's Rule Space Method called the Attribute Hierarchy Method (AHM). The AHM takes as a starting point the assumption that the construct of measurement is comprised of finer-grained "attributes" that have an ordered, hierarchical relationship. The specification of this relationship precedes and guides the specification of a "reduced form" Q_r matrix. While many DCM applications assume that all attributes are independent and/or non-hierarchical, in the AHM, a hierarchical dependence among attributes is central to the theory. In our view this feature makes the AHM an appealing candidate for the modeling of learning progressions, which also make explicit the hierarchical distinctions in student understanding as it becomes more sophisticated. Applications of the AHM to date have involved traditional multiple-choice items that are scored dichotomously (Leighton et al., 2004; Gierl, Wang, & Zhou, 2008). The application of the AHM to polytomously scored OMC items represents a novel extension.

[2] For detailed arguments in support of this perspective, see Michell (1990, 2008).

[3] For books, see Leighton & Gierl (2007); Tatsuoka (2009); Rupp et al. (2010). For an example of journal articles, see a special issue of the *Journal of Educational Measurement* co-edited by Dibello & Stout, Volume *44*(1), Winter 2007. For conference symposia, peruse the programs of the annual meetings of the National Council for Measurement in Education between 2007 and 2010.

There are two stages to the AHM. In the first stage, an attribute hierarchy is specified based upon the construct of measurement and then is used to characterize the cognitive features of items through a Q_r matrix. This makes it possible to generate distinct expected item response patterns that characterize the pre-specified attribute combinations comprising the hierarchy. Once this has been accomplished, in a second stage, expected response patterns are compared to observed item response patterns using either a parametric or a nonparametric statistical classification approach. The result is a set of probabilities that characterize the likelihood of a student with a given response pattern having a level of understanding consistent with a hypothesized level within the attribute hierarchy. Along with these probabilities, one can also generate hierarchy fit indices and estimates of reliability at the attribute level. In what follows we illustrate the first stage of the AHM as it could map to the ESS learning progression and OMC items. We then illustrate the gist of the second stage and provide a hypothetical illustration of how the results from this stage could be used diagnostically.[4]

Stage 1: Specifying a Learning Progression as an Attribute Hierarchy

We begin by translating the qualitative descriptions that distinguish the levels of our existing ESS learning progression (Figure 1) into attributes that can be coded dichotomously as either present or absent in any given test-taker.

A1: Student recognizes that there is some systematic nature to objects in the sky.
A2: Student knows that the Earth orbits the Sun, the Moon orbits the Earth, and the Earth rotates on its axis.
A3: Student is able to coordinate apparent and actual motion of objects in sky.
A4: 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, phases of the Moon, and the seasons.

The proper grain size of these attributes will always be a matter for debate. For example, the attribute A2 could easily be split into three smaller attributes. The more finely specified the attributes, the easier they are to code as present or absent. On the other hand, the larger the number of attributes, the harder they are to distinguish with a finite number of test items, and the more difficult they are to summarize as a diagnostic assessment of student understanding. We will return to this issue in the concluding section of this chapter.

Next we specify a hierarchy among these attributes. In this example, the hierarchy is fairly straightforward and mirrors the hierarchy implicit in the original ESS learning progression: A1 → A2 → A3 → A4. These attributes are conjunctive—a student must possess an attribute lower in the hierarchy (e.g., A1) in order to possess an attribute that is higher (e.g., A4). The combinations of these four attributes can be used to define the levels of the ESS learning progression.

Level 1 = No attributes
Level 2 = A1

[4] The actual implementation of the AHM approach with these OMC items is outside the scope of the present chapter but will be the focus of a forthcoming manuscript.

Level 3 = A1 & A2
Level 4 = A1 & A2 & A3
Level 5 = A1 & A2 & A3 & A4

The simple attribute hierarchy above leads to the specification of two matrices. An "adjacency" matrix

$$A = \begin{vmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{vmatrix}$$

and a "reachability" matrix

$$R = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix}.$$

The *A* matrix represents all the direct dependencies between attributes, with each row and column combination above the main diagonal representing a unique attribute combination. In this example, the first row of the matrix has the following interpretation: knowing that the Earth orbits the Sun, the Moon orbits the Earth, and the Earth rotates on its axis (A2, 2nd column) depends directly upon first knowing that there is a systematic nature to objects in the sky (A1, 1st row). The other nonzero cells in the A matrix have a similar interpretation. The *R* matrix represents both direct and indirect dependencies. Hence row 1 of the matrix indicates that attributes A2, A3, and A4 all depend on attribute A1. For A2 the dependency is direct (as indicated in the *A* matrix); for A3 and A4 the dependency is indirect. The *A* and *R* matrices can be manipulated through the use of Boolean algebra to specify the "reduced" Q matrix Q_r. In applications of the AHM with traditional multiple-choice items, a Q_r matrix has dimensions *a* by *I*, where *a* represents the number of attributes, and *I* represents the number of items. Because OMC items are scored polytomously, the associated Q_r matrix will be considerably more complicated. For our set of items, instead of a 4 attribute by 12 item matrix, it will be a 4 attribute by 55 item option matrix, since each item-specific option is given a separate column (and items in this set contained 4 or 5 options). For ease of presentation, we illustrate in Figure 4 an excerpt of the Q_r matrix for the ESS OMC items using only items 2 and 3 that were shown in Figure 2.

Figure 4. Excerpt of the Q_r matrix associated with ESS Attribute Hierarchy

	2A	2B	2C	2D	2E	3A	3B	3C	3D	3E
A1	1	1	1	1	1	1	1	1	0	1
A2	1	1	1	1	1	0	1	1	0	1
A3	1	0	0	1	1	0	1	0	0	0
A4	0	0	0	1	0	0	0	0	0	0
Level	4	3	3	5	4	2	4	3	1	3

Note: Columns represent OMC item options; rows represent hypothesized attributes of test-takers that must be present to select each item option. The row and column labels, as well as the indication of the learning progression level corresponding to each item option, are included to make the matrix easier to interpret.

The interpretation of the Q_r matrix for OMC item 2 ("Which is the best reason for why we experience different seasons?") option A ("the earth's orbit around the sun makes us closer to the Sun in the summer and farther away in the winter") is as follows: in order to select this response, a student should possess attributes A1, A2 and A3. However, attribute A4 is not a prerequisite to selecting option A. The columns for the other possible response options have similar interpretations. The presence of a "1" in a given row indicates that the associated attribute is a prerequisite for the response; the presence of a "0" indicates an attribute that is not a prerequisite. The Q_r matrix leads naturally to the specification of an expected response matrix for OMC items, where each row of the matrix represents the expected response to each OMC option for students with each conceivable attribute combination. Note that what is expected at the item option level hinges upon the central hypothesis that the attribute structure and its relationship to items is accurate. Figure 5 shows the excerpt from an expected response matrix that would correspond to the Q_r matrix in Figure 4.

Figure 5. Excerpt from an Expected Response Matrix for ESS OMC Items

Hypothetical Examinee	**Expected Responses by Item [2] [3]**	**Attributes [A1 A2 A3 A4]**	**ESS Level**
1	$\left[\frac{1}{5}\frac{1}{5}\frac{1}{5}\frac{1}{5}\frac{1}{5}\right]$ [00010]	0000	1
2	$\left[\frac{1}{5}\frac{1}{5}\frac{1}{5}\frac{1}{5}\frac{1}{5}\right]$ [10000]	1000	2
3	$\left[0\frac{1}{2}\frac{1}{2}00\right]\left[00\frac{1}{2}0\frac{1}{2}\right]$	1100	3
4	$\left[\frac{1}{2}000\frac{1}{2}\right]$ [01000]	1110	4
5	[00010] [01000]	1111	5

Note: The row and column labels along with the last column ("ESS Level") are included to make the matrix easier to interpret.

The expected item option responses for OMC items 2 and 3 are given within the brackets in the second column Figure 5. Take a hypothetical examinee with a level 1 understanding of ESS according to our learning progression. This is a student who does not yet have attributes 1 through 4. Yet for item 2, all possible response options require at least one of these attributes, so we could reasonably assume that such a student would be guessing among the available response

options; hence we insert a 1/5 for each expected response. (An alternative would be to give the item options associated with fewer attributes – or lower levels of the learning progression - higher probabilities than those with more.) In contrast, item 3 does include a response option (D) that requires no attributes. Hence the expected response string for this hypothetical examinee is [00010].

Notice that it is the combination of the attribute hierarchy (the *A* matrix) and the Q_r matrix (Figure 4) that are used to generate conditional expected item responses—the student by item response combinations we would expect to observe *if* the hypotheses underlying both the *A* and Q_r matrices were true.

In this example, a strategy for modeling OMC items with floor effects, ceiling effects, and multiple options comes into clearer focus.

- When the ability of a student is below that of the lowest available OMC option, assume that the test-taker is guessing (*floor effect*; e.g., expected response patterns of hypothetical examinees 1 and 2 for item 2.)
- When the ability of a student is above that of the highest available OMC option, assume that the test-taker will choose the highest available option. (*ceiling effect*; e.g., expected response pattern of hypothetical examinee 5 for item 3).
- When there are multiple options at a student's level, assume the student has an equal chance to pick either option (e.g., expected response patterns of hypothetical examinees 3 and 4 for item 2.)

Stage 2: Classifying Students Probabilistically into Attribute Profiles

Establishing the expected response matrix marks the culmination of the first stage of the AHM approach. In the second stage, one must establish the criteria that will be used to classify students into learning progression levels on the basis of their observed item response patterns. The purpose here is to facilitate the probabilistic mapping of observed item responses to the expected responses of students at each level of the ESS learning progression. A starting point is to simulate item responses for imaginary students at each level of the learning progression (i.e., with each possible combination of attributes). We simulate data under the constraint that the learning progression is true for comparison with item responses from real students who are, of course, unlikely to provide responses that perfectly match our initial hypothesis.

To illustrate the process of simulating such a dataset, suppose we wished to simulate item responses for N students, uniformly distributed across the five levels of the ESS progression. (Note that no assumption is being made here that students in actual school settings would be uniformly distributed across all five levels—the point is to characterize all possible item response patterns that could, in theory, be observed.) The item responses that would be expected for students at each level of the learning progression were illustrated previously using an excerpt for an expected response matrix associated with items 2 and 3 (Figure 5). We return to this example for the context of simulating item response vectors, where by a "vector" we mean a sequence of item responses. For the test as a whole, each vector would consist of a sequence of option choices for 12 items; for the example here the sequence is only two items long.

For item 2 ("Which is the best explanation for why we experience different seasons on earth?"), we would expect students who are at level 5 of the learning progression to choose answer D, which is a level 5 response. For item 3 ("Which best describes the movement of the Earth, Sun and Moon"), we would expect the level 5 students to select the highest level option (B). So for these two items, the response vector we would expect for all students at level 5 would be DB[5], and we would simulate this response pattern for N/5 students in our dataset. For students at level 4, things become a bit more complicated. For item 3 there is only one response option at level 4 (B), but for item 2 there are *two* possible response options at level 4 (A and E). It follows that there are two equally plausible response vectors: AB or EB. Each vector would need to be simulated for half of the N/5 students generated to be at level 4 in the dataset. Now consider students at level 3. On both items 2 and 3 there are two possible response options at level 3. This means that four response vectors would be equally plausible: BC, BE, CC, CE. Each vector would be simulated for one quarter of the N/5 students generated to be at level 3 in the dataset. Finally, for students at levels 1 and 2, there are no response options available at these levels for item 2. On the other hand, for item 3 there is one associated response option per learning progression level (option A is level 2; option D is level 1). In simulating item responses for these students, we can assume that when their level of understanding is below the available response options, they will guess. Hence, for both level 2 and 1 students there will be five plausible response vectors: AA, BA, CA, DA, EA for level 2, and AD, BD, CD, DD, ED for level 1. Each of these vectors would be simulated for one fifth of the N/5 students generated to be at levels 2 and 1 in the dataset. The simulated data set that would result from this process is summarized in Figure 6.

[5] To make this presentation easier to follow, we have simplified matters by expressing the response to each OMC item in terms of the response choices A to E. In terms of the underlying mathematical specification of the model, the actual response vector for "DB" would be written in binary code as <[00010][01000]> as indicated in Figure 5.

Figure 6. Simulated Dataset Based on Idealized Item Responses to Items 2 and 3

Distinct Item Response Vector	Learning Progression Level (Attributes)	Simulated Sample Size	Plausible Item Response Vector	Total Score (Item 2 Level + Item 3 Level)
1	5 (A1 & A2 & A3 & A4)	N/5	DB	9
2	4 (A1 & A2 & A3)	N/10	AB	8
3	4 (A1 & A2 & A3)	N/10	EB	8
4	3 (A1 & A2)	N/20	BC	6
5	3 (A1 & A2)	N/20	BE	6
6	3 (A1 & A2)	N/20	CC	6
7	3 (A1 & A2)	N/20	CE	6
8	2 (A1)	N/25	DA	7
9	2 (A1)	N/25	AA	6
10	2 (A1)	N/25	EA	6
11	2 (A1)	N/25	BA	5
12	2 (A1)	N/25	CA	5
13	1 (None)	N/25	DD	6
14	1 (None)	N/25	AD	5
15	1 (None)	N/25	ED	5
16	1 (None)	N/25	BD	4
17	1 (None)	N/25	CD	4

This example illustrates that the simulation of distinct response vectors corresponding to each hypothesized learning progression level in the OMC context becomes more and more involved with increases to (a) the number of items, (b) the complexity of the attribute structure, (c) the number of item floor effects, and (d) the number of items with multiple options linked to the same attributes/levels. The last column of Figure 6 shows the total score that would result from adding together the scored item responses (learning progression levels) for each expected response vector. One can see from that the total score could be a potentially misleading statistic if it were to be used for diagnostic classification, as it does not necessarily provide an accurate ranking of these simulated students in terms of the learning progression levels used to generate the simulated data.

The step from simulating a dataset with deterministic item responses to using the information in this dataset as a basis for classifying the likelihood of attribute patterns associated with observed response vectors can be rather complicated, and multiple approaches have been suggested (c.f., Leighton et al., 2004; Gierl et al., 2008). While the details are outside the scope of this chapter, the basic idea can be communicated by returning to the example of the two students we encountered previously, Liz and Andrew. If we consider just items 2 and 3, then the observed response vector for Liz was AC (which would be scored as a level 4 and a level 3 response), and the observed response for Andrew was BB (which would be scored as a level 3 and a level 4 response). Neither of these response vectors are among those that would be expected if the attribute hierarchy were true. If both students were actually at level 3 of the learning progression (i.e., they had mastered attributed A1 and A2 but not A3 and A4), then each student will have chosen one answer that constitutes an “error” (according to the model) in a positive direction. If

both students were actually at level 4 (i.e., they had mastered attributed A1, A2 and A3 but not A4) then each student will have chosen one answer that constitutes error in a negative direction. To find out which scenario is most plausible, more information would be needed about the overall probabilities that students will "slip" (give a response that is lower than expected) or "guess" (give a response that is higher than expected); this information would come from analyzing response patterns for the full sample of student respondents.

If, after comparing expected and observed response vectors element by element across all items and students, we were to find few instances in which there was a match between expected and observed responses, this would provide evidence against the hypothesized attribute hierarchy, which in turns raises questions about the validity of the learning progression. To evaluate this possibility, one can compute a Hierarchical Classification Index (HCI). The HCI takes on values between -1 and 1, and according to simulation work by Cui & Leighton (2009), values above 0.7 are interpreted as an indication of acceptable fit. When misfit is found, then one must either revise the attribute hierarchy, the hypothesized relationship between items and this hierarchy, the items themselves, or all of the above.

Imagine now that we have computed the HCI for these OMC items and have convinced ourselves that, on the basis of the data that has been gathered, the hypothesized learning progression is at least tenable. How would the results from applying the AHM be used to facilitate diagnostic inferences about Andrew's understanding of ESS? An example of a diagnostic profile display that might be provided to Andrew's teacher is shown in Figure 7. This display indicates the probability that Andrew has each of the attributes that comprise the ESS learning progression. From this the teacher might conclude that Andrew is likely to have attributes A1 and A2 and, therefore, generally thinking about ESS with a level 3 understanding according the learning progression. There is some evidence that Andrew is starting to coordinate the apparent and actual motion of objects in the sky (attribute A3), but he is not yet doing so consistently. Note that this conclusion is unlikely to differ substantially from the one that would be reached through a more subjective visual inspection of Andrew's response pattern (Figure 3)—this should come as no surprise. The point of the model in this context is not to replace teacher judgment, but to complement it. If the probability profile was greatly at odds with the judgment a teacher would have reached through careful inspection of the observed responses, this would be a reason for concern.

Figure 7. Andrew's Attribute Profile for ESS Learning Progression

Probability	A1	A2	A3	A4
1.00				
.90	***			
.80				
.70		***		
.60				
.50				
.40			***	
.30				
.20				
.10				
.00				***

A1: Student recognizes that there is some systematic nature to objects in the sky.
A2: Student knows that the Earth orbits the Sun, the Moon orbits the Earth, and the Earth rotates on its axis.
A3: Student is able to coordinate apparent and actual motion of objects in sky.
A4: 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, phases of the Moon, the seasons.

Discussion

In this chapter we have illustrated a novel method for the psychometric modeling of OMC items. At heart, building any psychometric model is about comparing observed and expected student responses. The process of delineating what is expected forces the developer of a learning progression to make some formal commitments about the actual appearance of more or less sophisticated expressions of conceptual understanding. In this chapter we have described how this process might unfold when applying a specific DCM, the AHM.

A strength of the AHM is that it requires the developer of a learning progression to be very explicit about the specific pieces of student understanding—the "attributes"—that are changing as a student progresses from naïve to sophisticated levels of understanding. This essentially involves breaking down level descriptors into what amounts to a sequence of binary codes, the combinations of which define movement from one level to the next. This process is followed to generate a Q_r matrix, which formally maps assessment items to the specific attributes students are expected to possess in order to answer each item correctly. Use of the AHM approach focuses attention on the link between hypothesized levels of a learning progression and the corresponding expectations for item response *patterns*.

In our present application involving the OMC format, we noted the challenges presented by floor and ceiling effects and multiple response options linked to the same learning progression level. In Briggs et al. (2006), two different IRT-based approaches were suggested for the psychometric modeling of OMC items: the Ordered Partition Model (Wilson, 1992) and the Multiple-Choice Model (Thissen & Steinberg, 1997). It would still be possible to take one of these approaches in stage 2 of the AHM after simulating a sample of expected response vectors under the preliminary assumption that the specified attribute hierarchy is correct. In such a scenario, the AHM could be viewed as a complement to an IRT-based approach. However, doing so negates one of the

motivations we noted earlier for applying a DCM—that it makes no assumption of continuity for the construct of measurement. A different tact that could be taken during stage 2 of the AHM modeling approach would be to view the activity as one of "pattern matching" and invoke a neural network approach or Tatsuoka's Rule Space Method to classify students into learning progression levels. In this case, the AHM would constitute a genuine alternative to an IRT-based approach.

Because all DCMs (of which the AHM is a specific instance) take a confirmatory modeling approach, much hinges upon the ability to evaluate model fit. Though considerable progress has been made over the past few years, indices of model fit (e.g., the HCI) and their interpretation are not yet well-established for DCMs. When a DCM produces output suggesting low probabilities for a student being classified in any level of a learning progression, this raises important question about fit of the student to the model and vice-versa. It will be through qualitative investigation of these discrepancies that progress can be made in our understanding of how students actually learn about scientific phenomena. Beyond internal evaluations of model fit, another alternative is to compare student classifications that would result under a more exploratory model specification. For example, Steedle & Shavelson (2009) demonstrate a case in which an exploratory modeling approach that did not begin with an a priori learning progression hypothesis (i.e. an exploratory latent class model) resulted in diagnostic classifications with substantively different interpretations about what students appeared to know and be able to do when compared to a more confirmatory diagnostic model akin to the AHM.

A potential weakness of taking a DCM-based approach is that this class of models is intended for applications in which there is a desire for very fine-grained diagnoses, where the attributes involved can be very precisely specified as "present" or "absent." It is unclear whether such fine-grain specification is possible (or even desirable) for some of the learning progressions under development in science education. In general, the more qualitative and holistic the learning progression, the less amendable it is likely to be to a DCM-based approach. For example, we have found that the force and motion learning progression (Alonzo & Steedle, 2009) is much harder to map using the AHM than was the ESS learning progression described in this chapter. Taking an IRT-based approach is sometimes viewed as a solution to this problem because it is thought to provide for inferences at a larger grain size (since constructs are typically specified in terms of multiple attributes), but it may be harder to defend the diagnoses that result after the θ continuum has been chopped into pieces through a process that may or may not follow from any substantive theory (i.e., standard-setting panels).

Regardless of the approach that is chosen for the psychometric modeling of an assessment item format such as OMC, the approach to be taken will need to satisfy at least two criteria. First, the approach must facilitate diagnostic classifications along an underlying learning progression. The classification should have formative usefulness for classroom instruction. Second, the approach must enable the developer of a learning progression to evaluate whether the initial hypothesis of the learning progression, and its instantiation using assessment items, can be supported empirically. Hence, a program of study around the use of a DCM to model OMC items would require at least two distinct strands: one that hinges upon the technical quality of the model being specified (in part through simulation work) and another that hinges upon an examination of the extent to which stakeholders (i.e., teachers) make use of the diagnostic information that emerges

from the model. It will be the evidence that emerges from these two strands of research that moves the concept of learning progressions from an interesting to a validated idea.

References

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*, 389-421.

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

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith, Jr. & R. S. Smith (Eds.), *Introduction to Rasch measurement* (pp. 143-166). Maple Grove, MN: JAM Press.

Atwood, R. K., & Atwood, V. A. (1996). Preservice elementary teachers' conceptions of the causes of seasons. *Journal of Research in Science Teaching, 33,* 553–563.

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: Lawrence Erlbaum Associates.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novicks (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bisard, W. J., Aron, R. H., Francek, M. A., & Nelson, B. D. (1994). Assessing selected physical science and earth science misconceptions of middle school through university preservice teachers: Breaking the science "misconception cycle." *Journal of College Science Teaching, 24,* 38–42.

Bock, R. D. (1997) A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*(4), 21-32.

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

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 429-449.

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

Dickinson, V. L., Flick, N. B., & Lederman, N. G. (2000). *Student and teacher conceptions about astronomy: Influences on changes in their ideas*. (ERIC Document Reproduction Service No. 442652)

Fischer, G. H. (1977). Linear logistic trait models: Theory and application. In H. Spada & W. F. Kampf (Eds.), *Structural Models of Thinking and Learning* (pp. 203-225). Bern: Huber.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychomerika*, *46*, 59-77.

Furuness, L. B., & Cohen, M. R. (1989, April). *Children's conceptions of the seasons: A comparison of three interview techniques*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco. (ERIC Document Reproduction Service No. ED306103)

Gierl, M., Wang, C., & Zhou, J. (2008) Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *The*

Journal of Technology, Learning and Assessment, 6(6). Retrieved from http://www.jtla.org

Jones, B. L., Lynch, P. P., & Reesink, C. (1987). Children's conceptions of the Earth, Sun, and Moon. *International Journal of Science Education, 9,* 43–53.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

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.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge, UK: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, 41(3), 205-237.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, *6*(1), 7–24.

Minstrell, J. (n.d.). *Facets of students' thinking*. Retrieved from http://depts.washington.edu/huntlab/diagnoser/facetcode.html

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, H. Niedderе (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 110-128). Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften.

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

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different persons employ different solution strategies. *Psychometrika, 55*, 195-215.

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

Newman, D., Morrison, D., & Torzs, F. (1993). The conflict between teaching and scientific sense-making: The case of a curriculum on seasonal change. *Interactive Learning Environments, 3,* 1–16.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

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.

Rupp, A., & Templin, J. (2008). Unique characteristics of cognitive diagnosis models: A comprehensive review of the state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*, 219-262.

Rupp, A., Templin, J. & Henson, R. (2010). *Diagnostic measurement: theory, methods, and applications*. New York: The Guilford Press.

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* (pp. 422–425). Ithaca, NY: Cornell University.

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.

Samarapungavan, A., Vosniadou, S., & Brewer, W. F. (1996). Mental models of the Earth, Sun, and Moon: Indian children's cosmologies. *Cognitive Development, 11,* 491–521.

Schneps, M. H., & Sadler, P. M. (1987). *A Private Universe* [DVD]. Harvard-Smithsonian Center for Astrophysics. (Available from Annenberg/CPB, http://www.learner.org/resources/series28.html)

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 Perspective, 4,* 1-98.

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.

Steedle, J.T., & Shavelson, R.J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46,* 699-715.

Summers, M., & Mant, J. (1995). A survey of British primary school teachers' understanding of the Earth's place in the universe. *Educational Research, 27*(1), 3–19.

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* (pp. 499–511). Ithaca, NY: Cornell University.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the Rule Space Method*. New York: Routledge.

Thissen, D. & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden and R. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 51-65). New York: Springer-Verlag.

Thissen, D., & Wainer, H. (Eds.). (2001) *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

van der Linden, W. J., & Hambleton, R. (Eds.) (1997). *Handbook of modern Item Response Theory*. 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: Lawrence Erlbaum Associates, Inc.

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

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.
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: Lawrence Erlbaum Associates.
Wilson, M. (2009). Measuring progressions: assessment structures underlying a learning progression. *Journal of Research on Science Teaching*, *46*, 716-730.
Wright, B. D (1997) A history of social science measurement. *Educational Measurement: Issues and Practice*, *16* (4), 33-45.
Zeilik, M., Schau, C., & Mattern, N. (1998). Misconceptions and their change in university-level astronomy courses. *The Physics Teacher*, *36*, 104–107.

Figure 1: A Learning Progression for Student Understanding of Earth in the Solar System

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 2. OMC Items Associated with ESS Learning Progression

2) Which is the best explanation for why we experience different seasons (winter, summer, etc) on Earth?

		Level
A.	The Earth's orbit around the Sun makes us closer to the Sun in the summer and farther away in the winter.	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.	3
C.	The Earth's rotation on its axis makes us face the Sun in the summer and away from the Sun in the winter.	3
D.	The Earth's tilt causes the Sun to shine more directly in the summer than in the winter.	5
E.	The Earth's tilt makes us closer to the Sun in the summer than in the winter.	4

3) Which best describes the movement of the Earth, Sun, and Moon?

		Level
A.	The Sun and Moon both orbit the Earth; the Earth rotates on its axis.	2
B.	The Moon orbits the Earth; the Earth orbits the Sun; the Earth rotates on its axis.	4
C.	The Moon orbits the Earth; the Earth orbits the Sun.	3
D.	The Earth, Sun, and Moon do not move, but other objects in the sky orbit around them.	1
E.	The Earth rotates on its axis.	3