USING POLYTOMOUS ITEM RESPONSE THEORY MODELS TO VALIDATE LEARNING PROGRESSIONS

by

Rajendra Chattergoon

B.A., Rutgers University, 2007

M.S.Ed., City University of New York, 2011

A thesis submitted to the

Faculty of the Graduate School of the University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

School of Education

Department of Research and Evaluation Methodology

2020

Committee Members:

Derek C. Briggs, Ph.D.

Lorrie A. Shepard, Ph.D.

Erin M. Furtak, Ph.D.

Ben R. Shear, Ph.D.

Kathy Perkins, Ph.D.

Jim Minstrell, Ph.D.

Chattergoon, Rajendra (PhD., Research and Evaluation Methodology School of Education)
Using Polytomous Item Response Theory Models to Validate Learning Progressions
Thesis directed by Professor Derek C. Briggs

Learning progressions (LPs) are "descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (National Research Council, 2007). One challenge that arises in LP research is the collection of evidence to ensure that ordered levels of an LP are an adequate representation of how learning occurs for all students. These LP validation studies involve the identification, accumulation, and interpretation of evidence collected using assessment tasks to make claims about the development of student thinking hypothesized by the LP.

This dissertation develops a novel method using item response theory (IRT) to evaluate the order of LP levels with items that have answer choices mapped to the levels of an LP. I use two IRT models – the nominal response model (NRM; Bock, 1972) and the partial credit model (PCM; Masters, 1982). The NRM is used to evaluate and revise partial credit scoring schemes. The PCM is used to explore whether a quantitative scale can be developed that maps onto the LP hypothesis. Empirical data from the Diagnoser assessment system (Thissen-Roe, Hunt, & Minstrell, 2004) are analyzed to illustrate how to apply this method with data collected from items with responses mapped to levels in a model of student thinking.

The findings illustrate how analyzing patterns among NRM response option curves relative to the content of the items reveal how partial credit could be assigned to the answer choices and how analysis of a PCM item-person map can be used to evaluate the quality of a latent ability scale. These results also uncover additional areas for investigation, including more research on the psychometric characteristics of items with answer choices mapped to models of

student thinking. The use of the NRM and PCM advanced in this dissertation provides a new method of providing psychometric information that can be used for LP validation studies.

Acknowledgements

This dissertation would not be possible without the support of my academic mentors. All of the members of my committee have been extremely supportive throughout this process, contributing invaluable feedback to shape my research. Jim generously shared the Diagnoser data that I analyze in this study. Lorrie and Erin both encouraged me to use my chemistry teaching experience and content expertise to interpret the data in this dissertation. Ben introduced me to the NRM and suggested it might be useful given my interests. Kathy encouraged me to think about alternative ways of describing learning. Derek read countless drafts, provided detailed feedback, and carved out time to help me understand foundational concepts better and craft a defensible investigation. His mentorship throughout my graduate career has been invaluable, and my approach to research is inspired by his guidance. Most importantly, Derek has been incredibly patient, kind, and understanding throughout this lengthy process, while never compromising on his expectations for quality. For that, I am immensely grateful.

I am also fortunate to have had additional mentors and teachers to explore the ideas in this dissertation in practical settings. Scott Marion's mentorship during a summer internship with the National Center for the Improvement of Educational Assessment helped me appreciate how assessment is used in the real world. Maddy Keehner and Gabrielle Cayton-Hodges from the Educational Testing Service helped deepen my understanding of the relationship between cognitive models and assessments. Elena Diaz-Bilello supervised my work as a graduate research assistant for years, always providing me rich, authentic opportunities to conduct studies with districts and states. Her supervision always sought to empower me and build my confidence in my abilities. I am incredibly grateful to all of my professors in the School of Education, who

gave me the solid theoretical and technical foundation to make this dissertation possible. I am also grateful to my past and contemporary fellow graduate students – Jessica Alzen, Michael Turner, and Amy Burkhart – who allowed me to share in their dissertation writing experiences so that I could understand what this process really looks like. I'd also like to thank Gillian Bayne, who first suggested that I should pursue a Ph.D., and Panorea Panagiosoulis and Althea Hoard who gave me first teaching job in New York City.

Finally, I would not have completed this dissertation without the constant support and love from my immediate and extended family. Though they viewed graduate school with a layer of mystery, my parents were always adamant that I finish. The support from my grandparents, numerous aunts, uncles, and cousins has been constant. Denise McCoy and Kieran Chattergoon, my beloved wife and son, gave me the time and space to complete this degree, while ensuring that I had numerous opportunities to spend quality time with them. Without their enduring support, success in graduate school would not have been possible.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

In the last two decades, there has been an explosion of research that attempts to integrate the learning sciences with educational assessment. One result has been the emergence of a research program to create, validate, and apply learning progressions[1] to a variety of content domains, especially those related to science and math. In the report *Taking Science to School*, the National Research Council (NRC, 2007) defined learning progressions (LPs) as "descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (p. 214). Subsequent researchers have expanded on this preliminary definition,[2] but all agree that using LPs to represent student knowledge requires a hypothesis of how student thinking develops along an ordered continuum. Adopting a developmental perspective on learning is not new. However, the coupling of assessment design with a theory of how students learn represents a departure from alternative test development approaches that designed tasks to determine whether or not students mastered discrete sets of knowledge and skills.

One challenge that arises in LP research is the collection of evidence to ensure that the ordered levels of an LP are an adequate representation of how learning occurs for all students. The levels of an LP can be viewed as an ordered set of stages of achievement that students may

[1] Learning progressions are also called learning trajectories, especially in the domain of math.

[2] For example, Duncan and Hmelo-Silver (2009) in the editorial for a special issue of the *Journal of Research in Science Teaching* defined LPs in terms of four key theoretical and structural characteristics: 1) LPs are focused on a few foundational and generative disciplinary ideas and practices; 2) LPs are bounded by an upper anchor describing what students are expected to know and be able to do by the end of the progression and by a lower anchor describing the developers' assumptions about the prior knowledge and skills of learners as they enter the progression; 3) they describe varying levels of achievement as the intermediate steps between the two anchors; and, 4) LPs are mediated by targeted instruction and curriculum.

demonstrate along the pathway of mastering a learning goal (Duncan & Hmelo-Silver, 2009). However, prior research has found that students often reason differently, particularly in the "messy middle" of LPs designed to measure complex learning goals (Gotwals & Songer, 2009). There is usually a tension in LP research between evidence to support a coarse order of student ideas that can be used to describe learning for the majority of students and evidence that reveals the rich diversity of student thinking. Both kinds of evidence must be synthesized to improve interpretations about what students know and can do relative to an LP. Evidence to describe the order of learning becomes easier to collect and analyze at scale when assessments are designed to embed information about stages of student thinking into plausible response options. Despite the potential of LPs to improve educational assessment, the analysis of tasks and items designed using LPs is a relatively new area of applied psychometrics. The purpose of this dissertation to develop a method to evaluate the order of LP levels using items with answer choices mapped to the levels and illustrate how this method can be used to interpret student response data.

1.1. Learning Progression Validation.

LP research is highly interdisciplinary. LP researchers frequently combine findings from the learning sciences with principles from educational measurement to describe how LPs represent student learning. Gotwals and Alonzo (2012) organize most LP research into four overlapping strands: defining LPs, developing assessments to elicit student responses relative to an LP, modeling and interpreting student performance relative to an LP, and using LPs. Research programs have emerged to create LPs, develop and use assessments to elicit evidence about student thinking, and interpret data relative to the LP hypothesis. Studies that interpret empirical

data for the purpose of evaluating an LP hypothesis are often called LP "validation" studies (Corcoran, Mosher, & Rogat, 2009). These studies use empirical data to explore whether the LP is a sufficiently accurate representation of how student thinking develops.

LP validation is closely related to the concept of test validation. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) distinguish between the validity of a test and the process of validation:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for proposed uses that are evaluated, not the test itself (p. 11).

This perspective views test validation as a "process" that requires the accumulation of evidence to support the interpretation of test scores for specific uses. The validation of tests designed using LPs is an ongoing process of accumulating evidence to support the meaning of test scores relative to the LP hypothesis.

In validation efforts, however, there is often a distinction between prioritizing the *interpretation* of test scores versus prioritizing the *uses* of the tests. A common, pragmatic approach in operational test development is to specify a logical interpretative argument that details the inferences leading from the test scores to the conclusions to be drawn and any decisions to be based on those conclusions (Kane, 1992). An alternative perspective argues that the use of tests in real-world contexts, especially in classrooms, should be the primary focus of validation efforts (Brookhart, 2004; Moss, 2003) and that validation efforts should focus on the

identification and mitigation of systematic errors in the interpretation of test scores (Chatterji, 2003). The latter perspective recognizes the practical reality that tests may be used by students and teachers in multiple, complex ways that impact future learning choices, self-worth, and a host of other cognitive and affective understandings about identity (Stiggins, 1999). Although these two perspectives suggest disagreements over the best way to engage with the process of validation relative to educational testing (Newton & Shaw, 2015), these competing prioritizations are not necessarily incompatible (Shepard, 2016). The more meaningful and defensible the numbers associated with test scores are, the more appropriately practitioners can use information from the assessments to support student learning in classrooms.

LPs can be used to imbue greater meaning into test scores (Shepard, 2018) and can therefore improve interpretative arguments for classroom assessments. Gorin (2006) has recommended incorporating information about student thinking into the test development process during the design of educational assessments rather than waiting until after data has been collected as a way to improve the information test scores provide about student ability. One of the strongest examples of the integration of LPs with assessment design is the ordered multiple-choice (OMC) item format developed by Briggs, Alonzo, Schwab, and Wilson (2006). OMC items resemble traditional multiple-choice items, but the answer choices are all mapped to levels of an LP. Although OMC items are some of the most-well known examples of how LPs can inform assessment design, there are many outstanding questions about how to interpret data collected using these innovative tasks relative to the levels of an LP.

1.2. Methods of Validating Learning Progressions.

LP validation involves the identification, accumulation, and interpretation of evidence collected using assessment tasks to make claims about the development of student thinking hypothesized by the LP. Keehner and colleagues (2016) distinguish between external or "offline" sources of empirical evidence and internal or "online" sources. Offline sources of evidence are collected outside the task, either temporally or physically, and they include cognitive-psychometric modeling of student response data and experimental manipulation of task characteristics. Online sources are collected while the respondent is completing the task, and they include methods that produce verbal data from respondents (e.g., playtesting, cognitive interviews, or think-alouds) and methods that collect fine-grained behavior from respondents (e.g., clickstream data or eye tracking). The accumulation of robust offline and online data is needed for strong validity claims. However, online validity investigations or experimental studies are often costly, time-consuming, and yield data that can be difficult to interpret. For these reasons, cognitive-psychometric modeling is often the preferred method used by researchers to provide validity evidence for LPs, especially in the initial stages of LP research.

Cognitive-psychometric modeling involves the incorporation of cognitive features directly into the mathematical formulation of the test score. This can be done through the modeling of item response data scored using partial credit scoring schemes (Andrich, 1978; Masters, 1982; Muraki, 1992; Samejima, 1969), diagnostic classification modeling using item-by-skill matrices (Rupp, Templin, & Henson, 2010), or the use of complex explanatory models that incorporate cognitive attributes into the calculation of probabilities of a correct response to an item (De Boeck & Wilson, 2004). A wide variety of psychometric models have been

developed for these purposes, but there is comparatively little research on how these models can be used to improve interpretations of what students know and can do relative to LPs using data collected from OMC items.

When the purpose of psychometric modeling is to collect and analyze evidence for the validity of an LP, it is helpful to select psychometric models that permit items and respondents to be ordered along a latent ability continuum and then mapped to the structure of the LP. A common psychometric framework used to provide validity evidence for an LP is item response theory (IRT). Developed separately by Birnbaum (1968), Lord (1952), and Rasch (1960), IRT models the probability of an item response as a nonlinear function of person ability and item characteristics. If item difficulties and student performance order themselves in ways that are consistent with the predictions implied by an LP, this is often taken as confirmatory evidence for the validity of the LP (Corcoran et al., 2009). There are a wide range of IRT models, and researchers have typically chosen models based on either tradition (e.g., Rasch models) or to match the complexity of the student response data (e.g., multidimensional or explanatory IRT models). As described in Chapter 3, the results of cognitive-psychometric modeling of OMC items may reveal that LP levels overlap or are highly related (e.g., Liu & Lesniak, 2005), the LP hypothesis needs to be revised to better account for instructional variables like content standards or teachers' curricula (e.g., Todd & Kenyon, 2016), or that there is little evidence to suggest that student ideas can be ordered at all (Steedle, 2008). While these investigations provide the field with tenuous, research-based, hypotheses for developmental constructs, the field lacks a general method to explore the order of levels in an LP when data has been collected using OMC items.

This dissertation contributes to the literature on LP validation by developing an IRT-based method to interpret the order of LP levels using student response data collected from OMC

items. The evidence produced from the application of this method can be used in LP validation studies to support or revise initial hypotheses about the LP. Before assessments linked to LPs can be used as part of large-scale assessment systems designed to measure new and challenging standards like the *Next Generation Science Standards* (NRC, 2014), important questions about whether LPs and the accompanying assessments adequately represent the development of student thinking must first be answered. This is the overarching goal of many studies that seek to validate LPs, and this dissertation provides an empirical illustration of an analytic approach that can help advance this research endeavor.

1.3. Overview of Dissertation.

This study is divided into six chapters in addition to this introduction. Chapter 2 provides conceptual background on the three components of educational assessments: a model of student thinking (e.g., an LP), tasks designed to collect observations about student thinking (e.g., OMC items), and methods to interpret response data relative to the model of student thinking (e.g., IRT). Chapter 3 situates this dissertation study within the broader research literature by reviewing prior research on the cognitive-psychometric modeling of OMC items. Chapter 4 presents a new method to explore the order of levels in an LP using two IRT models – the nominal response model (Bock, 1972) and the partial credit model (Masters, 1982). Chapter 5 introduces the empirical data used in this study. Chapter 6 illustrates how the IRT models can be used to interpret the empirical data. Chapter 7 concludes with a discussion of the results, identifies limitations of this study, and suggests areas for future research.

Chapter 2

Background

The purpose of this chapter is to present an overview of the conceptual foundations of educational assessment by describing how LPs can be adapted to design OMC items. I organize this chapter according to the three pillars of the assessment triangle (NRC, 2001), which is a conceptual tool that describes the elements that should inform the design of an educational assessment (see Figure 2.1). The first pillar is a model of how student thinking develops in a content domain (green box). These "little" models are created using a "big" theory of how learning occurs (orange box), and they can be used to design observations (blue box), which are the second pillar of the assessment triangle. Observations are the kinds of tasks that will prompt students to say, do, or create something that demonstrates knowledge and skills linked to the model of student thinking. The third pillar consists of methods of interpretation (purple box) that connect observations to a model of student thinking, often through the use of a statistical model.

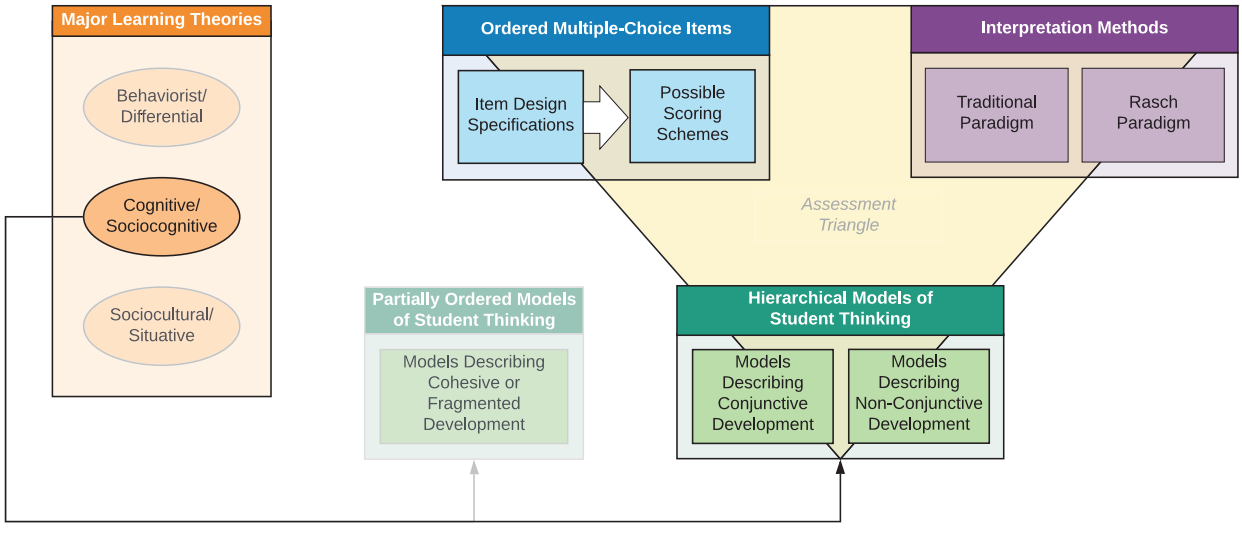


Figure 2.1. Using the assessment triangle for LP validation.

In the remainder of this chapter, I describe each of the components in Figure 2.1 in detail. I begin by situating LPs within the major, big theories of learning. Next, I distinguish between partially ordered and hierarchical little models of student thinking and locate LPs within these types of models. I then describe how hierarchical models of student thinking can be used to design and score OMC items. I conclude by introducing two general perspectives on psychological measurement and discuss how the use of psychometric models from both of these traditions can contribute information to test the hypothesized order of levels in an LP.

2.1. Connecting Big and Little Learning Theories.

The most authoritative sources of knowledge about how people learn (NRC, 2000; National Academies of Science, Engineering, and Medicine [NASEM], 2018) identify three broad, historical perspectives on learning. The earliest modern conception of learning comes from the behaviorist psychologists of the early twentieth century (e.g., Thorndike, 1931). Behaviorists conceived of learning as the study of observable behaviors and the stimulus conditions that control them. By mid-century, the new field of cognitive science conceptualized learning as an active process of mental construction. Early cognitive theorists like Piaget (1936) sought to identify and understand the individual components of mental function and to test theories through empirical investigations. In recent years, qualitative and quantitative research across multiple disciplines by researchers like Vygotsky[3] (1978) demonstrated that learning happens within complex developmental, cognitive, physical, social, and cultural systems. This

[3] Vygotsky was a contemporary of Piaget. However, his work was not widely known until after his death in 1934. In contrast, Piaget influenced numerous students and collaborators during his lifetime.

sociocultural perspective shifts the unit of analysis away from an individual learner and instead onto the process of learning within a community.

These three big learning theories often appear in the literature by different names to emphasize different features. For example, Greeno, Pearson, and Schoenfeld (1996) identified four "perspectives" on the nature of the human mind: differential, behaviorist, cognitive, and situative. In describing the relationship between a theory of learning and a model of student thinking, the authors of *Knowing What Student Know* (NRC, 2001) used Greeno and colleagues' perspectives to illustrate how the selection of a big learning theory has implications for what should be assessed and how the assessment process should occur. Shepard, Penuel, and Pellegrino (2018) used this example to clarify that little learning theories sit within big theories of learning. Penuel and Shepard (2016) have argued that the big sociocognitive and sociocultural learning theories are the most compelling theories to design coherent educational experiences for students through curriculum, instruction, assessment, and teacher professional development.

LPs are an example of a fine-grained little learning theory most clearly compatible with the sociocognitive perspective on learning. Sociocognitive models of learning attend to the social nature of learning and to the discipline-specific ways that core ideas and practices are developed over time (Shepard et al., 2018). To be used in ways more compatible with the sociocognitive (Shepard, 2018) or sociocultural (Lehrer & Schauble, 2015) perspectives on learning, LPs should be thought of as tools that require further adaptation. LPs can be adapted in local instructional contexts to improve teaching and learning, or they can be adapted by assessment developers to design tests that can potentially provide better information about what students may know.

Other LP adaptations have been described in the literature. Duschl, Maeng, and Sezen (2011) distinguish between "evolutionary" and "validation" LPs. Evolutionary LPs identify mid-

levels or milestones that can be used to design instructional interventions. Validation LPs are those that seek to "validate" the initial sequences and levels of a progression, and they are most useful to develop and test models of student thinking for the purpose of creating assessments. Shavelson and Kurpius (2012) distinguish between "curriculum and instruction" LPs, which are similar to evolutionary LPs, and "cognition and instruction" LPs. Cognition and instruction LPs involve the psychological analysis of cognition related to subject-matter learning, and they attempt to provide a valid and practically useful way of portraying the pathway of cognitive development for the purposes of instruction. These distinctions should be thought of as heuristics, since validation LPs are often developed from research into how student thinking develops in classrooms and evolutionary LPs are frequently based on some empirical evidence about how student ideas can be ordered.

This study adopts the perspective that little models of student thinking, like LPs, can be used to design higher-quality assessments that enable the collection of better observations about what students know and can do. To be used in ways compatible with sociocognitive learning theories, assessments designed using LPs should be subject to an ongoing process of validation that involves the accumulation of evidence that an LP is an appropriate representation of how student thinking develops during learning (Corcoran et al., 2009). This evidence may come from multiple sources, including qualitative studies of teaching and learning in classrooms or quantitative studies that attempt to evaluate the characteristics of students and assessment tasks along a scale. This dissertation is an example of the latter study because it explores how assessment items can be analyzed to explore hypotheses about the order of levels. The next section locates LPs within little models of student thinking.

2.2. Using Models of Student Thinking to Describe the Development of Student Ideas.

The report *Knowing What Students Know* (NRC, 2001) strongly recommended that educational assessment be coupled with developmental models of student thinking. Drawing on the extensive research on learning summarized in *How People Learn* (NRC, 2000), it defined a "model of cognition" as "a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain (e.g., fractions)" (p. 44). I use the term "model of **student thinking**" rather than "model of **cognition**" to underscore that these little learning theories are connected to the big theories of learning. That is, these little models can include elements that are consistent with behaviorist, sociocognitive, or sociocultural learning theories.

In the years following the publication of *Knowing What Students Know*, assessment developers began using models of student thinking to attempt to improve inferences about what students know. As illustrated in Figure 2.2, models of student thinking can be classified as partially ordered or hierarchical depending on how student ideas are represented relative to one another. Partially ordered models of student thinking consist of taxonomies of disciplinary ideas that students may possess. Ideas are ordered along a dimension of disciplinary correctness, often with the implicit assumption that "incorrect" ideas, or misconceptions, may have differing degrees of sophistication. These models are created primarily from research on "conceptual change," which is the process learners undergo as they restructure prior knowledge to understand advanced concepts in the various disciplines (Vosniadou, 2007). Partially ordered models can be useful to surface the diversity of ideas that students may bring with them into the classroom. Hierarchical models of student thinking represent sequences of increasingly more sophisticated ideas that appear as students engage with the process of learning. These models couple research

from the conceptual change literature with research on how learning occurs in classrooms. Although an individual's process of learning may not be linear nor follow a pre-determined pathway, hierarchical models of student thinking attempt to identify sequences of productive stages of learning that can be adapted for the purposes of classroom instruction and assessment. Both types of models describe how students develop understanding of a disciplinary idea.

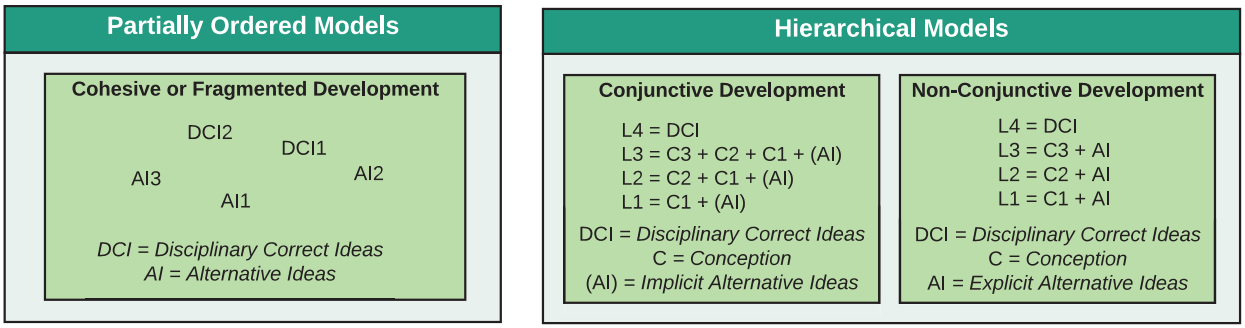


Figure 2.2. Types of models of student thinking.

The two types of models of student thinking depicted in Figure 2.2 conceptualize the development of student thinking differently. Partially ordered models of student thinking describe how student thinking develops from alternative ideas to disciplinary correct ones. Alternative ideas are generally not ordered in terms of sophistication, and the goal of assessment is to surface these ideas so they can be addressed directly during instruction. Hierarchical models of student thinking describe learning as either the gradual building of ideas (conjunctive development) or as the successive replacement of less sophisticated ideas with more sophisticated ones (non-conjunctive development). In developing hierarchical models of student thinking, researchers often specify a sequence of productive "conceptions" that students may have as they learn about a disciplinary idea. Assessment relative to hierarchical models of

student thinking, like LPs, has the potential to substantively describe how much students have learned during instruction along a quantitative scale that has desirable properties (Briggs & Peck, 2015), while also providing clear information to educators about "what to do next" given a particular diagnosis of student thinking (Heritage, 2008; Penuel, 2015). Although both partially ordered and hierarchical models of student thinking can be used for assessment, partially ordered models may be more challenging to use for teaching and learning, as I describe next.

2.2.1. Partially Ordered Models of Student Thinking. Partially ordered models of student thinking emerged as researchers began applying research on conceptual change to the process of learning in schools. Early conceptual change researchers were largely interested in how an individual's knowledge is restructured as a student develops understanding of a specific concept. Borrowing from Thomas Kuhn's (1962) theory of change in science, early researchers framed the study of conceptual change as "a paradigm shift in pupils' thinking" (Driver & Easley, 1978) that required a "Kuhnian state of crisis" (Posner, Strike, Hewson, & Gertzog, 1982) in order for learning to occur. Posner and colleagues' (1982) influential description of conceptual change argued that students bring alternative frameworks, commonly called misconceptions, to learning and that these misconceptions need to be identified and eradicated during the process of learning or else they could impede the development of conceptual understanding. Subsequent researchers (e.g., Sinatra & Pintrich, 2003; Smith, diSessa, & Roschelle, 1993) critiqued Posner and colleagues' perspective on conceptual change and advocated for an alternative model that viewed students' alternative ideas as productive resources that could be leveraged during instruction (e.g., Hammer & Elby, 2003).

Criticism of Posner and colleagues' theory of conceptual change surfaced an unresolved debate in the conceptual change literature over whether or not the development of student thinking could be described as cohesive or fragmented. Proponents of the cohesive view claimed that students' knowledge is "theory-like" (e.g., Vosniadou, 1994; Vosniadou, et al., 2008), meaning that individuals' knowledge structures consist of a relatively coherent body of domain-specific concepts. Learning and development is seen as the reorganization of this theoretical mental structure. In contrast, proponents of the fragmented view (e.g., diSessa, 1988, 1993, 2008) argued that the knowledge system of learners consists of an unstructured collection of many simple elements that originate from superficial interpretations of physical reality. Learning occurs as these pieces of knowledge are collected and systematized into larger wholes.

Representations of partially ordered models of student thinking emerged as researchers began using conceptual change research to create taxonomies of ideas for the purpose of assessment. Figure 2.3 presents an example of a partially ordered model of student thinking. This figure is a condensed display of the disciplinary correct and alternative ideas associated with the Force Concept Inventory (Hestenes, Wells, and Swackhamer, 1992). The left-hand column of Figure 2.3 displays the disciplinary correct ideas associated with force and motion along with potential misconceptions. Although the letters and numbers are nominal labels, the partial ordering of ideas comes from the distinction between the bolded, "correct" ideas and the misconceptions printed in regular text below them. Partially ordered models can be used to create concept inventories, which are multiple-choice tests consisting of items where the response options correspond to one or more correct or alternative ideas. The right-hand column of Figure 2.3 indicates the number of a multiple-choice item from the Force Concept Inventory and the letter of the answer choice(s) associated with the misconception described on the left.

	Inventory Item
0. Kinematics	
Kl. position-velocity undiscriminated	208,C,D
K2. velocity-acceleration undiscriminated	20A; 21B,C
K3. nonvectorial velocity composition	7C
1. Impetus	
I1. impetus supplied by "hit"	9B,C; 22B,C,E; 29D
I2. loss/recovery of original impetus	4D; 6C,E; 24A; 26A,D,E
I3. impetus dissipation	5A,8,C; 8C; 16C,D; 23E; 27C,E; 29B
I4. gradual/delayed impetus build-up	6D; 8B,D; 24D; 29E
I5. circular impetus	4A,D; 10A
2. Active Force	
AFl. only active agents exert forces	11B; 12B; 13D; 14D; 15A,B; l8D; 22A
AF2. motion implies active force	29A
AF3. no motion implies no force	12E
AF4. velocity proportional to applied force	25A; 28A
AF5. acceleration implies increasing force	17B
AF6. force causes acceleration to terminal velocity	17A; 25D
AF7. active force wears out	25C,E
3. Action/Reaction Pairs	
AR1. greater mass implies greater force	2A,D; 11D; 13B; 14B
AR2. most active agent produces greatest force	13C; 11D; 14C
4. Concatenation of Influences	
CI1.largest force determines motion	l8A,E; 19A
CI2. force compromise determines motion	4C, 10D; 16A; 19C,D; 23C; 24C
CI3. last force to act determines motion	6A; 7B; 24B; 26C
5. Other Influences on Motion	
CF. Centrifugal force	4C,D,E; l0C,D,E
Ob. Obstacles exert no force	2C; 9A,B; 12A; 13E; 14E
Resistance	
R1. mass makes things stop	29A,8; 23A,B?
R2. motion when force overcomes resistance	28B,D
R3. resistance opposes force/impetus	28E
Gravity	
G1. air pressure-assisted gravity	9A; 12C; 17E; 18E
G2. gravity intrinsic to mass	5E; 9E; 17D
G3. heavier objects fall faster	1A; 3B,D
G4. gravity increases as objects fall	5B; 17B
G5. gravity acts after impetus wears down	5B; 16D; 23E

Figure 2.3. A taxonomy of misconceptions probed by the Force Concept Inventory.[4] (reproduced from Hestenes et al., 1992)

Since the conceptual change literature emphasizes the development of individual cognition, partially ordered models of student thinking that are developed exclusively from research on conceptual change may be limited in their relevance for instruction aligned to sociocognitive or sociocultural principles. Both learning theories recognize that students develop

[4] Hestenes and colleagues present a separate table describing the disciplinary correct ideas and answer choices corresponding to the disciplinary correct ideas. Figure 2.3 is a representation that focuses exclusively on the misconceptions probed by the Force Concept Inventory.

understanding of a disciplinary idea within an environment, and part of the goal of instruction is to build on students' existing ideas as they move towards the learning goal. For example, classroom assessment researchers like Heritage (2008) argue that for instructors to provide effective feedback to students, instructors need to have models that describe a "continuum of how learning develops in any particular knowledge domain so that they are able to locate students' current learning status and decide on pedagogical action to move students' learning forward" (p. 2). Partially ordered models of student thinking may help instructors identify the kind of alternative ideas that are present among students in a class, but these models may provide little support for deciding what to do next during instruction. A stronger ordering of alternative ideas may help instructors design activities to help students gradually increase the sophistication of their thinking, even if these students may not yet demonstrate mastery of the learning goal.

2.2.2. Hierarchical Models of Student Thinking. Alternative models of student thinking describe the attainment and gradual development of an ordered sequence of conceptions. This approach was first promoted by Jean Piaget and his students. Piaget's (1936) classic *The Origins of Intelligence in Children* described sequential stages of the development of intelligence, and Piaget was interested in how "the entire sequence of stages" could illuminate "the reality of the evolution of schemata" (p. 399). His approach emphasized structure, order, and continuity in the learning process, not necessarily linearity or the development of thinking within a specific discipline like math or science. Piaget was interested in understanding how student understanding unfolds naturally during schooling, and he did not focus on instructional interventions that could improve the practices of teaching and learning in classrooms.

The conceptual change literature emerged as researchers began exploring nuances inherent in the transitions between the stages identified by Piaget and others for the purpose of helping students develop understanding of a specific idea. The main distinctions between the Piagetian and the conceptual change mechanisms for learning are that Piaget emphasized "bottom-up, conservative, additive, and largely unconscious mechanisms" while conceptual change emphasized "top-down, radical, deliberate, and intentional learning mechanisms" (Vosniadou, 2007, p. 50). Drawing on Piaget's research, the conceptual change literature, and research on student learning in classrooms, education researchers have developed domain-specific, hierarchical models of student thinking, of which LPs are one example. Hierarchical models view learning as the conjunctive or non-conjunctive development of ideas.

Hierarchical models that describes conjunctive learning are those where ideas build on one another successively as students develop understanding of a concept. These models describe how student thinking becomes more complex by incorporating more and more sophisticated ideas. They are accompanied by an implicit understanding that students may have one or more alternative ideas at each level of the progression, but alternative ideas are less productive for the purpose of strategically moving students towards the learning goal. Examples of conjunctive models of student thinking include many of the LPs from science education (e.g., Stevens, Delgado, & Krajcik, 2010) and learning trajectories in math education (e.g., Confrey, 2012). Conjunctive models characterize learning as the additive accumulation of concepts.

Figure 2.4 displays an example of a hierarchical model of student thinking that describes conjunctive learning. Stevens and colleagues (2010) developed this LP using design-based research (Collins, Joseph, & Bielaczyc, 2004) to describe the development of student ideas about the nature of matter in classrooms. The LP in Figure 2.4 emphasizes the additive building of

connections among ideas. To understand atomic structure, for example, the LP proposes that students first develop an understanding that atoms are spherical structures in Level 1 then progress to understanding the components of atoms in Level 2, understanding models that describe the arrangement of electrons in Level 3, and finally understanding the relationship between the structure of the atom and energy in Level 4. These concepts build on another as indicated by the numbering of the levels and their vertical orientation, but there may be multiple pathways for students to progress through the states as indicated by the lines between the levels.

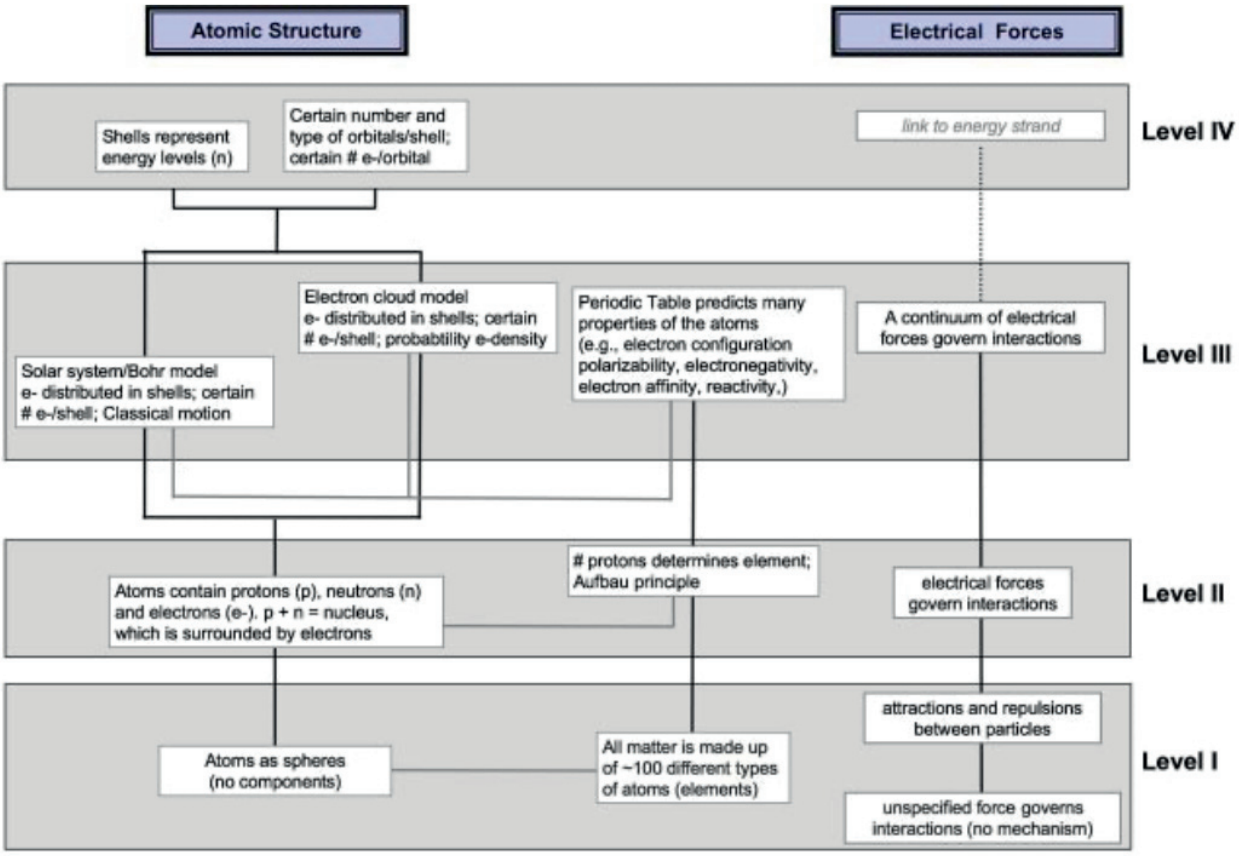


Figure 2.4. A hypothetical LP for atomic structure and inter-atomic interactions. (reproduced from Stevens et al., 2010)

An alternative hierarchical model of student thinking describes non-conjunctive learning where the developmental hypothesis proposes that more sophisticated ideas replace less sophisticated ones as learning occurs. Ideas do not necessarily build on one another. These models consist of a distinct conception at each level accompanied by one or more explicitly stated alternative ideas. Examples of non-conjunctive models of student thinking include Alonzo and Steedle's (2008) Force and Motion LP and the models of domain expertise developed by Chi and Glaser (1985). In describing the development of expertise in problem solving, Chi, Glaser, and Reese (1982) state that "experts and novices organize their knowledge in different ways" since "experts possess schemata of principles that may subsume schemata of objects" and "novices possess only schemata of objects" (p. 70). This perspective hypothesizes that less sophisticated novice ideas may be subsumed by more sophisticated expert ones.

One set of hierarchical models of student thinking that may describe non-conjunctive learning are the facet clusters developed by Jim Minstrell and his colleagues. These models were inspired by diSessa's (1988, 1993, 2008) "knowledge-in-pieces" theory of conceptual change:

> *Facets* are used to describe students' thinking as it is seen or heard in the classroom. Facets of students' thinking are individual pieces or constructions of a few pieces of knowledge and/or strategies of reasoning. While facets assumes a "knowledge in pieces" perspective like that of diSessa (1993), the pieces are generally not as small as … assumed by diSessa. ... They are convenient units of thought for characterizing and analyzing students' thinking (Minstrell, 2000, p. 47).

Facets are intended to describe student thinking as it is seen or heard in the classroom. Figure 2.5 displays an example of a facet cluster for student thinking about atoms (DeBarger, Ayala, Minstrell, Kraus, & Stanford, 2009). In contrast to the LP in Figure 2.5 that describes how

students' ideas about the nature of matter build on one another, facet clusters order ideas in terms of how problematic they are from an instructional perspective.

00 All matter is made up of atoms, which are too small to be seen even with a powerful light microscope. Atoms cannot be created or destroyed by ordinary chemical or physical means.

- 01 The student understands that all matter is made up of atoms.
- 02 The student understands that atoms are tiny (too small to see even through a light microscope).
- 03 The student understands that atoms cannot be created or destroyed by chemical reactions.

---

40 Student believes that atoms are created (or destroyed) through ordinary daily events.

- 41 When a substance is used or burned, atoms are destroyed, disappear or are turned into a form of energy.
- 42 When a new substance is created, atoms are created.

50 The student does not have an accurate sense of the scale of an atom as compared with objects they can see.

- 51 Compares the size of the atom to objects visible to the naked eye like a grain of sand, speck of dust, tip of a pin, hair, etc. …
- 52 Compares the size of the atom with objects visible with a microscope like bacteria, a virus, blood cell, etc. …

80 The student thinks that not all matter is made up of atoms.

- 81 Matter is infinitely divisible (in theory) – if we kept cutting a substance, we would always get just a smaller amount of that same substance (until it is no longer strong enough to hold together).
- 82 Living things are not made of atoms.
- 83 Atoms do not exist because they cannot be seen.

Figure 2.5. Facet cluster for student thinking about atoms. (reproduced from www.diagnoser.com)

Facet clusters suggest an order among facets to make them useful for instruction. Minstrell, Anderson, & Li (2016) indicate that the two-digit numerical codes associated with each facet roughly rank facets from least urgent (20s to 40s) to most urgent to address (80s to

90s) during instruction, and these authors carefully distinguish facet clusters from traditional LPs that specify a clear progression of ideas:

> The problematic Facets … represent the multiple and variable ideas or learning constructs that students actually demonstrate on their way to understanding the learning goal. For this reason, their ordering within the cluster is a rough (not fixed) ranking rather than a clear progression (Minstrell et al., 2016, p. 56).

The facet clusters *may* suggest non-conjunctive development since there is no presumption that facets with lower numbers build on facets with higher numbers. For example, students who have not yet demonstrated understanding of the scale of atoms but understand that atoms are the building blocks of matter (facet 50) may no longer possess the more problematic idea that matter is made of components other than atoms (facet 80). As students' ideas become less urgent to address during instruction (lower facets), very problematic alternative ideas (higher facets) may be replaced by other less problematic ideas (intermediate facets). In Figure 2.5, distinct conceptions are indicated by the labels that end with "0" and the alternative ideas associated with each level are represented by the digits that end in "1" through "9." That is, the order of facets is defined relative to the first digit of the label and not the second.

Regardless of whether a hierarchical model of student thinking presumes conjunctive or non-conjunctive development, these models provide a representation of sequential states of student thinking so that this order can be used in classrooms to support teaching, learning, and assessment. The goal of assessment relative to ordered models of student thinking is not simply to surface student ideas, like with partially ordered models of student thinking, but instead to use the information provided by assessment to directly inform instructional decisions. In the research literature, hierarchical models of student thinking are often called LPs or learning trajectories,

but there are non-LP examples of hierarchical models like Minstrell's facet clusters. In the next section, I describe how assessment developers have adapted hierarchical models of student thinking to create a specific type of task – the OMC design introduced in the previous chapter.

2.3. Developing Ordered Multiple-Choice Items.

The OMC item design developed by Briggs and colleagues (2006) is one attempt to connect the design of assessment tasks to a hierarchical model of student thinking. OMC items are multiple-choice items that have each answer choice linked to a level in a hierarchical model of student thinking. An intended advantage of OMC items is that they combine features typical of both selected-response and constructed-response items. Selected-response items require students to choose an answer from a fixed set of options. Often, students can respond to many of these tasks in a short period of time, ensuring relatively high reliability. Constructed-response items require students to provide a response to a prompt without providing options. These items may yield richer information about student thinking. However, constructed-response items can be time-consuming to score, and their reliability may be poor since fewer constructed-response items can be given at a particular time. The idea behind the OMC item design is to maintain the ease of scoring and reliability advantages of selected-response items while providing the content-rich information typically associated with open-ended tasks. However, because of the complicated OMC design described below, it can be difficult to distinguish among ordered levels when using these items even if many OMC items can be administered at the same time.

To illustrate how a hierarchical model of student thinking can be used to develop OMC items, it is helpful to distinguish between the *design* of assessment tasks to collect observations

about student thinking and the *scoring* of those observations. A hierarchical model of student thinking can be used to create item design specifications where selected-response answer choices are connected to distinct levels. The order of these levels can then be used to develop partial credit scoring schemes that assign ordinal scores to the responses selected by individuals. Separating the item design and scoring phases is common in some applications of educational measurement. For example, Wilson's (2005) construct modeling framework distinguishes the "items design" that specifies criteria for how the items should be written from the "outcome space" that describes how observations should be categorized and scored. Similarly, Mislevy, Almond, and Lukas' (2003) evidence-centered design framework distinguishes between "task models" that structure the situations needed to collect observations and "evidence rules" that identify, summarize, and score evidence collected from tasks.

2.3.1. Designing Ordered Multiple-Choice Items. To write an OMC item, an answer choice must be associated with a level in a hierarchical model of student thinking. However, it is often challenging for content creators to write items where all possible ordered states are represented among the answer choices. Briggs and colleagues developed OMC items so that each answer choice could be mapped to a level of a hierarchical model of student thinking, but these authors used only a handful of possible designs. Items 1-3 in Figure 2.6 provide examples of some of the designs used by Briggs and colleagues. The first item represents the ideal scenario where each answer choice is mapped to a distinct level of a four-level hierarchical model of student thinking. The second item has answer choices mapped to 3 of the 4 levels, where two choices are mapped to the same level and one of the response options is linked to the highest

level of the model of student thinking. The third item is similar to the second one, but the highest level is not available as an answer choice.

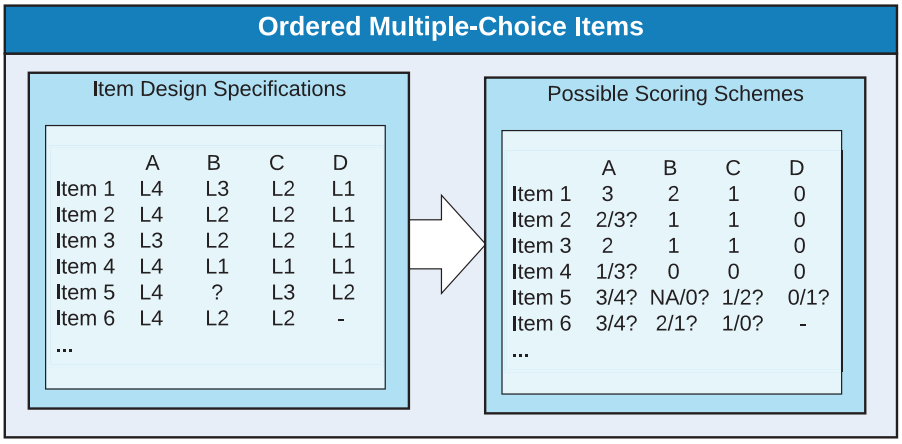


Figure 2.6. Designing and scoring ordered multiple-choice items.

Incorporating other design features results in additional types of OMC items, as illustrated by Items 4-6 in Figure 2.6. Item 4 is similar to a traditional multiple-choice item that has just one answer choice mapped to the highest level, but all of the other response options are linked to a single lower level. Item 5 has one response option linked to an unknown level, indicating that the item designer is unsure of how the answer choice may be connected to the hierarchical model of student thinking. Item 6 has one fewer response option than the other five illustrative items. Items 4-6 in Figure 2.6 vary the number of levels mapped to response options, incorporate an unknown level, and vary the number of response options, respectively.

Minstrell and his colleagues used facet clusters to design multiple-choice items with design specifications similar to Items 1-6 in Figure 2.6. These "Diagnoser" items are available for free on www.diagnoser.com (Thissen-Roe, Hunt, and Minstrell, 2004). Figure 2.7 presents three examples of Diagnoser multiple-choice items designed to evaluate student thinking relative

to the atoms facet cluster displayed in Figure 2.5. All items in Figure 2.7 have a response option associated with the scientifically accurate idea, but the items vary in terms of how the other facets are mapped to the distractors. In the first item, only three facets (00, 40, and 80) of the four that are possible (00, 40, 50 and 80) are available as response options. The second item only has response options associated with facets 00 and 80, and the third item has a response option associated with an "unknown" facet. These items also allow the number of answer choices to vary across items. Since Diagnoser and OMC items have a similar design, I will use the more concise and generalizable term "OMC items" to refer to both kinds of assessment tasks.

A piece of paper burns in a closed flask. As it burns, does the number of atoms in the flask increase, decrease, or remain the same? a. The number and type of atoms increase. [Facet 42] b. The number and type of atoms decrease. [Facet 41] *c. The number and type of atoms remain the same. [Facet 03]* d. The number of atoms remains the same but the types of atoms change. [Facet 40] e. None of the above. The paper is not made up of atoms. [Facet 80]	Three students were discussing atoms and cells during biology class. Brett: "Non-living things are made of atoms. Living things are made of cells, not atoms." James: "Only some non-living things are made of atoms." Steve: "All living and non-living things are made of atoms." With which student do you agree? a. Brett [Facet 82] b. James [Facet 80] *c. Steve [Facet 01]*	Assume you have the technology to cut a piece of aluminum foil into the smallest piece of aluminum possible. What would you be left with? Choose the best response. a. An incredibly tiny piece of aluminum foil. [Facet 81] *b. A single atom of aluminum that is too small to be seen. [Facet 01]* c. Nothing, because there is nothing to see when you get that small. [Facet 83] d. None of the above. [Facet Unknown]

Note: Answer choices associated with scientifically accurate ideas are italicized.

Figure 2.7. Multiple-choice items designed using the facet cluster for atoms. (reproduced from www.diagnoser.com)

The OMC items described in this section are just one attempt to connect the design of assessment tasks to a hierarchical model of student thinking. This innovative item type incorporates information about student thinking into the test development process during the

design stage (Gorin, 2006) rather than waiting until analysis. Additional designs are possible if the item type is varied. For example, a similar process could be used to create an alternative task template that separates the question stem in a multiple-choice item from the response options. The item stem could ask a question about a concept and an accompanying response bank could contain several possible responses that are mapped to levels in a hierarchical model of student thinking. Alternative designs may introduce new challenges, but the key takeaway is that the use of a hierarchical model in the design phase offers guidance for how observations can be scored.

2.3.2. Scoring Ordered Multiple-Choice Items. Although scoring OMC items should be straightforward relative to the levels of the LP, it is often difficult to know in advance whether levels of an LP are sufficiently distinct and ordered correctly to support grouping students into discrete developmental levels. The right-hand panel in Figure 2.6 illustrates some challenges that arise when using a hierarchical model of student thinking to assign scores to student responses. In the ideal scenario reflected in the design for Item 1, all levels of the hierarchical model of student thinking are mapped to answer choices and the observations can be scored using ordinal numbers (i.e., 0, 1, 2, and 3) to indicate increasing understanding of the disciplinary idea. Across all items, we would have evidence in the ideal assessment that scores of 0 reflect lower ability than scores of 1. This example illustrates that the scoring scheme associated with a hierarchical model of student thinking is the quantitative instantiation of the qualitative little learning theory.

One important challenge[5] of scoring OMC items involves establishing that scores associated with lower levels are actually indicative of less ability than scores associated with

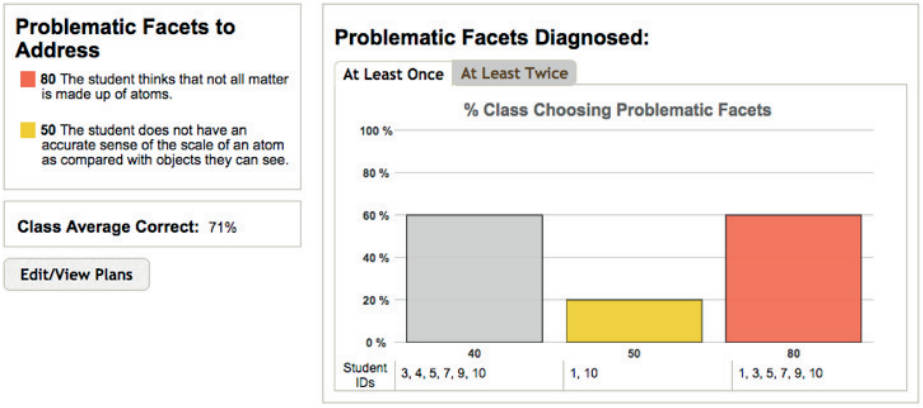
[5] Briggs and Alonzo (2012) describe additional challenges involved with the scoring and analysis of OMC items, including floor and ceiling scoring effects that may arise depending on the design of the OMC item. For example, Item 3 in Figure 2.6 exhibits a ceiling effect since students do not have the opportunity to select a response option associated with the highest level of the hierarchical model of student thinking (i.e., earn a score of 3).

higher levels, especially since the model of student thinking is often a hypothesis. Gotwals and Songer (2009) describe this as the challenge of articulating a clear order of levels in the "messy middle" of LPs. That is, hypotheses about the order of levels in the middle of LPs may be particularly susceptible to misspecification. Interpretation challenges become even more complex as items begin to deviate from the ideal design (e.g., Items 2-6 in Figure 2.6). These challenges mean that it may be difficult to rely exclusively on the little learning theory to score OMC items. As such, one important aspect of LP validation studies with OMC items should involve exploration into the meaning of partial credit scores both within and across items.

2.4. Interpreting Responses to Ordered Multiple-Choice Items.

When using the assessment triangle to develop educational assessments, the purpose of interpreting scored observations is to connect the observations back to the model of student thinking. In the early stages of research, it may be helpful to share descriptive information about student performance on OMC items to explore how teachers use this information. For example, Figure 2.8[6] represents an example of a teacher score report from the Diagnoser assessment system that displays a descriptive summary of class performance. The upper half of the score report presents an aggregate summary of students' performance followed by detailed information about the performance of individuals.

[6] I created this display using the tools available on www.diagnoser.com. Diagnoser (Thissen-Roe et al., 2004) is a website that contains all of the facet clusters that Minstrell developed, along with distractor-driven assessment items, for free use by teachers and students. After completing a Diagnoser question set, students receive a short report consisting of the percent of items answered correctly, a description of the facets they have mastered, and a description of problematic facets.



Student Id * (from class 273-21-)	Date Completed	Self Rating	% Correct	1	2	2e	3	4	5	6	7
				90%	80%		75%	80%	60%	50%	50%
1	05/23/2018	4	38	80	01		51,02	52	03	80	82
2	05/23/2018	1	100	01	01		02	02	03	03	01
3	05/23/2018	3	71	01	81		02	02	40	03	01
4	05/23/2018	2	86	01	01		02	02	40	03	01
5	05/23/2018	3	71	01	01		02	02	03	40	82
6	05/23/2018	1	100	01	01		02	02	03	03	01
7	05/23/2018	3	57	01	01		02	02	40	40	82
8	05/23/2018	1	100	01	01		02	02	03	03	01
9	05/23/2018	2	71	01	01		02	02	03	40	82
10	05/23/2018	4	13	01	83		51,80	52	41	40	80

* Only students who started the assignment are listed. Only the data of students who finished are included in the summary graphs above.

Figure 2.8. Example teacher score report.

Although thoughtfully designed and useful for the purpose of making class-level decisions, score reports that attempt to incorporate models of student thinking by only relying on descriptive summaries of students' raw scores can be challenging to use during for flexibly tailoring instruction to the needs of individual students. First, note that the score report presumes that the distinctions between the ideas represented by the three facets 40, 50, and 80 are all substantively meaningful. However, as described above, the two-digit numerical codes indicate a rough ranking of instructionally problematic ideas. This means that ideas 40 and 50 may be more similar to one another from an instructional perspective than idea 80. In other words, facets 40 and 50 may not be meaningfully distinguishable from one another and evidence that they overlap may be used to support collapsing these two facets into a single category. Second, information

about students' performance does not include a specific facet "diagnosis" that aggregates information about student performance across all items. Teachers are given the number of items that a student answered correctly (the "% correct" column in Figure 2.8) and the facets associated with each answer choice a student selected. This abundance of data may be difficult to interpret relative to the model of student thinking. For example, prior research has shown that interpreting and acting on student responses is extremely challenging for teachers to effectively engage in formative assessment (e.g., Schneider & Gowan, 2013; Frohbieter, Greenwald, Stecher, & Schwartz, 2011). Displays of student response data derived from psychometric modeling can help improve the interpretation of students' classroom assessment scores.

Consider Figure 2.9, which is an example of an item-person map that displays a histogram of student ability estimates in the leftmost panel and estimates of the difficulty of the items in the center using a probability scale. This visualization is an alternative representation of the table of student responses at the bottom of Figure 2.8 that can be adapted to provide improved information about what students know and can do. In the ideal scenario, the 7 Diagnoser items would each be written so that students could select a response option associated with one of the four facet levels (i.e., the design for Item 1 in Figure 2.6), the scores for these items would reflect progressively sophisticated understanding of the topic of atoms, and this order would hold across all items on the test. Each dot in Figure 2.9 represents the probability where selecting a response associated with one level becomes relatively more likely than the preceding level, given that the respondent has already attained the lower level. Across all items, the dots are ordered and do not overlap to support dividing the scale into four sections[7] (<-2, -2

[7] For this illustration, I divide the scale by taking the mean of the intersection parameters across all items, but alternative approaches can be used that combine content expertise with results from psychometric analyses.

to 0, 0 to 2, and >2). These regions of the scale can then be assigned labels relative the hierarchical model of student thinking used to develop the items and answer choices.

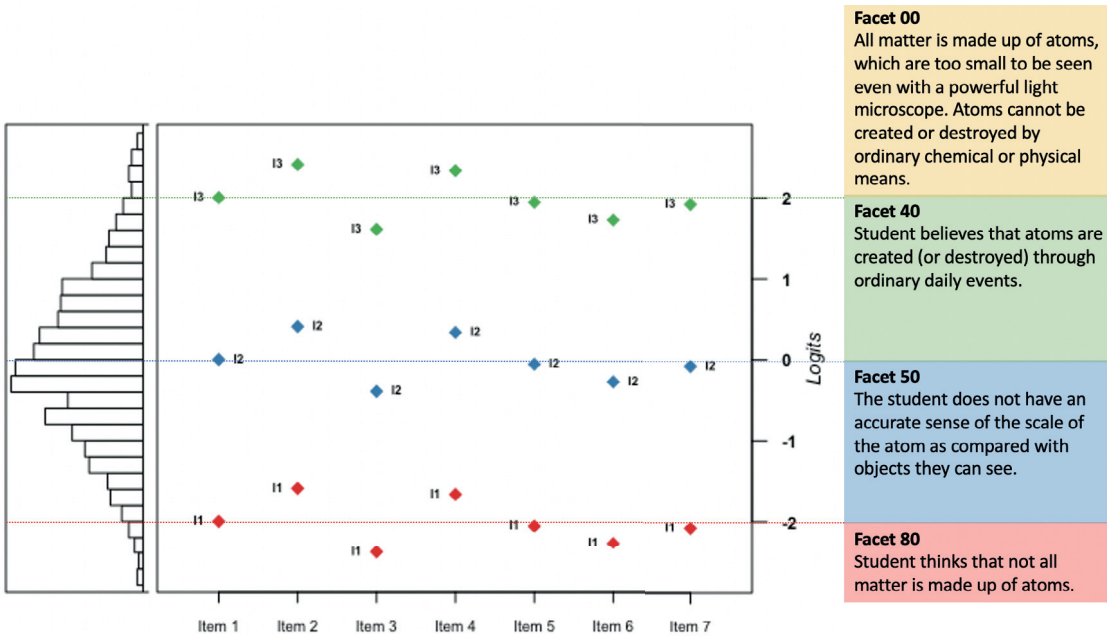


Figure 2.9. Example item-person map for LP validation studies.

The construction of a scale using items with a strong design could be used to support stronger interpretation of student scores relative to the model of student thinking. If the Diagnoser items followed the pattern displayed in Figure 2.9, there would be evidence that facets 40 and 50 are meaningfully distinguishable, and criterion-referenced interpretations could be assigned to students' test scores. For example, all students with scores that fall below the red line may need support understanding that all matter is made up of atoms (facet 80), and a productive instructional strategy may be to help these students understand the scale of atoms compared to objects that they can see (facet 50). If these kinds of probability scales can be established and maintained, they can also be used to provide criterion-referenced information about what

students have learned during the course of instruction. To improve the interpretation and uses of test scores from assessments designed using a model of student thinking, it is necessary to move towards the psychometric modeling of student response data.

Psychometric models describe the probability of responding to an item as a function of student ability and item characteristics. Psychometric models are sometimes referred to as "measurement models" because they attempt to measure an underlying latent variable. Wilson (2005) states "the measurement model must help us understand and evaluate the scores that come from the item responses and hence tell us about the construct, and it must also guide the use of the results in practical applications" (p. 16). In addition to the improved interpretations of test scores that can result from the application of psychometric models, the process of specifying a model and evaluating its fit may offer a systematic way to validate and refine a hypothesized LP (Briggs & Alonzo, 2012). Despite this potential, psychometric models can be challenging to use when response data has a complicated structure like that generated by OMC items.

A researcher may select a psychometric model on the basis of implicit or explicit perspectives on the measurement of latent variables. Andrich (2004) distinguishes between two dominant perspectives on measurement that inform the selection of a psychometric model: the "traditional" paradigm and the "Rasch" paradigm. In the traditional paradigm, the most appropriate psychometric model is the one that is best able to account for the data. Model selection occurs by fitting and using formal statistical tests to compare competing models. The challenge with this approach is that, in general, the more parameters a psychometric model contains the better it will explain the data. For this reason, complicated models tend to be favored by those with a traditional perspective on measurement. In the Rasch paradigm, the most appropriate psychometric model is one that satisfies criteria independent of any data. Rasch

(1960) was interested in models that permit "specific objectivity," which is the property that comparisons between persons must be generalizable beyond the specific conditions (i.e., items) under which they were observed. Conceptually, the idea is to objectively evaluate the progress of individuals even if the instruments used to collect observations change. This perspective emulates the additive nature of some measures in the physical sciences (e.g., the properties of density or force), which must be discovered indirectly rather than directly observed (Briggs, 2013; Luce & Tukey, 1964). Rasch models are chosen because they permit objective measurement, and the process of evaluating the fit of data to a Rasch model can illuminate the conditions that generate data that permit objective measurement.

Researchers have developed a wide variety of psychometric models consistent with both measurement paradigms. Models have been developed for data that can be scored dichotomously (e.g., correct v. incorrect responses) or polytomously (e.g., partial credit scoring schemes developed from LPs). These models include Item Response Theory models (van der Linden & Hambleton, 1997), cognitive diagnosis models (Rupp et al., 2010), and latent class models (Lazarsfeld & Henry, 1968). There are also unidimensional and multidimensional versions of most psychometric models. The application of psychometric models to response data that has been polytomously scored using a model of student thinking permits exploration into the design features of OMC items and by extension the order of levels in the model of student thinking.

As discussed in detail in the next chapter, most LP researchers have used Rasch models to interpret data relative to LPs by extending the progress mapping (Masters & Forster, 1996) and construct modeling (Wilson, 2005) measurement frameworks developed by students of Rasch. Progress and construct maps are qualitative descriptions of how an unobserved latent variable varies from one extreme to another. LPs are similar to these maps, but LPs consist of

ordered, discrete levels of sophistication. Because the Rasch model has the potential of yielding an interval scale that can convey instructionally useful information about student growth (Briggs, 2013; Briggs & Peck, 2015), Rasch models are often desirable for LP validation. However, during the early stages of LP validation efforts when researchers may still be exploring the order and distinctiveness of levels in LPs, more complex non-Rasch models may yield useful information about the structure of LPs. This information can be used to distinguish among possible messy middles. In the next chapter, I review how prior researchers have applied psychometric models to analyze data from OMC items to evaluate the order of student ideas.

Chapter 3

Literature Review

One goal of linking the results from psychometric analysis back to a model of student thinking is to investigate whether there is quantitative evidence to support the order and distinctiveness of the hypothesized levels of student ideas. In their review article, Gierl, Bulut, Guo, and Zhang (2017) describe the general steps involved in these kinds of alignment studies:

> First, plausible algorithms, rules, or procedures must be specified by content specialists. Second, plausible but incorrect distractors must be produced using these rules. Third, the misconceptions identified by the content specialists are [evaluated to see if], in fact, the same misconceptions are held by the students. Proper alignment of the assumptions is critical for creating distractors that measure plausible misconceptions (p. 1104).

The analysis of items that occurs in LP validation research is part of a broader research agenda that seeks to connect the results from psychometric analysis to substantive information about student thinking. The goal of this chapter is to review the methods other researchers have used to analyze OMC or OMC-like items. I conclude this chapter by identifying the contribution of this dissertation by situating this study within the research literature.

3.1. Literature Search Procedure.

To identify publications for inclusion in this review, I first conducted a broad survey of the research literature. The goal was to access citations of empirical studies. I input the terms *multiple-choice*, *distractors*, and *psychometric* with additional phrases like *order*, *OMC*,

cognition, *learning progression*, *learning trajectory*, and *concept inventory* into Google Scholar[8], which is a free web search engine that indexes full text or metadata of scholarly literature. This process uncovered a preliminary list of citations, two of which were highly cited by other researchers. As of May 2020, Sadler's (1998) article describing an IRT analysis of concept inventory items was cited in 302 later publications, and the article introducing the OMC format by Briggs and colleagues (2006) was cited by 341 subsequent sources. I conducted a second round of review by examining all publications citing these two articles.

I then filtered the citations. First, I excluded studies that were not published in English. Second, I excluded studies that analyzed assessments agnostic of a model of student thinking. There is a sizable test development literature that explores design features of items (e.g., the optimal number of multiple-choice response options) without attempting to connect analyses to a model of student thinking. Third, I excluded conference papers or technical reports that were later published in peer-reviewed outlets. Fourth, I excluded studies that did not score data in multiple categories relative to the model of student thinking. When items are designed using a model of student thinking (e.g., OMC items), the items can be scored using two categories (e.g., correct or incorrect) or scored using more than two categories by incorporating information about how answer choices are mapped to student ideas. Scoring data polytomously has the potential of providing stronger evidence to validate an LP because the scoring scheme is more strongly connected to the model of student thinking. Applying these exclusion criteria resulted in the retention of 42 studies, including 34 peer-reviewed articles, 6 dissertations, and 2 book chapters. I then read these articles and organized them into the schema displayed in Figure 3.1.

[8] Google Scholar indexes most peer-reviewed online academic journals and books, conference papers, theses and dissertations, preprints, abstracts, and technical reports. Although it has been criticized for including poor-quality publications, Google Scholar has been found to be comparable to subscription-based databases in some fields of study (e.g., Kulkarni, Aziz, Shams, & Busse, 2009).

Method Used to Interpret the Order of Student Ideas

Type of Model of Student Thinking	Exploratory Investigations	Confirmatory Investigations
Partially Ordered Models	**non-Rasch IRT Modeling** Sadler (1998, 2005) Battisti et al. (2010)	
	Descriptive Methods Morris et al. (2006, 2012) Ishimoto et al. (2017)	
	Rasch-Based Approaches Herrmann-Abell & DeBoer (2011, 2014) Wren & Barbera (2014) Wind et al. (2015, 2019) Laliyo et al. (2019)	**Multiple Approaches** Jorion et al. (2015)
	Diagnostic Classification Modeling Bradshaw & Templin (2014) Shear & Roussos (2017)	**Latent Class Analysis** Steedle (2008) Steedle & Shavelson (2009)
Hierarchical Models		**Polytomous Rasch Modeling** Briggs et al. (2006) Lee et al. (2009, 2011) Liu et al. (2011a, 2011b) Brown & Wilson (2011) Wallace (2011) Rivet & Kastens (2012) Weinberg (2012) Hadenfeldt et al. (2013, 2016) Fulmer et al. (2014, 2015) Plummer & Maynard (2014) Kuo et al. (2015) MacPherson (2015) Chen et al. (2016) Duckor et al. (2017) Duncan et al. (2017) Morell et al. (2017) Castle (2018) Furtak et al. (2018) Testa et al. (2019)
		Multiple Approaches Circi (2015) Chen et al. (2017)

Figure 3.1. Studies connecting analysis of OMC items to a model of student thinking.

The studies in Figure 3.1 are grouped using two dimensions of the assessment triangle framework introduced in the previous chapter. The rows in Figure 3.1 distinguish between analyses of assessments designed using partially ordered models of student thinking (e.g., concept inventories) and analyses of tests using hierarchical models of student thinking (e.g., tests consisting of OMC items). The columns in Figure 3.1 distinguish between exploratory investigations into how student ideas in the model of student thinking can be ordered relative to one another and confirmatory investigations that seek to provide validity evidence for the sequence of ideas in a hypothesized model of student thinking.[9] In each column and cell, I loosely organize the studies in chronological order to illustrate how the studies build on one another. The remainder of this chapter discusses these studies in detail, reviewing the methods other researchers have used to interpret the order of student ideas in a model of student thinking.

3.2. Exploratory Investigations into the Ordering of Student Ideas.

The exploratory investigations identified in the left column of Figure 3.1 use evidence from psychometric modeling to develop hypotheses about how ideas may be ordered in a partially ordered model of student thinking. The methods used to interpret data include non-Rasch IRT modeling, descriptive methods, Rasch-based approaches, or diagnostic classification modeling. These studies typically focus on the analysis of a subset of items on a test rather than all of the items that comprise the assessment. Sadler (1998) was the first to interpret the order of multiple-choice distractors relative to a partially ordered model of student thinking. He used

[9] In all of the studies listed in Figure 3.1, content experts were used to design the items and tests, but they may not have been involved in the interpretation of observations relative to the model of student thinking. Although the integration between content expertise and item analysis is not the focus of this review, I do discuss this relationship in the next chapter (see section 4.4.2).

graphs of response option curves that display how the probability of selecting answer choices vary across a latent ability continuum. Other researchers extended Sadler's method.

3.2.1. Using non-Rasch IRT Models to Analyze Response Option Curves. The goal of Sadler's (1998) study was to illustrate how the psychometric analysis of astronomy concept inventory multiple-choice items can reveal ordered patterns among student ideas. He analyzed response option curves produced by the multiple-choice IRT model (Thissen & Steinberg, 1984) relative to statements about what students may know. Figure 3.2 displays an example of the response option curves for an item that asked students to identify the main reason for it being hotter in the summer than in the winter. The item has five response options, and the lines in Figure 3.2 display the probability of selecting each answer across the latent ability continuum. The sixth line represents the latent probability of a "do not know" (DK) category. Two of the five response options were linked to misconceptions about the seasons (M1 and M2).

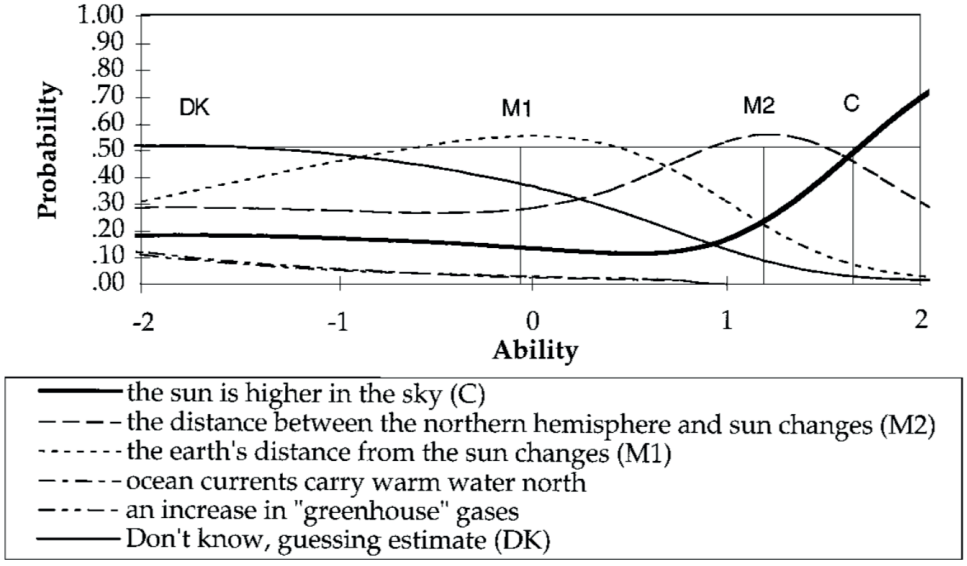


Figure 3.2. Multiple-choice model response option curves for a sample item. (reproduced from Sadler, 1998)

The expected form for response option curves is that the correct answer should monotonically increase while all the curves corresponding to distractors should monotonically decrease (Haladyna, 1994). The curves in Figure 3.2 deviate from the expected form. The dark black line (C) is the curve for the correct response, and it follows the expected trend; students with greater ability have a higher probability of selecting the correct response. However, the curves for the two options linked to misconceptions (M1 and M2) do not monotonically decrease. Students with an estimated ability of 0 are most likely to choose the option linked to M1 while those with an estimated ability of 1.2 are most likely to choose the option linked to M2. Sadler interpreted this result as suggestive of an order among the misconceptions: "the responses to this item in the student population suggest a progression in thinking about the earth's seasons from not discriminating between answers (don't know) to a solar distance model [M1] to a hemispheric difference model [M2] to solar altitude model [C]" (Sadler, 1998, p. 277). The significance of Sadler's study is that he illustrated how alternative ideas in partially ordered models of student thinking could have a meaningful empirical order.

In later writing, Sadler (2005) discussed how response option curves from multiple-choice tests could be used to create instructional resources like "curriculum maps." Curriculum maps describe developmental relationships among concepts. These maps can help teachers simplify the scope and content of their courses. Figure 3.3 is an example of a curriculum map that describes the development of students' understanding of the concepts of light and color. This representation is similar to the example item-person map displayed in Figure 2.9. The right-hand side of both displays presents a qualitative progression of student ideas connected to a quantitative ordering of item parameters and students along a common scale. Sadler was more interested in improving substantive interpretations about locations along the scale than in

developing a scale that has properties that allow for quantitative comparisons between students (e.g., quantifying the amount of knowledge Student A has relative to Student B).

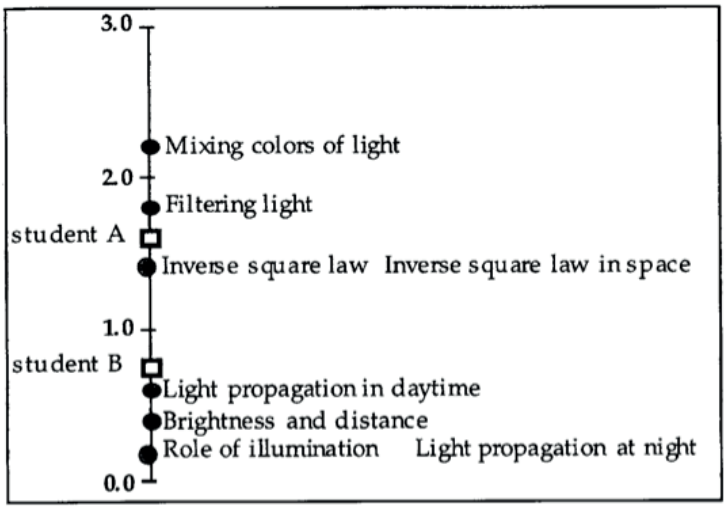


Figure 3.3. Example curriculum map.
(reproduced from Sadler, 2005)

Subsequent researchers replicated Salder's method of analyzing response option curves using data from different concept inventories. Battisti and colleagues used data from the conceptual inventory of natural selection to illustrate how IRT can be used to improve the quality of items on concept inventories (Battisti, Hanegan, Sudweeks, & Cates, 2010). They used the multiple-choice model to produce response option curves similar to those displayed in Figure 3.2 but compared and contrasted the shape of response option curves for pairs of items. For example, they recommended that item writers avoid using non-scientific terminology to describe misconceptions because the response curves associated with distractors using everyday language were flat (i.e., non-discriminating) compared to distractors that used scientific terms.

3.2.2. Using Descriptive Statistics to Analyze Response Option Curves. Rather than applying an IRT model to produce response option curves, Morris and colleagues derived them using students' total scores as a proxy for students' ability and the proportion of students selecting an option instead of probability (Morris, Branum-Martin, Harshman, Baxter, & Mazur, 2006; Morris, Harshman, Braum-Martin, Mazur, Mzoughi, & Baker, 2012). Figure 3.4 presents examples of descriptive response option curves for an "efficient" item, meaning one that has answer choices that reflect different sets of student ideas, and an "inefficient" item that just classifies students dichotomously. The efficient item illustrates a progression in student thinking about force from ignoring initial momentum (answer choice 2), to recognizing the components of momentum but adding them incorrectly (answer choice 3), to accurately understanding the nature of an impulse relative to Newton's First Law of Motion (answer choice 4). Morris and colleagues (2012) later argued that descriptive response option curve analysis provides instructionally relevant information about the order of student ideas consistent with the design of the concept inventory that can be overlooked when items are scored and modeled dichotomously.

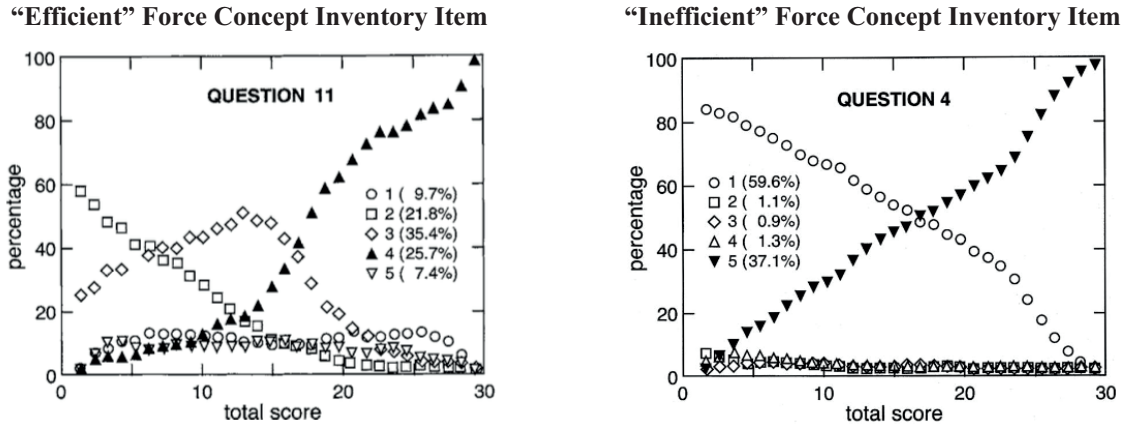


Figure 3.4. Examples of descriptive item response option curves. (reproduced from Morris et al., 2006)

Ishimoto and colleagues extended Morris and colleagues' (2006, 2012) analysis by exploring differences in the shape of descriptive response option curves produced from samples of Japanese and American students (Ishimoto, Davenport, & Wittmann, 2017). They found that response option curves had similar shapes for Japanese and American students across all items, providing some evidence to support the conclusion that misconceptions about force and motion transcend the two countries despite differences in culture, language, and education. These researchers also discussed how differences in the shapes of some of the curves could be used to identify translation issues or cultural differences in the teaching and learning of force concepts.

3.2.3. Rasch-Based Approaches to Response Option Curve Analysis. Herrmann-Abell and DeBoer used response option curves produced by the Winsteps Rasch modeling computer program (Linacre, 2020) to explore the order of student ideas. Herrmann-Abell and DeBoer (2011) developed 91 multiple-choice items written so that the answer choices correspond to a learning goal or a misconception related to middle school chemistry standards. Their pilot sample included 13,360 students, and they used linking items with test forms consisting of 26 to 30 items to ensure that each item was answered by an average of 4,000 students but individual students did not have to answer all 91 items. In a later study, they used a similar design to collect data from 23,744 students using 186 items written to assess energy forms, transformation, transfer, and conservation (Hermann-Abell & DeBoer, 2014). Herrmann-Abell and DeBoer analyzed their item banks using the simple Rasch model by comparing the distribution of students' abilities to the distribution of the difficulty of the items.

To explore hierarchies among misconceptions, Herrmann-Abell and DeBoer analyzed "option probability curves" for some items on the test. Option probability curves, like those

depicted in Figure 3.5, plot simple Rasch model estimates of student ability on the x-axis and the empirical proportion of students selecting each response option on the y-axis. Herrmann-Abell and DeBoer did not provide details on the specific options used to produce these plots, but their graphs resemble the "empirical category curves" described in the Winsteps manual (Linacre, 2020). Empirical category curves display the proportion of students selecting each response option for selected ranges along the ability scale. Winsteps also permits smoothing the curves using cubic splines. Figure 3.5 indicates that students with the lowest ability may think that atoms do not move (option B), those with the next lowest ability may believe atoms in physical objects move but those in air do not (option D), students with higher ability may think that atoms in air move but those in physical objects do not (option C), and those with the most sophisticated understanding believe that all atoms move (option A). Herrmann-Abell and DeBoer used these displays and prior research from science education to recommend alternative instructional sequences for the topics assessed by their item banks.

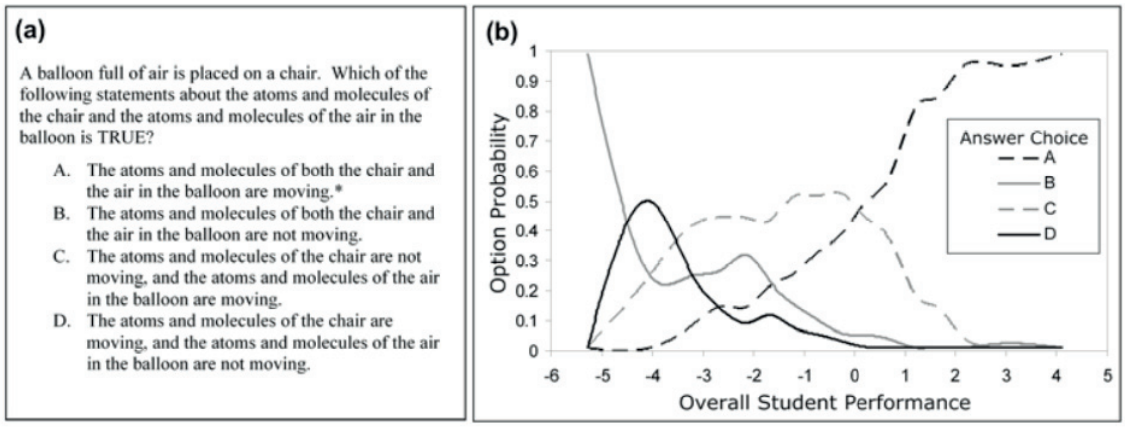


Figure 3.5. Example option probability curves.
(reproduced from Herrmann-Abell & DeBoer, 2011)

Wind and colleagues extended Herrmann-Abell and Deboer's Winsteps-based approach to explore how students' misconceptions about physical science may shift during instruction. Wind and Gale (2015) compared pre and post response option curves. To produce plots like the one displayed in Figure 3.6, Wind and Gale rounded Rasch model ability estimates to the nearest integer value (-3 to 4), computed the frequency of students selecting each item's answer choice for each integer value, and then converted this frequency into a proportion by dividing it by the total number of students at each point on the scale. The graphs in Figure 3.6 plot the rounded estimates of student ability on the x-axis and the proportion of students selecting each an answer choice on the y-axis. On the pre-test, the distractors monotonically decrease. However, after being exposed to instruction that emphasizes using free-body diagrams to calculate net forces on objects, the proportion of students selecting responses B and D increased among low ability students. Wind and Gale hypothesized that this may be due to students' unsuccessful attempts to use free body diagrams to answer the item. Wind and colleagues used a similar approach in another study to illustrate how patterns of student responses varied across classrooms for items on which the researchers observed qualitative differences in student performance (Wind, Alemdar, Lingle, Moore, & Asilkalkan, 2019).

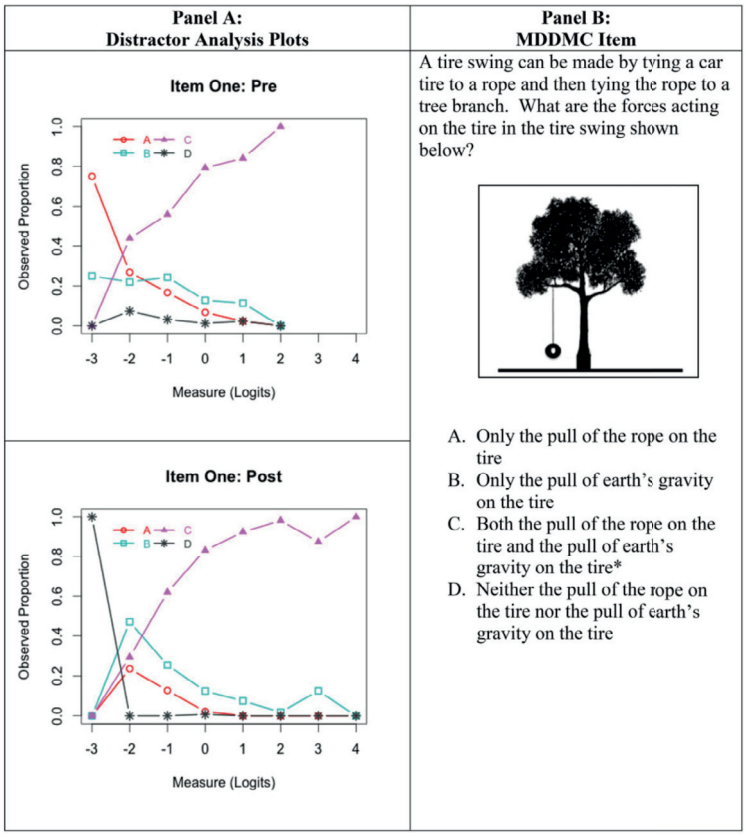


Figure 3.6. Example distractor analysis plot.
(reproduced from Wind & Gale, 2015)

Other researchers used variants of this Winsteps-based method of analyzing response option curves to support test score interpretations for new concept inventories or to identify productive instructional strategies. Wren and Barbera (2014) designed a thermochemistry concept inventory and used the empirical category curves from Winsteps to help understand whether or not distractors were providing information about an alternative conception. Laliyo, Botutihe, and Panigoro (2019) analyzed pairs of empirical category curves from Winsteps to identify dominant misconceptions among lower ability students. Laliyo and colleagues then identified themes among these misconceptions to help teachers identify productive instructional strategies (e.g., translating representational diagrams into reaction equations).

3.2.4. Using Diagnostic Classification Models to Group Students. Some researchers have developed or applied methods to improve the classification of students into meaningful groups relative to a partially ordered model of student thinking. Bradshaw and Templin (2014) developed a psychometric model that combines features from diagnostic classification models with those from the nominal response IRT model (Bock, 1972). The nominal response model is a simplified version of the multiple-choice model used by Sadler. Bradshaw and Templin's model alters the nominal response model by specifying categorical latent variables associated with misconceptions that decrease the probability of selecting the correct response. It permits a composite ability estimate that reflects the number of correct answers selected along with identification of the misconceptions associated with the incorrect answers.

Figure 3.7 is a graphical display of the kinds of response option curves produced from Bradshaw and Templin's model. Unlike the response option curves presented earlier in this chapter, the curves for the incorrect response options in Figure 3.7 always monotonically decrease. The pattern of misconceptions indicated by the 1's and 0's above each panel determine the probability of selecting an answer choice. In other words, the "order of the probability an examinee selects a given incorrect option is dependent upon his or her misconceptions and is invariant with respect to his or her ability" (Bradshaw & Templin, 2014, p. 417). For example, the first row in Figure 3.7 indicates an absence of the misconception associated with response D [100], and students have a low probability of selecting response D across all four panels. The curves in the second row indicate that students with low ability have a higher probability of selecting a misconception associated with response D. The absence of a curve for response E in the four rightmost panels (i.e., [010], [110], [011], and [111]) indicates that option E does not discriminate well between examinees who do and do not have this misconception.

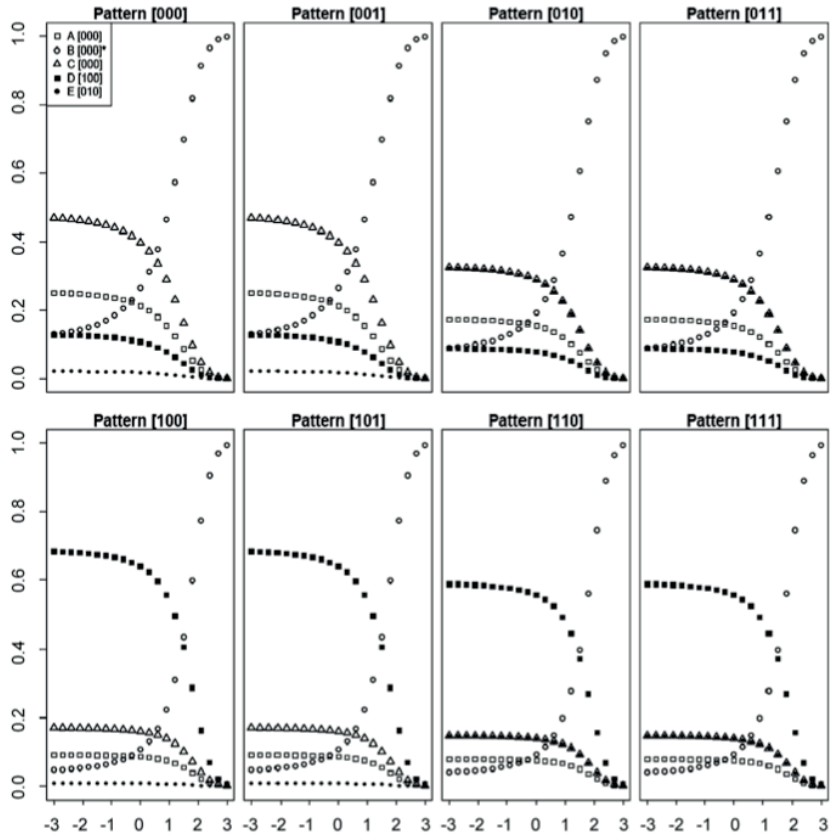


Figure 3.7. Example response option curves for a diagnostic classification model. (reproduced from Bradshaw & Templin, 2014)

Shear and Roussos (2017) explored how information from the diagnostic classification modeling of a geometry test can contribute validity evidence for a test using "item probability plots" like the one displayed in Figure 3.8. The numbers above the panels indicate the desired understanding (first digit) or presence of the misconception (second digit). Students who had a high probability of having a misconception [01] had a high probability of selecting a response option associated with a misconception across all items on the test, but students who had a high probability of having both the correct idea and misconception [11] had differing probabilities of selecting response options for items 1-5 and 6-12. Shear and Roussos used the latter result to revise the diagnostic classification model to measure an additional misconception. Note that the

focus here is on the measurement of misconceptions, and "incorrect" responses reflect answer choices that are both incorrect and inconsistent with the misconception of interest.

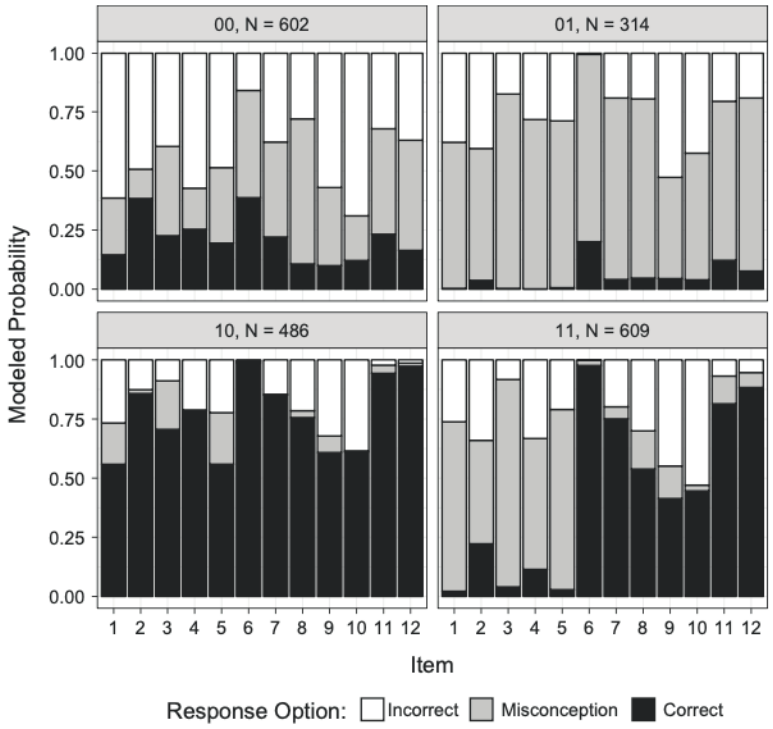


Figure 3.8. Example item probability plots.
(reproduced from Shear & Roussos, 2017)

Attempts to apply diagnostic classification models to interpret data from concept inventories highlight the importance of having a strong understanding of the relationship between items and the ideas being measured by them. In this measurement approach, the relationship between an item and the attribute being measured is quantified by a "Q matrix" (Tatsuoka, 1983). The specification of a Q matrix typically requires an integration of learning theory, content expertise, and empirical evidence. Misspecification of the Q matrix results in poor classification accuracy rates that compromise the validity of inferences resulting from the application of diagnostic classification models (Madison & Bradshaw, 2015). Although

diagnostic classification models may be promising tools to classify students into meaningful groups, there should first be a strong understanding of the relationship between the design of the test relative to the student ideas being measured so that a defensible Q matrix can be constructed.

3.3. Confirmatory Investigations into the Ordering of Student Ideas.

The confirmatory investigations listed in the right-hand column of Figure 3.1 use evidence from psychometric modeling to evaluate intended interpretations of test scores relative to models of student thinking. The methods used to interpret data include polytomous Rasch modeling, latent class modeling, or approaches that combine IRT analysis with diagnostic classification modeling. These studies can be grouped into two categories: 1) studies that use polytomous Rasch models to evaluate hierarchical models of student thinking; and, 2) comparative investigations that either compare models of student thinking or multiple methods.

3.3.1. Using Polytomous Rasch Models with Construct Maps. Since the publication of *Knowing What Students Know*, some researchers have used "construct maps" to integrate models of student thinking into educational assessments. Construct mapping is a Rasch measurement approach developed by Wilson (2005) and promoted by his students and colleagues. This approach consists of four building blocks: 1) specification of a construct map that describes how the attribute of measurement varies from one extreme to another; 2) item design specifications that describe how the construct informs the design of the items; 3) the outcome space that describes how observations collected from the items are categorized and then scored; and 4) the measurement model which relates the scores back to the construct.

Early applications of construct mapping focused on improving the design of assessment items through the development of construct maps with ordered levels. Building on Wilson's description of a construct map, Briggs and colleagues considered construct maps representations of "unidimensional continua with distinct levels" where each level "reflects a hierarchical stage through which students pass as they gain a qualitatively richer understanding about a given construct" (Briggs et al., 2006, p. 38). Imposing levels onto a construct map facilitated the design of OMC items where answer choices could be associated with levels of the map. Shortly after OMC items were introduced, Liu and colleagues developed "explanation multiple-choice items," which are similar to OMC items in that each answer choice is linked to a level in a hierarchical construct map, but items are paired such that the first item asks students about a particular phenomenon and the second asks students to select a response that describes their reasoning (Lee & Liu, 2009; Liu, Lee, & Linn, 2011a, 2011b; Lee, Liu, & Linn, 2011).

Initial methods used to interpret data collected from OMC or explanation multiple-choice items involved fitting a polytomous Rasch model and interpreting item parameters relative to the hierarchical model of student thinking. Briggs and colleagues (2006) suggested using the ordered partition model, which is a Rasch model that permits plotting a response option curve for each response option. Liu and colleagues used the partial credit model to analyze the explanation portion of their items. Both Rasch models assume that there is preexisting evidence to support the ordering of scores assigned to the response categories.

Liu and colleagues initially focused on interpreting cumulative probability curves produced by the partial credit model. Cumulative probability curves, like those displayed in the left panel of Figure 3.9, plot values of the latent variable on the x-axis and the probability of obtaining score $k + 1$ or higher from score k or lower on the y-axis. The "thresholds" indicated

by the letters A-E are the estimated value of the latent variable that intersects with the 50% probability on the *y*-axis. The letter A corresponds to the lowest level of the construct map (irrelevant) and E the highest (complex). Liu and Lee (2009) interpreted the relative magnitudes of the thresholds using item-person maps like the one displayed in the right panel of Figure 3.9. They noted that the short distances between the thresholds for the two lowest levels of the construct map (no-link and irrelevant) offer evidence that those two levels could be combined.

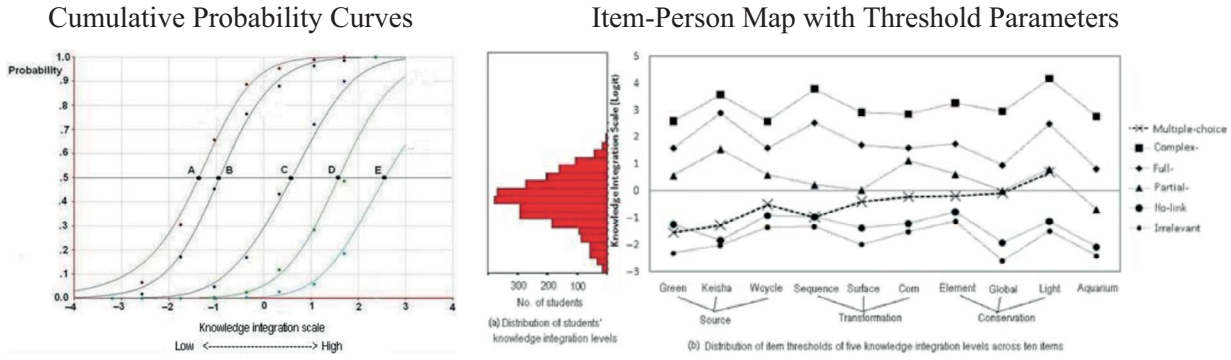


Figure 3.9. Example cumulative probability plots. (reproduced from Lee and Liu, 2009)

Subsequent research by Liu and colleagues formalized connections among empirical parameters and the construct map. Alternative item-person map displays, like the one in Figure 3.10, grouped threshold parameters relative to levels hypothesized by the construct map (Liu, Lee, and Linn, 2011a, 2011b; Lee, Liu, & Linn, 2011). The section labeled "MC" consists of the threshold parameters for traditional multiple-choice items. The remaining regions plot the thresholds for constructed response items scored using the same five-level construct map displayed in the legend of Figure 3.9 (i.e., irrelevant, no-link, partial, full, and complex). Liu and colleagues interpreted the progression of threshold parameters in the item-person map as evidence to support the order of levels in the construct map.

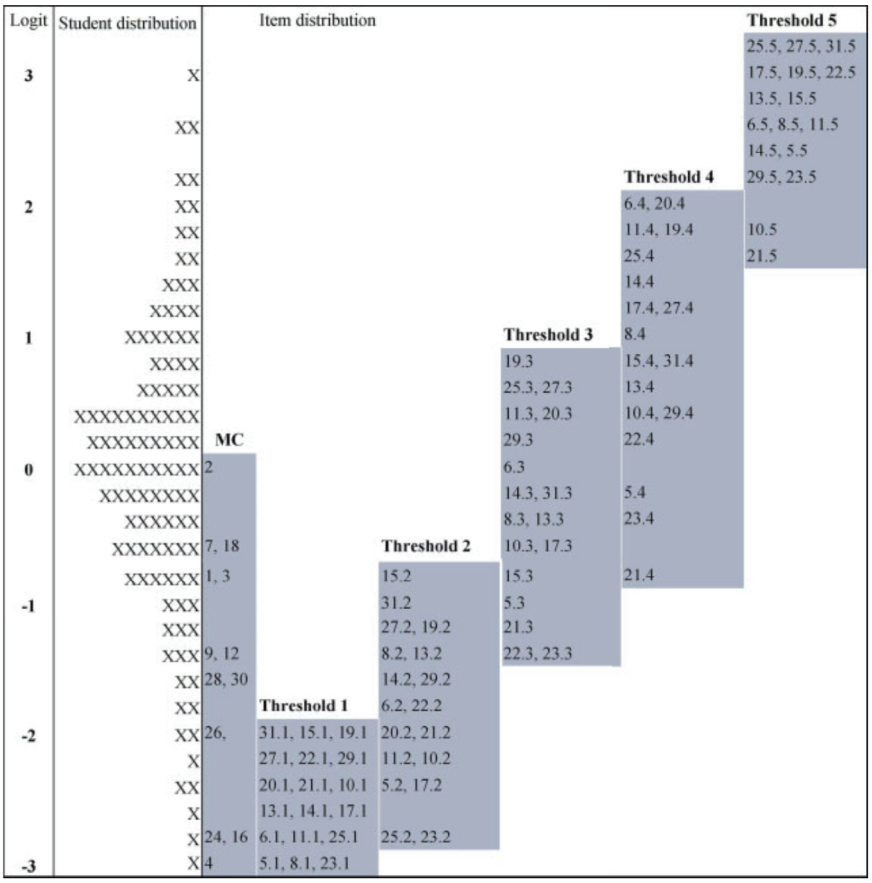


Figure 3.10. Example item-person map.
(reproduced from Liu, Lee, and Linn, 2011b)

Other researchers analyzed the order of threshold parameters for different constructs in a similar manner as Liu and colleagues. Rivet and Kastens (2012) developed items using a three-tiered construct map describing the development of students' analogical reasoning around physical models in earth science. Similar to the display in Figure 3.10, they grouped partial credit model threshold parameters in an item-person map together to provide evidence for the order of levels in the construct map. They also identified problematic items where the small distance between the item's threshold parameters (e.g., similar estimates for threshold 1 and threshold 2) supported collapsing the scoring categories. Plummer and Maynard (2014) developed items using a construct map that described how students' reasoning about celestial motion becomes

more sophisticated. They interpreted the order of threshold estimates from an item-person map relative to the design features of the items and illustrated how construct mapping could be used to specify an LP for celestial motion.

In the peer-reviewed publications reviewed so far, researchers omit the details of how they developed and revised their scoring approach in favor of advancing more important research contributions. Information about how scoring schemes were developed and revised can be found in some dissertation studies. MacPherson's (2015) dissertation used construct mapping to develop a set of "item bundles" to measure the scientific practice of argumentation. Item bundles presented students an opening prompt about a particular ecological scenario, asked students 1-2 multiple choice questions followed by open-ended items that required students to construct their own argument and compare and contrast two hypothetical arguments. Items were initially scored individually using a partial credit scoring scheme relative to the construct map, but MacPherson found that very few students earned partial credit. To improve scoring and interpretation of PCM parameters, MacPherson then used the construct map to combine items within a bundle to create "super-items" that were scored using partial credit. The revised item-person map revealed that the PCM threshold parameters were more separated and easier to map onto the construct map.

One challenge that occurs when interpreting item-person maps relative to a construct map is dividing the latent ability scale into distinct regions relative to the construct map. Figures 3.9 and 3.10 illustrate that there can be overlap among threshold parameters across items, resulting in difficulty placing students into distinct levels. To help resolve this issue, Brown and Wilson (2011) presented a general process of how construct mapping could be used with a model of student thinking to divide the latent ability scale into regions corresponding to levels of a hierarchical model of student thinking. The left panel of Figure 3.11 presents an example of a

construct map for developing conceptual understanding of scientific phenomena involving dynamic equilibrium. The dark bold vertical arrow along the center of the display illustrates the developmental continuum, and the descriptions to the right of the arrow are hypothesized levels of understanding ranging from "None" and "Acausal" (A) at the lower anchor to "Emergent" (E) at the upper anchor. The right panel of Figure 3.11 is an item-person map that connects partial credit item threshold parameters to the construct map. The thick horizontal lines in the right panel represent the weighted averages of the empirical thresholds across items (e.g., "A" represents the weighted average of the threshold parameters for items 2-9). Brown and Wilson's slices the latent ability scale into distinct regions that can be interpreted using the construct map.

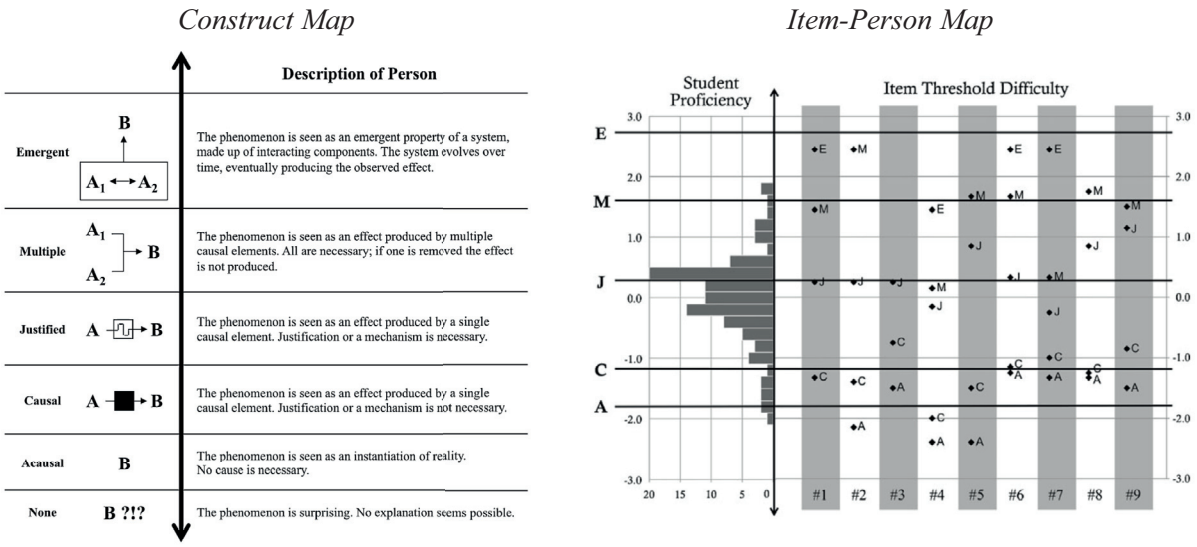


Figure 3.11. Connecting construct maps and item-person maps. (reproduced from Brown & Wilson, 2011)

Other researchers have used variants of Brown and Wilson's method of aggregating threshold parameters to improve understanding of the structure of hierarchical models of student thinking. Weinberg's (2012) dissertation used construct mapping to develop an assessment to

measure students reasoning about basic mechanical systems in the physical sciences. He interpreted the order of the mean partial credit threshold parameters as evidence to support the recovery of an empirical progression consistent with the construct map. Kuo and colleagues analyzed a subset of items designed using a construct map for scientific inquiry, interpreting an increase in the mean partial credit item thresholds as evidence that the items could differentiate among different levels of student performance (Kuo, Wu, Jen, & Hsu, 2015). Duckor, Draney, and Wilson (2017) used construct mapping to develop a test to measure classroom assessment literacy. They used partial credit model threshold parameters to divide the scale into distinct ordered "bands" that had some overlap among levels. Furtak and colleagues calculated pre and post differences in the magnitude of partial credit threshold parameters for OMC items to explore the impact of teachers' professional learning experiences (Furtak, Circi, & Heredia, 2018). Testa and colleagues developed LPs and OMC items for three big ideas related to learning quantum mechanics (Testa et al., 2019). Their analysis of the mean partial credit threshold parameters revealed that the LP could be confirmed for two big ideas, but the order of the threshold parameters for the third big idea did not support the LP hypothesis.

Some studies use multidimensional construct maps, where each level of an LP is conceptualized as a separate latent variable. Morell and colleagues generated a multidimensional item-person map, displayed in Figure 3.12, using a four-level LP for how students understand the structure of matter (Morell, Collier, Black, & Wilson, 2017). They interpreted the overlap in thresholds – particularly for the physical changes (PHS), explanations of physical changes (EPC), and explanation of chemical changes (ECC) levels – as evidence to disconfirm the sequence of levels in the LP. However, they illustrated how a substructure for the ECC level could be derived by comparing the design of the items to the order of threshold parameters.

Hadenfeldt and colleagues analyzed how the order of multidimensional partial credit threshold parameters for OMC items varied across grades to provide evidence for an LP that describes how students understanding of the concept of matter develops (Hadenfeldt et al., 2013, 2016). Duncan and colleagues used multidimensional item-person maps to first establish an ordering among LP levels and then to compute learning gains by construct (Duncan, Choi, Castro-Faix, & Cavera, 2017). Castle's (2018) dissertation compared the order of multidimensional threshold locations for scaffolded and non-scaffolded items designed using an LP for matter. Castle found that scaffolded items tended to be easier, especially for students with lower abilities.

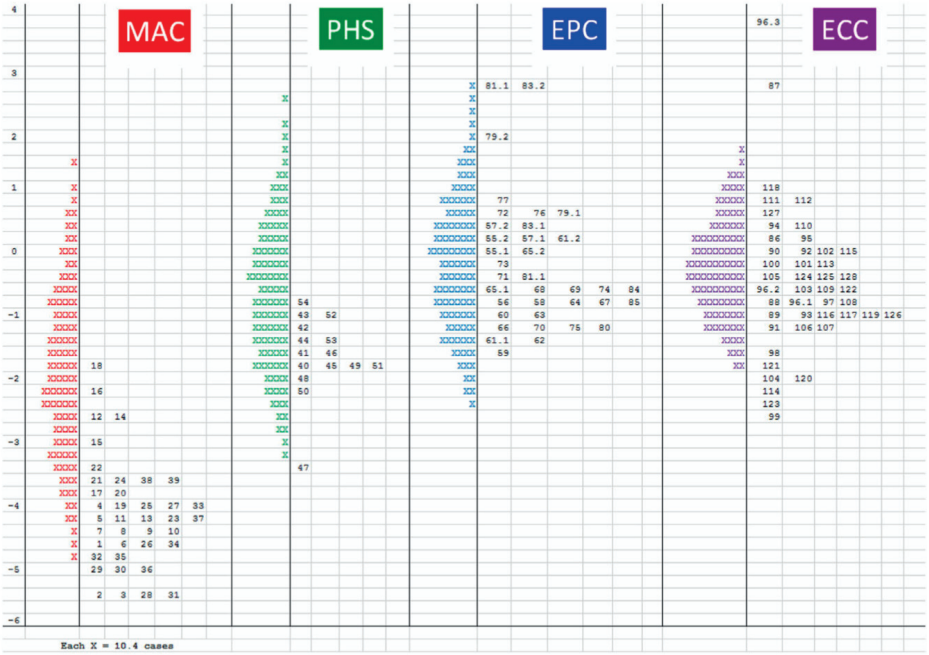


Figure 3.12. Example multidimensional item-person map. (reproduced from Morell et al., 2017)

So far, the methods used to interpret data relative to a model of student thinking have used threshold parameters produced from polytomous Rasch models. Threshold parameters are computed from cumulative probabilities, and they can be useful to determine whether or not

there is sufficient evidence to support distinct levels of ideas. However, they assume that scores can be ordered relative to a *hypothetical* model of student thinking. This is a strong assumption that should be evaluated as part of the assessment validation process. Some researchers have attempted to interpret threshold parameters with response option curves to clarify the structure of LPs. Gavin Fulmer and colleagues used Alonzo and Steedle's (2009) LP for force and motion to score items from the force concept inventory (Fulmer, Liang, & Liu, 2014; Fulmer, 2015). They presented threshold parameters and response option curves derived from the partial credit model to illustrate the expected ordering of responses relative to the LP. Fulmer and colleagues concluded that the empirical order of parameters supported the order anticipated by the LP.

There are a handful of studies that interpret category intersection parameters from polytomous Rasch models. Category intersection parameters are the points on the latent ability continuum where two consecutive response category curves intersect one another. Consider Figure 3.13, which displays partial credit model response option curves for two items that can be scored using three ordered categories. The left panel illustrates response option curves that follow the expected ordering, meaning the intersection between the response option curve for category 1 and 2 is smaller in magnitude than the intersection between the curves for categories 2 and 3. For the item in the right panel, the categories do not follow the anticipated order because the magnitude of the intersection between the curve for category 1 and 2 is larger than the intersection between the curves for categories 2 and 3. Like threshold parameters, category intersection parameters also assume that scores are ordinal. However, because they only compare adjacent categories (i.e., they are not cumulative probabilities), they can reveal when the order of categories may be incorrectly specified. For example, category intersection parameter reversals

like those displayed in the right panel of Figure 3.13 may indicate that two levels of an LP should be reversed rather than collapsed.

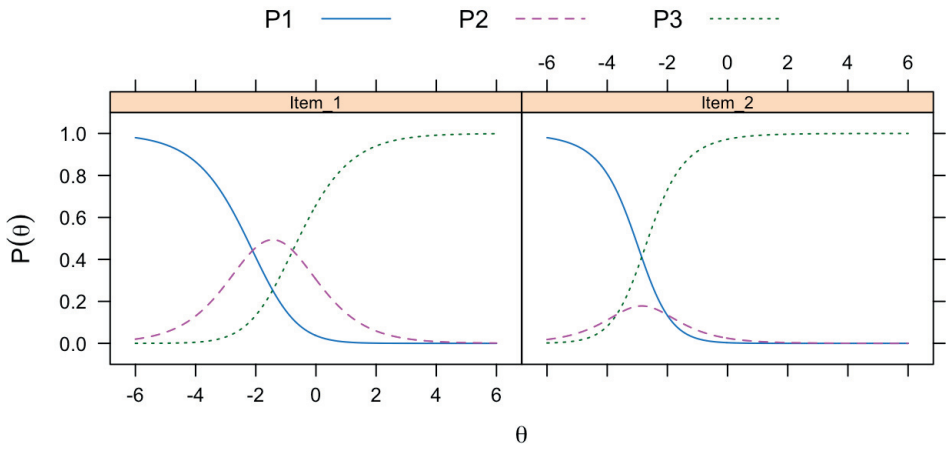


Figure 3.13. Examples of ordered and disordered category intersection parameters.

Category intersection parameters have not been widely interpreted relative to models of student thinking. Wallace's (2011) dissertation study compared the order of partial credit model threshold and category intersection parameters for an assessment of cosmology developed using construct mapping. Wallace uncovered examples of intersection parameters that did not follow the expected ordering, but he considered these estimates to be "counterintuitive" relative to the design of the items and elected to interpret threshold parameters instead. Chen and colleagues used the partial credit model to analyze items designed using an LP and found some evidence that items had intersection parameters in the wrong order relative to the LP (Chen, Gotwals, Anderson, & Reckase, 2016). They interpreted these reversals as suggestive of problems with the empirical data (i.e., there were too many or too few responses at a particular level relative to the proportion of responses at that level from other items) rather than problems with the LP or items.

3.3.2. Comparative Approaches to Interpretation. The remaining studies in the right-hand column of Figure 3.1 are examples of comparative investigations that test hypotheses about models of student thinking. Steedle's dissertation and subsequent research (Steedle, 2008; Steedle & Shavelson, 2009) compared two models of student thinking using latent class analysis. Steedle compared a facet cluster for force and motion to a force and motion LP. He then developed latent class models consistent with assumptions inherent in each model of student thinking, and fit response data to each model. The goal of these studies was to compare whether patterns of expected responses supported valid interpretations of LP diagnoses. Results from these studies indicated that students who had a scientifically accurate understanding of the concept of forces usually reasoned systematically across items on the assessment, but other students did not reason systematically. Steedle concluded from these results that student classifications into levels of an LP can therefore be invalid.

Other researchers used multiple interpretation methods with the same model of student thinking to compare and contrast information about how student thinking develops. Jorion and colleagues developed an analytic framework to evaluate the validity of concept inventory claims (Jorion, Gane, James, Schroeder, DiBello, & Pellegrino, 2015). They recommended first using descriptive analyses consistent with classical test theory and IRT modeling to determine the properties of the items and the overall reliability of the assessment. Any items that were identified as poorly functioning based on statistical criteria should then be removed from the analysis. Next, the results from exploratory and confirmatory factor analyses should be used to recover an empirical structure for the test comparable to the design. For example, if a concept inventory was designed to measure three misconceptions, the results from factor analysis should be able to recover three factors associated with these ideas. Once the structure of the tests is

confirmed, diagnostic classification models can be used to classify students into groups. Jorion and colleagues analyzed data from three concept inventories and found some evidence of overall test reliability but limited evidence that any of the concept inventories could reliably measure students' misconceptions.

Circi's (2015) dissertation used the partial credit model, attribute hierarchy model, and a generalized diagnostic classification model to analyze OMC items to provide evidence to support the levels of a force and motion LP. Her study revealed that there was little evidence that the items followed the hypothesized sequence of levels described in the LP, but it also identified some problems with the estimation procedure for the attribute hierarchy model. The artificial neural networks used in the estimation of attributes can produce varying attribute probability estimates when trained using the same data (Briggs & Circi, 2017). This finding indicated that it is important to provide evidence to support the different levels of a hierarchical model of student thinking independent of the results from a psychometric model (e.g., through designing items relative to a model of student thinking). Chen and colleagues analyzed the curricula of six high-achieving regions identified from international assessments and identified nine attributes related to the learning of number sense (Chen, Yan, & Xin, 2017). They then developed an LP using these attributes, scored items from international tests using the LP, and used data from the attribute hierarchy model and simple Rasch model to modify the original LP hypothesis.

3.4. Review of Methods to Analyze Items Designed Using a Model of Student Thinking.

The studies reviewed in this chapter illustrate that an important goal of using a psychometric model to connect items to a model of student thinking is to provide evidence for

the order of student ideas. Exploratory studies use response option curve analysis to help understand the order of misconceptions or use diagnostic classification models to explore how students are classified relative to misconceptions. They typically focus on the analysis of individual items relative to a model of student thinking. Confirmatory investigations solicit evidence to evaluate a clearly articulated hypothesis about how student thinking develops. These studies attempt to analyze all of the items on a test together, using displays like item-person maps. As the review in this chapter has shown, researchers have used a wide variety of methods to conduct these exploratory and confirmatory investigations.

All of the exploratory investigations interpreted data collected from items designed using partially ordered models of student thinking. Studies that analyzed response option curves sought to identify ordered patterns among student misconceptions for the purpose of supporting teaching and learning. This approach can be helpful to evaluate the scores assigned to data relative to a model of student thinking or to generate hypotheses about how student misconceptions might be ordered. However, it can be difficult to synthesize the rich qualitative insights obtained from response option curve analysis across all items on a test, especially if a goal of analysis is to create a scale that can be used to measure student thinking. Furthermore, applying more complex psychometric models like diagnostic classification models requires a strong understanding of the cognitive structure of all items on the assessment that can be challenging to obtain from item-specific response option curve analyses.

In contrast, the majority of confirmatory investigations used items that were designed using hierarchical models of student thinking like LPs. These studies all rest on a strong, implicit assumption that student ideas can be rank ordered relative to the *hypothetical* levels in the model of student thinking. This assumption is used to score data that is then modeled to obtain

empirical evidence that the ordering of ideas is an appropriate assumption. In the majority of confirmatory studies using the polytomous Rasch model, researchers used threshold parameters to illustrate that the rank order assumption is appropriate. However, threshold parameters are incapable of detecting violations of the rank order hypothesis because they will always be ordered since they are derived from cumulative probability estimates (Andrich, 2013, 2015). Category intersection parameters are more appropriate indicators of problems with the rank order hypothesis, but they can be difficult to interpret relative to the model of student thinking.

In LP validation research, stronger evidence to support interpretations of test scores relative to the order of ideas in a hierarchical model of student thinking can be produced by combining exploratory and confirmatory methods. Analysis of response option curves from non-Rasch nominal response IRT models can provide evidence to support partial credit scoring schemes relative to a hierarchical model of student thinking. Category intersection parameters derived from polytomous Rasch models can then be used to identify problematic items. They can also be analyzed across items using item-person maps like the ones displayed in this chapter to help impose levels onto the latent ability scale. This dissertation contributes to the literature on LP validation by developing a method, presented in the next chapter, to integrate evidence from the nominal response and partial credit IRT models to strengthen validity arguments for assessments designed using hierarchical models of student thinking.

Chapter 4

Methods

Building on the research reviewed in the previous chapter, this chapter presents a method to provide evidence to support interpretations of test scores relative to hierarchical models of student thinking. The first section introduces a general analytic framework that describes some types of validity evidence that cognitive-psychometric modeling can contribute to support LP validation studies involving OMC items. The second and third sections present the functional forms of the two polytomous IRT models I use in this study – the Nominal Response Model (NRM; Bock, 1972) and the Partial Credit Model (PCM; Masters, 1982) and discusses how to interpret empirical estimates of item parameters from these two IRT models. In the fourth section, I present a novel method that uses the NRM and PCM with OMC items for the purpose of evaluating hypotheses about the order of levels in hierarchical models of student thinking.

4.1. Analytic Framework for LP Validation Studies.

During the initial stages of LP research, validation of a test designed using the LP can help identify whether or not the LP is an appropriate representation of how student learning occurs. The *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) identify five sources of evidence that can support the validity of tests: evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence for the validity and consequences of testing. The *Standards* recognize that validation is an ongoing process and that the strongest validity

arguments integrate all five sources into a coherent account of the extent to which theory and empirical evidence support the intended interpretations of the test scores. Here, my focus is on the preliminary stage of building an evidentiary argument that begins after an LP is developed, an assessment is constructed, and data from the assessment is interpreted relative to the LP. The framework displayed in Figure 4.1 uses the assessment triangle (Figure 2.1) to begin integrating theoretical evidence based on the content of the test with empirical evidence from analysis of the internal structure of the test.

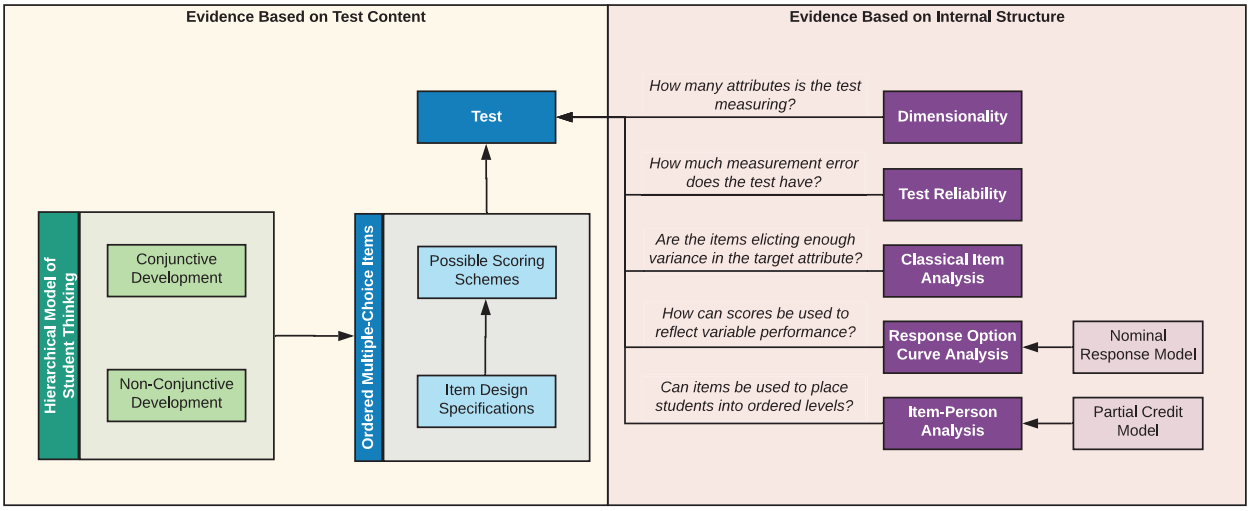


Figure 4.1. Analytic framework for LP validation studies.

The left-hand portion of Figure 4.1 represents *theoretical* evidence that the content of the test (blue boxes) is related to the construct that the test is intended to measure (green boxes). When OMC items are designed using hierarchical models of student thinking, the developmental hypothesis is used to create response options. However, as described in Chapter 2, there are many possible designs for OMC items and there are multiple options for the scoring of OMC items relative to the model of student thinking (e.g., assigning partial credit scores using the

levels of the model of student thinking or collapsing one or more categories together). Some item designs and scoring options may support stronger interpretations of test scores relative to the LP. Thus, two critical questions that arise when using a hierarchical model of student thinking to represent a construct and design a test consisting of OMC items are: a) whether or not the order of levels, particularly in the messy middle, is sufficient to support partial credit scoring; and, b) whether or not the items can be ordered to support criterion-referenced interpretations along a developmental scale (e.g., see Figure 2.9).

The right-hand side of Figure 4.1 represents the degree to which the relationships among the test components conform to the construct on which the proposed test score interpretations are based. Here, I consider a test to be a set of items. Patterns among observed responses to the items can be used to provide some *empirical* evidence that the test scores have meaning relative to a hypothesis of ordered student ideas. Analyzing correlations among the item responses can provide evidence to support hypotheses about the number of latent variables the test is intended to measure (i.e., the dimensionality of the test). Correlations between the scores on an item and overall scores the test can provide information about the reliability of the test, and reliability estimates can be used to quantify the amount of measurement error contained in the test scores. Methods to analyze the dimensionality and reliability of tests are well-established in the literature.[10] Test-level analyses are a useful first step in LP validation studies because they provide signal that the overall test is of reasonable quality.

[10] Factor analysis and structural equation modeling can be used to analyze the dimensionality of a test and explore hypotheses about how latent variables may be structured. Tabachnick and Fidell (2019) and Kline (2011) provide general overviews of exploratory and confirmatory factor analysis techniques. Reliability, as defined by the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014), refers to the consistency of test scores across instances of the testing procedure. It is often estimated using one or more reliability coefficients, often derived from classical test theory. The mostly commonly used reliability coefficient is Cronbach's (1951) alpha. See the research articles in Volume 34 Issue 4 of *Educational Measurement: Issues and Practice* for a contemporary discussion of issues related to reliability and dimensionality.

More important for LP validation, though, is analysis of the properties of the items because it is the items that are designed to reflect the model of student thinking. There should first be some evidence that the items are eliciting variance in the latent variable characterized by the hierarchical model of student thinking. This evidence can be obtained by analyzing the distribution of item difficulty, either through descriptive classical item analyses or through analyses of the items using an IRT model for dichotomously scored data. When a test consists of OMC items, additional evidence should be produced to support a partial credit scoring scheme and a scale that can be interpreted relative to the model of student thinking. Analysis of response option curves can provide insight into how partial credit scores could be assigned to the observed data. Analyzing item-person maps can provide evidence about the extent to which a scale can be established to distinguish among respondents relative to the hypothesized levels in the hierarchical model of student thinking. In the next section, I introduce two IRT models that were developed to support the latter two analyses.

4.2. The Nominal Response and Partial Credit IRT Models.

IRT is a useful mathematical framework for LP validation research because it permits both items and persons to be located along a continuous scale, facilitating comparison of the empirical order of item parameters relative to the order expected under an LP hypothesis. In general terms, IRT describes the probability of observing response category k of the random variable X_{pi} by person p to item i using an item response function (IRF):

$$P\left(X_{pi} = k\right) = f\left(\boldsymbol{\theta}_p, \boldsymbol{\xi}_i\right), \tag{4.1}$$

where $\boldsymbol{\theta}_p$ represents respondent-specific parameters and $\boldsymbol{\xi}_i$ represents item-specific ones (Briggs & Alonzo, 2012). Equation 4.1 is the general mathematical function for the IRF of any IRT model. IRT models have been developed for response data that can be scored using dichotomous or polytomous scoring schemes. Polytomous IRT models are most relevant for LP validation studies with OMC items because OMC items provide the strongest validity evidence when they are scored relative to the levels of a hierarchical model of student thinking.

Polytomous IRT models can have "difference" or "divide-by-total" functional forms, depending on how the model describes the probability of a response in category k (Thissen & Steinberg, 1986). Difference models express the right-hand side of equation 4.1 as the difference between the probability of a response in category k or above and the probability of a response in category $k+1$ or above. Divide-by-total models express the right-hand side of the IRF as the probability that the response is in category k divided by the sum of probabilities across all response categories. All polytomous IRT models have both difference and divide-by-total specifications. Most LP validation studies have used difference model specifications (see Chapter 3). However, scoring item responses ordinally relative to a model about student thinking implies, at a minimum, that the underlying latent variable represented by the LP can also be ordered. I use the divide-by-total specifications of the NRM and PCM because they can reveal violations of the intended ordering of levels unlike the difference model specifications.

4.2.1. Functional Forms of the NRM and PCM. Bock (1972) developed the NRM by extending choice models for binary response data to describe the probability of choosing among multiple alternatives. The IRF for the NRM (Bock, 1972) expresses the probability of choosing a response option in category k of item i as:

$$P(X_{pi} = k|\theta_p) = \frac{e^{a_{ik}\theta_p + c_{ik}}}{\sum_h^{m_i} e^{a_{ih}\theta_p + c_{ih}}},$$
$$\text{where } \sum a_{ik} = \sum c_{ik} = 0 \tag{4.2}$$

In the NRM, the probability of selecting a response is dependent on the ability of a person, which is represented by θ_p, and the characteristics of the item (a_{ik} and c_{ik}). The respondent-specific parameter θ_p is an estimate of a person's ability relative to an unobserved, latent variable that is assumed to influence responses to the items. The item-specific parameters include the response option specific slopes (a_{ik}) and intercepts (c_{ik}). Bock imposed the identification constraint that all a_{ik} and c_{ik} parameters within an item i must sum to zero to obtain estimates for the item parameters. Bock (1972, 1997) provides a complete derivation of the NRM.

An alternative specification of the NRM expresses Bock's response option slopes (a_{ik}) as "scoring functions." Chalmers (2012, 2020) derives the NRM from a multidimensional IRT model (Reckase, 2009) that distinguishes between item specific slopes (a_i), response option specific slopes (ak_{ik-1}), and response option intercepts (d_{ik-1}):

$$P(X_{pi} = k|\theta_p) = \frac{e^{a_i \times (ak_{ik-1}\theta_p) + d_{ik-1}}}{\sum_{k=1}^{m_i} e^{a_i \times (ak_{ik-1}\theta_p) + d_{ik-1}}},$$
$$\text{where } d_{i0} = 0,\ ak_{i0} = 0, \text{ and } ak_{ik-1} = m_i - 1 \tag{4.3}$$

The response option slope and intercept parameters from the Bock (1972) and Chalmers (2012, 2020) NRM specifications are linearly related to one another. The main differences between equations 4.2 and 4.3 are in the identification constraints and how the slopes are represented. The advantage of the Chalmers' specification is that the response option slope parameters can be compared to the expected scores for each category since the slope for the first category is constrained to be 0 ($ak_{i0} = 0$) and the slope for the highest category is constrained to be the number of categories minus 1 ($ak_{ik-1} = m_i - 1$). For this reason, the ak_{ik-1} parameters are also

called scoring functions. Additionally, the item-specific portion of the slope (a_i) characterizes the average discrimination of an item that allows comparison of slopes across all items.

Masters (1982) developed the PCM as an extension of Andrich's (1978) rating scale model so that it could be used with items that award partial credit for success on an item. The PCM fits within a general taxonomy of models that share a property Rasch (1960) described as specific objectivity, which essentially means that comparisons among persons should not depend on the choice of items and comparisons among items should not depend on persons. The IRF for the PCM expresses the probability of choosing a response in category k of item i as:

$$P\left(X_{pi} = k|\theta_p, x = k \text{ or } k'\right) = \frac{e^{\sum_{k=0}^{x}(\theta_p - b^*_{ikk\prime})}}{\sum_{r=0}^{m_i}\left(e^{\sum_{k=0}^{r}(\theta_p - b^*_{ikk\prime})}\right)}, \quad \text{where } \sum_{k=0}^{0}(\theta_p - b^*_{ikk\prime}) = 0. \tag{4.4}$$

Like the NRM, the PCM expresses the probability of selecting a response option as dependent on the ability of a person (θ_p) and characteristics of the item ($b^*_{ikk\prime}$). However, the PCM intersection parameters ($b^*_{ikk\prime}$) are the points on the latent trait continuum at which two consecutive response category curves (i.e., those for categories k and k') intersect. In this specification, $k' = k - 1$ indicating the estimated probabilities are dependent on a response being in category k or $k - 1$. The constraint used to identify the model sets the difference between the estimated ability of a person (θ_p) and the category intersection parameter for the lowest category (b^*_{i10}) equal to 0.

Category intersection parameters ($b^*_{ikk\prime}$) can be calculated from the NRM slope and intercept parameters (a_{ik} and c_{ik}) using the following equation (Bock, 1997):

$$b^*_{ikk\prime} = \frac{c^*_{ikk\prime}}{a^*_{ik\prime k}} = \frac{c_{ik} - c_{ik\prime}}{a_{ik\prime} - a_{ik}}, \tag{4.5}$$

where $b^*_{ikk\prime}$ is the crossing point between the response option curves for categories k and k', the c_{ik} and $c_{ik\prime}$ parameters are the corresponding NRM response option intercepts for those

categories, and the a_{ik} and $a_{ik'}$ parameters are the corresponding NRM response option slopes. When using the NRM, equation 4.5 can be used to calculate the intersections between curves for any two categories (e.g., k and $k-1$, k and $k+1$, or even k and $k+2$) because slopes and intercepts are estimated for every response category. Note that k' in equation 4.5 represents any other category, and it is not constrained to $k' = k-1$ like in the PCM. Equation 4.5 is most useful for LP validation studies when there is a pre-existing hypothesis that describes how the responses should be ordered.

Mathematically, the main differences between the NRM and PCM are in the constraints placed on the slope parameters (a_{ik}'s from equation 4.2 or a_i's and ak_{ik-1}'s from equation 4.3). The NRM parameterizes slopes for each response option within an item. The PCM constrains the NRM response option slopes (a_{ik}'s) to increase by 1 (Bock, 1997). Stated differently, the PCM constrains the item specific portion of the NRM slopes to be constant (e.g., $a_i = 1$) and the response option specific portion of the NRM slopes to be the expected scores for each response category (e.g., $ak_{i0} = 0, ak_{i1} = 1, ak_{i2} = 2$ etc.). This constraint introduces an order among the scores assigned to data modeled with the PCM that is absent from the NRM. When using the PCM, initial evidence for the order of partial credit scores comes from the theory used to design the items. More specifically, multiple-choice distractors that contribute to partial credit scoring should include some aspects of the construct being measured (e.g., by requiring some reasoning associated with the correct response) and demand more of the person than an incorrect response but less than the correct response (Andrich & Styles, 2011). As I discuss more fully later, empirical evidence that PCM intersection parameters are reversed indicates evidence through proof-by-contradiction that there are problems with the design or scoring of the item.

4.2.2. Assumptions of IRT Models. A fundamental assumption of all IRT models is local independence. Local independence means that the only association that exists between item scores is through the j-dimensional latent variable $\boldsymbol{\theta}$. Within any group of examinees all characterized by the same values of $\theta_1, \theta_2, \ldots, \theta_n$, the conditional distributions of the item scores are all independent of each other (Lord & Novick, 1968, p. 361). For local independence to hold, there should be strong evidence to support the dimensionality of a test. The forms of the NRM and PCM introduced earlier both assume unidimensionality (i.e., $j = 1$). When a test has been designed to measure more than one latent variable, multidimensional IRT models (Reckase, 2009) can be used that preserve local independence by specifying more than one latent variable. Yen's (1984, 1993) Q_3 statistic is one way to evaluate the appropriateness of the assumption of local independence. To calculate Q_3, the item and person parameter estimates from an IRT model are first used to generate a response matrix consisting of expected item responses for each person. Next, the deviations between the observed and expected responses are calculated. These deviations are then correlated across respondents to derive the Q_3 statistic between two items. Yen and Fitzpatrick (2006) and Chen and Thissen (1997) recommend flagging values of the Q_3 statistic greater in absolute value than 0.20 for local dependence among items.

When using IRT models in LP validation studies, there are also implicit assumptions that can be difficult to evaluate. First, selecting any IRT model assumes that the functional form of the model is an appropriate way of characterizing the data. Analyzing the fit of empirical data to a model can help identify whether or not the model is appropriate. Second, using IRT to analyze data assumes that the latent variable θ_p has a quantitative structure. Michell (1997) argues that this assumption is seldom made explicit or carefully evaluated. Applied LP researchers are typically more concerned with the instrumental task of creating hypotheses of how a latent

variable might be ordered with the implicit assumption that the latent variable is quantitative, and then analyzing item parameters relative to the hypothesis to evaluate assumptions about the psychological attribute. Conceptually, the attribute characterized by an LP is assumed to be a latent quantity, and the levels are considered important milestones along the scale. I proceed in this direction, recognizing that testing hypotheses about the ordering of levels in an LP using OMC items does not necessarily constitute evidence that the latent variable characterized by the LP has a quantitative structure.

4.2.3. Properties of IRT Models. All IRT models have the property of parameter invariance, which is the property that item characteristics and ability estimates are not dependent on the test or the group of respondents. Parameter invariance makes it possible to create item banks or to select items that closely match a respondent's ability in computer adaptive testing, but it only applies when empirical data can be shown to adequately fit an IRT model. If an IRT model fits the data, then parameter invariance should hold. As a check, one can randomly split the data into two subsamples and separately calibrate the item and person parameters. One would expect to find strong correlations among the parameter estimates. Weak correlations would indicate a violation of the parameter invariance property, suggesting that the one or more central IRT model assumptions have not been met.

Rasch (1960) was particularly interested in models that could measure the progress of individuals even if the instruments used to collect the observations change – what I described earlier as the property of specific objectivity. If a scale can be developed by coupling an LP with a Rasch model, then, in principle, items can be used to create tests that provide meaningful information about what students have learned during the course of instruction. The resulting

scale produced by applying a Rasch model may have equal interval properties (Briggs, 2013; Domingue, 2013) that can be used to define units relative to the levels of the LP (Briggs, 2019). Because these properties only emerge when a model within the Rasch family fits the data (i.e., when response functions have a common slope across items), researchers have proposed numerous fit statistics for Rasch models (see Fischer & Molenaar, 1995) that, under certain conditions can detect violations of this assumption. Infit and outfit mean square statistics (Smith, 1991; Wright & Masters, 1982) are two of the most widely used fit statistics, and they provide information about whether or not it is plausible to assume that all items are equally discriminating (Wu & Adams, 2013). The outlier sensitive fit (outfit) is sensitive for people who respond incorrectly to very easy items or who respond correctly to very hard items (i.e., the outliers). The information weighted fit (infit) removes the impact of outliers. Smith (1991) and Wright and Masters (1982) describe the calculation and properties of these fit statistics. The infit and outfit statistics have an expected value of 1, and the corresponding t-statistics give an indication of whether or not the differences are larger or smaller than expected by chance.

4.3. Interpreting Item Parameters from Polytomous IRT Models.

Chapter 3 reviewed how other researchers have interpreted the order of item parameters from psychometric models relative to models of student thinking, illustrating the prevalence of graphical analysis of response option curves or interpretation of the order of cumulative threshold parameters. Below, I discuss how Bock and Masters, respectively, originally envisioned interpreting the order of NRM and PCM item parameters. Their approach is agnostic to whether a model of student thinking was used to design the items. Item parameters from

polytomous IRT models can be useful sources of LP validity evidence when the items have been designed so that the response options are aligned to a hypothesis about ordered levels in a model of student thinking. OMC items are one example of items that have this characteristic. As I show below, there are two competing perspectives of how empirical polytomous IRT model item parameters can be interpreted to provide information about the structure of the latent variable.

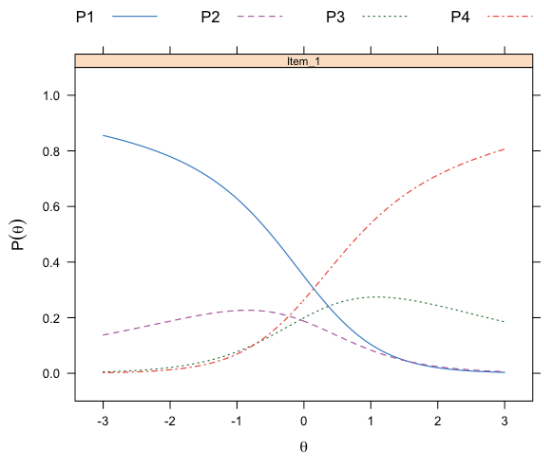
4.3.1. Response Option Slope Parameter Interpretation. One approach to evaluating the order of LP levels is to interpret the empirical order of NRM response option slope parameters (the a_{ik}'s in equation 4.2). Consider the item displayed in Figure 4.2, which is reproduced from Bock (1972, 1997)[11]. This item is from an achievement test that was designed to assess comprehension of vocabulary words. The item asks respondents to identify the correct definition for the word "domicile," which is a legal term for a person's residence (category 4). Bock grouped the distractors "legal document," "hiding place," and "family" together (category 2) because few individuals selected those options. Students that did not provide a response were assigned to the lowest category (category 1) and those who selected "servant" were assigned the second highest category (category 3). The test was not designed using a model of student thinking, and Bock does not provide information about how the distractors were constructed. In Figure 4.2, I also include estimates of the scoring functions (ak_{ik-1} parameters) and some category intersection ($b^*_{ikk\prime}$) parameters for the domicile item.

[11] To generate the empirical response option curves included in this section, I input the response option slope and intercept parameter estimates from the literature (Bock, 1972, 1997; Masters, 1982) into the simdata function within the mirt package (Chalmers, 2012, 2020) for the free, open-source programming language R (R Core Team, 2015). The simdata function simulates a dataset based on input values for the item parameters in an IRT model. After simulating a dataset, I then fit an IRT model to the data to recover the item parameter estimates.

Select the definition for the word. Do not guess. If you do not know the correct answer, omit the item.

DOMICILE

(1) No Response
(2) Legal document, Hiding place, Family
(3) Servant
(4) Residence



Cat.	c_{ik}	a_{ik}	ak_{ik-1} [95% C.I.] ($a_i = 0.64$)	$b^*_{ikk'}$
(1)	0.37	-0.97	0	-
(2)	-0.26	-0.56	0.63 [0.48,0.77]	1.53
(3)	-0.20	0.57	2.38 [2.24,2.53]	-0.05
(4)	0.09	0.96	3	-0.28

Figure 4.2. Example NRM response option curves.

The NRM item parameters are quantitative estimates of important features of the response option curves. The intercept parameters (c_{ik}) are estimates of the response option curve inflection points, or the points at which the curve changes direction. Consider the intercept for the curve associated with the correct response ($c_{i4} = 0.09$). At $\theta = 0.09$, the curve for response option (4) shifts from being concave up to being concave down. The slope parameters (a_{ik}) are the slopes of tangent lines at the inflection points. For the correct answer choice, $a_{i4} = 0.96$. The line tangent to the inflection point at $\theta = 0.09$ has a slope of 0.96. The scoring functions (ak_{ik-1}) can be derived from the a_{ik} parameters by adding the estimate of the lowest category to all the a_{ik} parameters and then scaling the resulting parameters so that the highest category equals the number of categories minus 1($m - 1$). To calculate intersection parameters ($b^*_{ikk'}$) from the NRM intercept (c_{ik}) and slope (a_{ik}) estimates using equation 4.5, one needs to

determine which two response curves should be compared. The intersection parameters in Figure 4.2 assume that the categories are 1-4 are ordinal rather than nominal, and correspond to the points along the curve where the curves for categories 1 and 2 intersect ($b^*_{i21} = 1.53$), 2 and 3 intersect ($b^*_{i32} = -0.05$), and 3 and 4 intersect ($b^*_{i43} = -0.28$).

In his presentation of the NRM, Bock emphasized the interpretation of response option slopes rather than intercepts, scoring functions, or intersection parameters. In interpreting NRM parameters for the domicile item, Bock (1972) states:

> As inspection of the derivative of the category characteristic with respect to θ shows, the largest algebraic value of $\hat{a}_i$ corresponds to the category with a monotonic increasing curve; whereas the smallest algebraic value corresponds to the category which is always monotonic decreasing. Curves of the remaining categories are non-monotonic and have a maximum at some finite value of θ…we see that the response "servant," although incorrect, is largely indicative of positive ability. This contrasts with the curve for category 2, which shows that response in any remaining alternative of this item is similar to omit in its implication for ability (p. 47-48).

Bock's observations regarding the similarity of the "servant" and "residence" choices and the "omit" and "legal document/hiding place/family" options, can also be observed by comparing the magnitude of the differences between the slope parameters. That is, the differences between a_{i1} and a_{i2} ($\Delta = 0.41$) and a_{i3} and a_{i4} ($\Delta = 0.39$) are smaller than the difference in slopes for the two intermediate categories a_{i2} and a_{i3} ($\Delta = 1.13$). In later writing, Bock (1997) referenced two studies that found NRM response option slopes are ordered in data that can be analyzed using a graded response model, which is a difference IRT model that assumes the response data

has an ordinal structure. For Bock, the principal goal of recovering an empirical order among the response option curves was to support the use of difference model IRT specifications.

In LP validation studies, an alternative goal of analyzing the slope estimates can be to compare the empirical ordering to a hypothesized ordering. This approach compares both the order and magnitude of estimated scoring functions (ak_{ik-1} parameters) to the empirical order. Consider the scoring functions for the "domicile" item in Figure 4.2. If a partial credit scoring scheme was appropriate for this item, the expected values for the scoring functions would be $ak_0 = 0$, $ak_1 = 1$, $ak_2 = 2$, and $ak_3 = 3$. Since the scoring functions for the middle categories (ak_1 and ak_2) are freely estimated, the empirical estimates for these parameters can be used to explore hypotheses about whether these categories follow an expected ordering. This feature can be attractive for exploring hypotheses about the ordering of levels in the "messy middle" of LPs. For the "domicile" item, we might be concerned with imposing constraints on the scores for the middle two categories (i.e., using a partial credit scoring scheme) since the empirically derived scoring functions are statistically different from the expected values of 1 and 2 ($ak_1 = 0.63$ and $ak_2 = 2.38$). This method presumes that the number of scoring categories are known in advance, and it may not be useful when it is unclear whether or not categories should be collapsed.

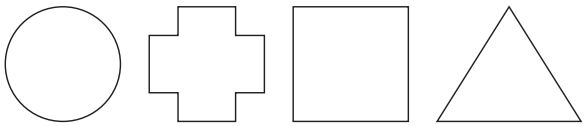
Inspecting the shapes of the NRM response option curves for the middle two categories reveals some additional challenges with using a partial credit scoring scheme for the "domicile" item. A content expert might favor a partial credit scoring scheme that recognizes students who are using alternative resources to solve this item. For example, a student who recognizes that domicile is a legal term may choose "legal document," another who may think that the prefix "dom" indicates a specific domain or location might select "hiding place," and a respondent who thinks the prefix "dom" relates to domestic matters might choose "family." Awarding these

students 1 point instead of 0 might indicate that they have some of the latent attribute being measured. Initial inspection of the response option curves could support using these scores since students have non-zero probabilities of choosing distractors associated with the middle two categories. However, the distractors associated with the middle two categories are never the most probable response for a student across the latent ability continuum. Students with $\theta = 0$ have a 0.20 probability of selecting distractors associated with the second lowest category but a 0.80 probability of providing no response to the item. Analyses involving the content of the item are most defensible when the items have been constructed using clear design specifications.

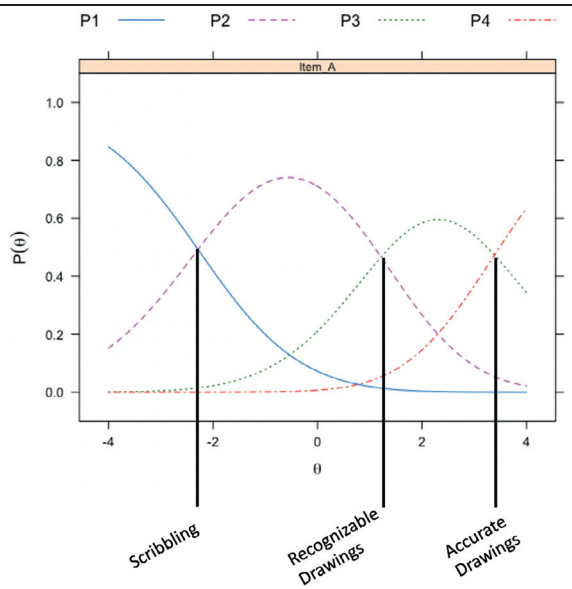
4.3.2. Category Intersection Parameter Interpretation. An alternative approach to evaluating the order of LP levels is to interpret the empirical order of category intersection parameters (the $b^*_{ikk\prime}$'s in equations 4.4 and 4.5). Consider the item displayed in Figure 4.3, which is reproduced from Masters (1982). This item is from a test designed to identify fine-motor learning problems among preschool children. The item asks students to copy a circle, cross, square, and triangle with a pencil. Children can earn three points for each shape, and four performance categories are defined from these points. A child who makes no response is placed in the first category. To be located in the second performance category, a child needs to just mark the paper (e.g., scribbling). Children in the third performance category are able to create recognizable copies of at least two of the four shapes, and this category defines a much higher level of fine-motor functioning than scribbling. To be placed in the highest performance category, students must accurately copy at least two shapes and create recognizable copies of the remaining two shapes. The highest performance category requires a relatively high degree of

coordination and defines an even higher level of functioning than the preceding categories. Like Bock, Masters does not provide information about how the items were designed.

Copy each shape with a pencil.



(1) No response.
(2) Student marks the paper once in response to any of the four shapes.
(3) Student produces recognizable copies of at least two shapes.
(4) Student produces accurate copies of at least two shapes and recognizable copies of two shapes.



Categories	$b^*_{ikk'}$
(1) and (2)	-2.29
(2) and (3)	1.22
(3) and (4)	3.39

Figure 4.3. Example PCM response option curves.

Masters interprets the magnitudes of the PCM category intersection parameters, which he calls "steps," relative to their locations along the latent ability scale. For the item displayed in Figure 4.3, Masters (1982) states:

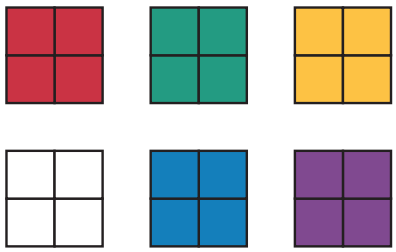
> The probability curves … [show] a wide range of abilities for which 1 is the most probable score … The first step is very easy and the second step is relatively hard, and so, for children between [-2.29 logits and 1.22 logits], the most probable outcome is completion of only the first step (scribbling) (p. 169).

Inspection of the magnitudes of the category intersection parameters in Figure 4.3 reveals that they are ordered along the latent ability continuum with $b^*_{i21} = -2.29 < b^*_{i32} = 1.22 < b^*_{i43} = 3.39$. The curves indicate that these are the values of θ where the probability of being in a higher performance category becomes more likely than a lower category. At $\theta = -2.29$, for example, students become less likely to provide no response to the item and more likely to mark the paper at least once. Similarly, at $\theta = 1.22$, students are less likely to just mark the paper and more likely to produce recognizable drawings. When the intersection parameters are ordered, distances between them can be used to provide substantive meaning to the latent ability scale by dividing it into regions where a response in each category is most probable. For a large portion of the ability scale ($-2 < \theta < 1$), scribbling is the most probable response. Since scribbling is a very easy fine-motor action to perform, only students who do not cooperate or who have difficulty holding a pencil would be not be placed in this performance category. Masters recommended improving the item by making the scribbling performance category more difficult to attain.

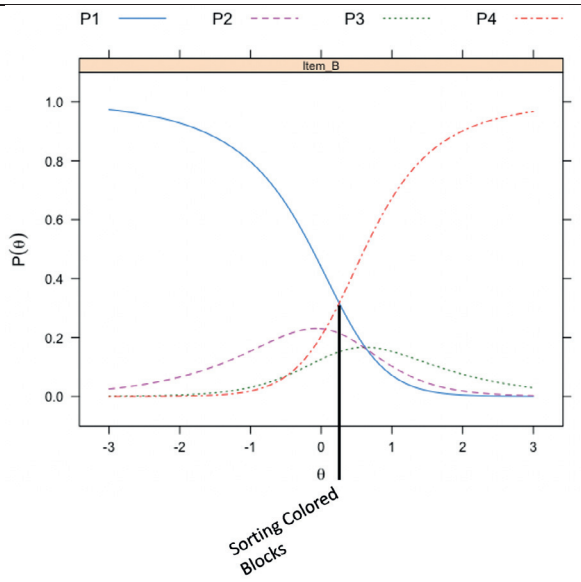
In LP validation studies with OMC items, comparing the magnitude of category intersection parameters provides a way to evaluate the quality of the developmental hypothesis used to design and score the items. Consider another item from Masters (1982) reproduced in Figure 4.4. This item is from the same fine-motor test as the item in Figure 4.3, but it asks students to arrange 24 blocks into six squares of four blocks each consisting of the same color. Students earn one point for identifying and grouping any four blocks of the same color and another point if the four blocks are arranged in a square, and these points are used to define the four performance categories listed in Figure 4.4. Like the item in Figure 4.3, the first performance category indicates non-cooperation or extreme difficulty with the task. To be placed in the second category, a student needs to find and group four blocks of the same color or

arrange some blocks into squares. The third category requires that the student groups all of the blocks into squares, while the fourth requires that at least three of the squares consist of the same color. Note that the latter three performance categories are substantively different from those for the item in Figure 4.3 and may require a different set of cognitive or fine-motor skills.

Each child is given a pile of 24 blocks and asked to arrange the blocks into six four-block squares, each of a different color.



(1) No response.
(2) Student identifies and groups any four blocks of the same color.
(3) Student groups the remaining blocks into squares.
(4) Student arranges at least three color groups into squares.



Categories	$b^*_{ikk'}$
(1) and (2)	0.65
(2) and (3)	0.60
(3) and (4)	-0.48

Figure 4.4. Example category intersection parameter reversal.

Examining the category intersection parameters ($b^*_{ikk'}$) in Figure 4.4 reveals that they do not follow the expected ordering of $b^*_{i21} < b^*_{i32} < b^*_{i43}$. Masters (1982) interprets the intersection parameters, or "steps," for this item as follows:

> The second step … is not significantly more difficult than the first, making the grouping of only one color a relatively improbable event. A child with a good chance of finding four blocks of the same color has a good chance of finding all blocks of the same color.

> The third step … is estimated to be easier than either of the preceding steps, meaning that every child is estimated to be less likely to complete the second step than to complete the third step *if they reach it* (p. 169).

Masters recommends improving this item by making the second and third steps harder but notes that $\theta = 0.2$ is a meaningful point on the scale because it signals a transition between students being more likely to provide no response to being able to group colored blocks into squares. Reversals like those illustrated in Figure 4.4 provide evidence to falsify the partial credit scoring scheme for an item. Andrich (2013, 2015) demonstrates that these reversals are a property of the assessment data and not an artifact of the model and recommends either revising the item or the scoring to eliminate the reversals. The reversals in Figure 4.4 suggest that the design or scoring or the item should be revised to more clearly distinguish among students' fine motor skills. Note that the domicile item (see Figure 4.2) also contains reversals, but the estimated response option slopes are all ordered providing a different perspective from the NRM on partial credit scoring.

4.4. A Method to Interpret Polytomous IRT Parameters Relative to an LP Hypothesis.

The preceding section illustrated two different perspectives on how to interpret the order of item parameters from polytomous IRT models. The NRM tradition focuses on evaluating the order and relative magnitudes of empirical estimates of response option slope parameters or the scoring functions. The slopes can reveal whether or not there is evidence to support assigning partial credit to responses, and inspection of the response option curves can reveal locations along the latent ability scale where students have a non-zero probability of selecting an answer choice. The PCM tradition constrains the NRM response option slopes and evaluates the order

and magnitudes of category intersection parameters. When the intersections parameters are ordered, they segment the latent ability scale into regions where respondents are most likely to provide a response associated with each of the successive partial credit scores. To the extent that the transitions between the categories are well-defined, the intersection parameters can be used to give meaning to the latent ability scale. If disordered, they indicate a problem with the scoring or design of the item. In the PCM tradition, the latent ability scale is more important than the shapes of the response option curves because it is the latent scale that permits independent comparisons among items or respondents if the data fits the model. The NRM tradition uses the latent ability scale for the purpose of comparing response option curves, but it is more agnostic about how the scale can be used to provide information about what respondents know and can do.

4.4.1. Using the NRM and PCM in Isolation. When the goal of analysis is to provide validity evidence for an LP with OMC items, using a single perspective on parameter interpretation can be challenging. LPs are hypotheses about how student ideas may develop over time, and they can be conceived as an attempt to provide substantive meaning to the latent ability scale. Interpreting response option slopes is helpful for understanding how scores could be assigned to answer choices, especially scores associated with levels in the "messy middle of LPs," but an exclusive focus on slope parameters and response option curves neglects the potential for the development of a scale that can be mapped to the LP hypothesis. Additionally, using statistical criteria in the interpretation of response option slopes assumes that there is enough power to detect statistically significant differences from the expected scores. On the other hand, inspection of intersection parameters can identify substantively meaningful regions of the latent ability scale, but the presence of intersection parameter reversals only signal

problems with the design or scoring of the item. Reversals provide limited evidence about how the item or the scoring scheme could be revised.

Consider a test consisting of OMC items, where all items can be scored in four response categories (1, 2, 3, or 4) relative to four levels of an LP. The numbers 1, 2, 3, and 4 represent nominal categories when using the NRM and ordered categories when using the PCM. The test developer hypothesizes that the levels of the LP are ordered but is unsure if this ordering would appear empirically. Hypotheses about the order of levels would be transferred to the scoring scheme for the OMC items, as described in Chapter 2. Figure 4.5 illustrates how incorrect scoring relative to the "true" ordering of the latent variable could impact PCM and NRM item parameters. Let us assume that the "true" progression of ideas is that level 3 represents a slightly less sophisticated understanding than level 2 (i.e., respondents require less ability to be placed in level 3 compared to level 2). If the LP hypothesis reversed the order of levels 2 and 3, this reversal would be reflected in the scoring scheme since levels 2 and 3 would be reverse scored.

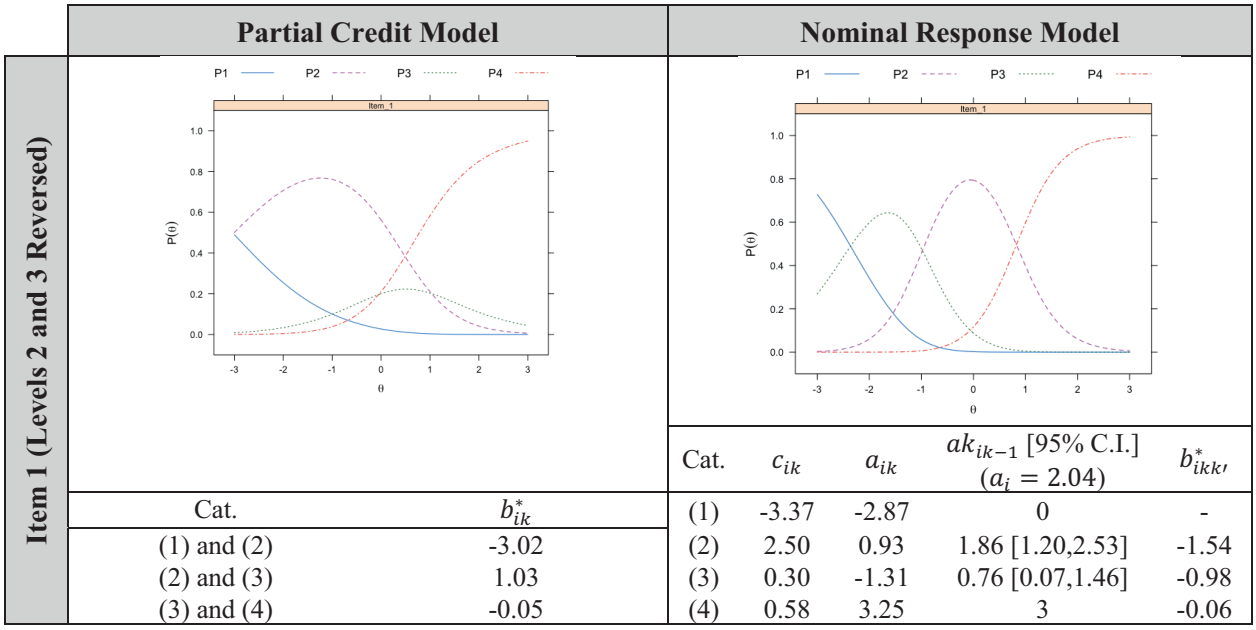


Figure 4.5. Illustrative examples of items modeled using the NRM and PCM.

If we only fit the PCM to the data, we would observe category intersection parameter reversals that signal a problem with the item. For item 1, the reversal appears because $b^*_{i32} = 1.03 > b^*_{i43} = -0.05$. This result might lead one to conclude one of the following about item 1: a) the transition from level 2 to level 3 is too easy and should be made more difficult; b) the transition from level 3 to level 4 is too easy and should be made more difficult; or, c) scores for levels 2 and 3 should be combined since the response associated with level 3 does not discriminate well along the latent ability scale unlike the response associated with level 2. The above conclusions all signal that something is wrong with either the scoring or design of the item, but it can be difficult to pinpoint the cause for the reversal if using only the PCM.

The NRM response option curves illustrate that the PCM reversals are caused by the way scores are assigned to the data. The response option curves clearly indicate that different answer choices are more probable at different regions of the latent ability scale. The order of response option slopes ($a_{i1} = -2.87 < a_{i3} = -1.31 < a_{i2} = 0.93 < a_{i4} = 3.25$) suggests that the middle two categories should be reversed to reflect the "true" order of LP levels. In contrast to the domicile item in Figure 4.2, rounding the scoring functions (ak_{ik-1}) for each category to the nearest whole number recovers the "true" scoring that should be used with the data (e.g., $ak_{i1} = 2$ and $ak_{i2} = 1$) since the scoring functions are not statistically significant. The NRM category intersection parameters calculated under both the mis-specified LP hypothesis ($b^*_{i21} = -1.54 < b^*_{i32} = -0.98 < b^*_{i43} = -0.06$) and using the true order of categories ($b^*_{i31} = -2.35 < b^*_{i23} = -0.98 < b^*_{i42} = 0.83$) are both ordered, but only the latter divides the latent ability scale into regions where selecting response options associated with each category are most likely.

Although the NRM can be used to derive an empirical partial credit scoring scheme, content experts should confirm that the estimated scores would be appropriate to use given the

LP hypothesis. For example, the NRM scoring functions for item 2 suggest that the lowest score should be assigned to category 2 and the second lowest score to category 1. However, if it is implausible that category 2 should be scored lower than category 1 (e.g., if category 1 indicates "no response" compared to a substantively meaningful category 2 response), then the design of the item may be problematic. Category intersection parameters for the NRM response option curves can also help content experts understand the meaning of the latent ability scale relative to the LP hypothesis. However, calculating NRM category intersection parameters requires one to decide on an upper and lower category (see equation 4.5). This decision may not always be straightforward, especially when using the NRM to explore how scores could be assigned to response options (e.g., see the item in Figure 4.2). Additionally, using category intersection parameters to impose meaning on the latent ability scale in LP validation studies may be less useful for the NRM compared to the PCM since the NRM tradition does not emphasize the potential for creating a scale with desirable properties.

4.4.2. Using the NRM and PCM Together with OMC Items. The NRM can help determine the best way to assign partial credit scores to answer choices relative to a model of student thinking because it imposes the fewest restrictions on the item parameters, treating each answer choice as a nominal category. When NRM response option curves are interpreted relative to the latent ability scale using a model of student thinking, it becomes possible to explore how scores could be assigned to answer choices to reflect variable performance along the latent continuum. The PCM can then be used to explore whether it is possible to construct a scale that can be interpreted relative to the levels in the hierarchical model of student thinking. Figure 4.6 presents a method to analyze NRM and PCM item parameters relative to a model of student

thinking. This approach is most useful when the model of student thinking includes a hypothesis of how ideas develop over time (e.g., the model is an LP), and this hypothesis is used to design items that can be scored relative to the categories (e.g., OMC items).

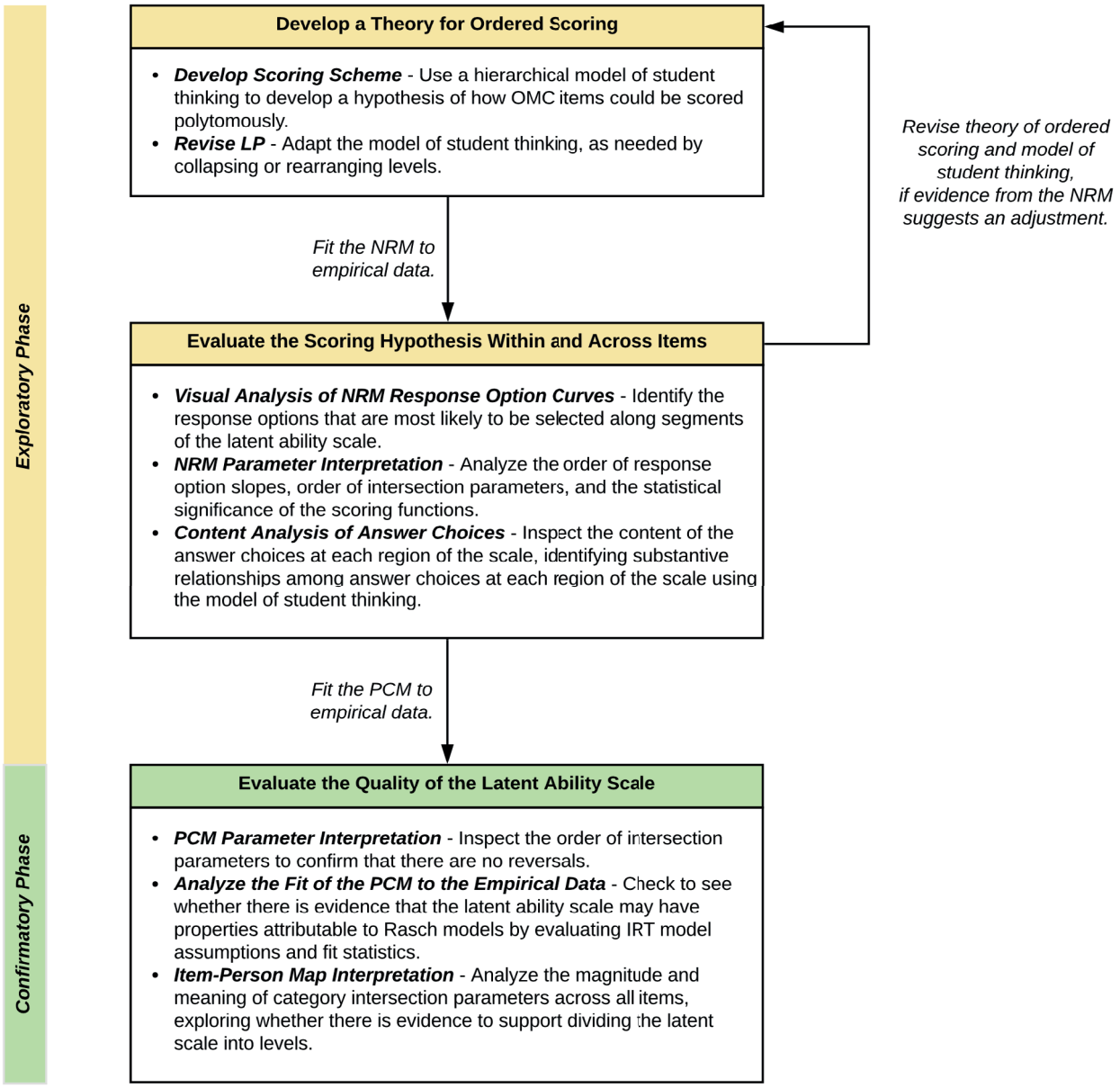


Figure 4.6. A method to analyze OMC items relative to a model of student thinking.

The first two stages explore how scores can be meaningful relative to both the model of student thinking and the latent ability scale. The first stage is to develop a hypothesis for how OMC answer choices should be scored relative to the hierarchical model of student thinking. If the OMC item design specifications are clear, then this hypothesis clearly connects partial credit scores to the model of student thinking. The second stage evaluates the scoring hypothesis both within and across items to identify whether the partial credit scores assigned to response options divide the latent ability scale into regions where successive ordered response options are most probable. Interpreting the response option curves along with the empirical order, statistical significance, and magnitude of differences among the estimated NRM slope parameters can help confirm whether the partial credit scores are appropriate to use for each item.

To analyze the content of the answer choices relative to the latent ability scale, I use the descriptions for disciplinary core ideas, crosscutting concepts, and scientific and engineering practices detailed in *A Framework for K-12 Science Education* (NRC, 2013) to clarify the conceptual ideas about atoms described by the facet cluster. The framework describes in detail grade-band endpoints for disciplinary ideas associated with the structure and properties of matter (PS1.A). The descriptions for grades 2 and 5 are most similar to the descriptions contained in the facet cluster for atoms. For example, "matter can be described and classified by its observable properties" and "a great variety of objects can be built up from a small set of pieces" (grade 2 endpoint) and "matter of any type can be subdivided into particles are too small to see, but even then the matter still exists and can be detected by other means" and "the amount (weight) of matter is conserved when it changes form, even in transitions in which it seems to vanish" (grade 5 endpoint). I begin with the disciplinary core ideas because they are the most concrete, detailed representations of the ideas described in the facet cluster. This approach also permits me to

leverage my content expertise in chemistry to understand how the more abstract crosscutting concepts and scientific and engineering practices might be layered onto the facet cluster.

Inspection of the disciplinary ideas reveals that they tie together concepts related to the crosscutting concept of scale, proportion, and quantity and the scientific practice of using mathematics and computational thinking. Emphasis on observable properties of matter, building objects from smaller pieces, and visible and invisible components of matter in the descriptions of the disciplinary core ideas tentatively apply ideas related to the crosscutting concept of scale, proportion, and quantity. The *Framework* suggests that this crosscutting concept can be used to design scientific phenomena and possibly assessments: "in considering phenomena, it is critical to recognize that what is relevant at different measures of size, time, and energy and to recognize how changes in scale, proportion, or quantity affect a system's structure or performance" (p. 84). The grade 2 endpoint related to building up larger objects from smaller ones and the grade 5 endpoint related to understanding that the amount of matter is conserved when it changes its form both require a mathematical thinking. The *Framework* provides a rough sketch of a progression for this practice that recognizes that "as soon as students learn to count, they can begin using numbers to find or describe patterns in nature" and then "a significant advance comes when relationships are expressed using equalities first in words and then in algebraic symbols." Although the facet clusters were designed prior to the development of the *Framework*, the three dimensions of the framework can be used to improve our understanding of the model of student thinking, specifically how the ideas in the facet cluster may develop over time.

Integrating the science content into these analyses can help resolve lingering questions about whether or not answer choices should be collapsed for the purpose of scoring or whether there is enough power to interpret the statistical significance of the scoring functions. The latter

situation is especially problematic in LP research where investigations are conducted in classrooms with relatively small samples of students (i.e., <1000 cases). As the number of students selecting a response option for an item decreases or the number of item parameters increases (e.g., by adding more response options), it becomes more difficult to accurately estimate the NRM parameters and evaluate their statistical significance. See Appendix A for a more detailed discussion of sample size issues related to estimation of the NRM parameters. These first two stages in Figure 4.6 are meant to be iterative and exploratory, flexibly connecting interpretation of the item parameters to the observations and to the model of student thinking as envisioned by the assessment triangle. The outcome of the exploratory phase is a partial credit scoring scheme that then be used with the PCM.

The third stage is a more confirmatory phase that evaluates the quality of the latent ability scale relative to the hierarchical model of student thinking. I consider this stage to be more confirmatory because the PCM imposes stricter requirements on the data than the NRM for the purposes of producing a latent ability scale with desirable properties (i.e., specific objectivity). Any category intersection parameter reversals like those present for the items in Figure 4.2 and 4.4 would signal that the partial credit scoring scheme developed in the exploratory phase may either need revision or that the items with the reversals need to be rewritten. Evaluation of the fit, assumptions, and properties of the PCM can help distinguish whether there are additional problems with the items. Item-person maps, which are also called Wright Maps in the Rasch tradition, were discussed extensively in Chapter 3, and they plot item parameter estimates and person ability estimates along the same scale. I use PCM category intersection parameters on the item-side of the map, rather than PCM threshold parameter estimates used in most of the item-person maps in Chapter 3. Item-person maps facilitate comparison of intersection parameters

across all items on the test, and they can provide information about whether it is reasonable to divide the latent ability scale into regions corresponding to levels of the LP. Taken together, this use of the PCM can reveal whether or not there is reasonable evidence to map a quantitative latent ability scale onto a qualitative hierarchical model of student thinking like an LP.

Using the NRM in conjunction with the PCM has the potential of providing stronger evidence to support the validity of OMC test score interpretations relative to hierarchical models of student thinking. The strength of using the NRM to analyze OMC items is best realized when all answer choices for an item can be analyzed independently relative to the model of student thinking. That is, answer choices should not be collapsed together prior to fitting the NRM because OMC item designers typically have an initial hypothesis that the answer choices capture meaningfully distinct ideas. The major limitations of using the NRM to analyze data are that it requires a large amount of data to estimate item parameters and the parameters can be difficult to interpret when answer choices are not collapsed into meaningful categories prior to analysis. Collecting large amounts of student response data for LP validation could be feasible when assessments that are thoughtfully designed are administered in digital environments. The major difference between the use of the PCM promoted in Figure 4.6 and prior research is in the interpretation of category intersection parameters rather than the cumulative threshold parameters. PCM threshold and intersection parameters both assume that the scores assigned to the response categories are ordinal. However, the threshold parameters reflect cumulative probabilities, while the intersection parameters reflect conditional probabilities. Comparing the magnitudes of adjacent threshold *or* intersection parameters can reveal information about the distinctiveness of an item's response categories. However, threshold parameters will always be ordered according to the scoring scheme, in contrast to intersection parameters that can be

reversed if there are problems with the assumption that the scores are ordinal. In the next chapter, I introduce empirical data from the online Diagnoser assessment system (Thissen-Roe et al., 2004), which I use to provide an empirical illustration of the novel method for LP validation introduced in this chapter.

Chapter 5

Data

This chapter introduces the empirical data that I use to illustrate how the NRM and PCM can be used to improve interpretations of test scores in LP validation studies. The first section introduces the design and intended use of the online Diagnoser assessment system (Thissen-Roe et al., 2004), which is the digital platform used to collect the data. The Diagnoser data is well-suited for an LP validation study because: these data were collected using tests consisting of multiple-choice items with response options aligned to statements of student thinking; the model of student thinking used to develop the items includes a hypothesis about the order of student ideas; and, a relatively large amount of student responses were collected in a digital environment permitting adequate estimation of NRM parameters (see Appendix A). The second section describes how the data were prepared for analysis and presents descriptive statistics. The third section describes the software that I use to estimate the NRM and PCM parameters.

5.1. The Diagnoser Assessment System.

Diagnoser is an online assessment system[12] that contains instructional resources for science instruction (Thissen-Roe, et al., 2004). Diagnoser provides numerous resources for teachers and students, including assessments (elicitation questions and question sets) that can be integrated with a teacher's curriculum (learning goals and lessons). To use the resources as intended, teachers would align their learning goals with those identified on Diagnoser, administer

[12] All Diagnoser resources can be accessed electronically at the following website: http://www.diagnoser.com/. The descriptions of the components of Diagnoser are taken from those provided on the website.

elicitation questions to surface student ideas about the topic, teach one or more developmental lessons that challenge and build on student thinking, assign a question set to evaluate students' progress toward attaining the learning goal, teach a prescriptive lesson that targets specific problematic ideas identified by the assessment, and then assign a second question set to conclude the unit and assess student thinking. Diagnoser recommends that teachers adapt or use these resources as needed. For example, teachers can choose whether to administer the question sets at the beginning or end of an instructional unit.

Although Diagnoser is almost 20 years old, its content and resources remain relevant for science educators, especially if supplemented to address newer standards. Diagnoser's content consists of resources for disciplinary ideas in biology, chemistry, and physics.[13] These resources include facet clusters (see Figure 2.5 for an example facet cluster), sample lessons, reports that summarize performance on question sets (see Figure 2.8 for an example teacher score report), and functionality to assign and access assessment data online. Diagnoser's learning goals, curricular materials, and assessments were aligned to the *National Science Education Standards* (NRC, 1996) and the *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993). These standards have been replaced with a new vision for science education described in *A Framework for K-12 Science Education* (NRC, 2013) and the *Next Generation Science Standards* (NGSS Lead States, 2013). The development of science assessments for these new standards remains a daunting task because of their complexity. The National Research Council (2014) recommended developing assessment systems that integrate multiple sources of evidence from classroom and monitoring assessments with indicators of students' opportunity to learn. Since Diagnoser question sets are classroom assessments designed

[13] These big ideas include force and motion, energy, waves, properties of matter, the atomic structure of matter, changes in matter, and human body systems.

using a model of student thinking, information from these question sets can be used to provide information about what disciplinary ideas students know and have learned during instruction. For this potential to be realized, there should first be sufficient information to support interpretation of the question set scores relative to a model of student thinking.

5.2. Empirical Data.

The empirical data I use in this dissertation consists of student responses to the items in two Diagnoser question sets designed to assess student thinking relative to a facet cluster for student thinking about atoms (see Figure 2.5). Both question sets consist of distinct multiple-choice items with answer choices that are aligned to facets. Question set 1 consists of 7 items. Question set 2 consists of 9 items, of which 2 are scored conjointly. Appendix B provides the full text of each item. The items and facet cluster were developed by Jim Minstrell and his colleagues (DeBarger et al., 2009). Minstrell shared anonymized student data from 2010-17 with me on November 25, 2017 as a set of comma-separated value (csv) files. Variables included unique teacher and student IDs, the date a student answered the question set, the letter of the multiple-choice response a student selected for each item on the question set, and self-reported ratings of confidence in understanding the material (1 = great, 2 = pretty good, 3 = OK, 4 = not very good, 5 = uncomfortable, and 6 = terrible). These files contained no missing data.

5.2.1. Data Preparation. To prepare the data for analysis, I systematically filtered the data to exclude some student records. First, I removed duplicate student records that occurred when a teacher reassigned the same question set to a class. I retained the oldest record to avoid

test-retest effects that might influence students' responses. Second, I removed student records associated with teachers who were likely exploring Diagnoser rather than using it as a classroom assessment tool. In another empirical study involving Diagnoser data, Steedle (2008) retained data for teachers that assigned at least 20 question sets per year. This decision was appropriate because Steedle was modeling data for multiple facet clusters. Using the same criterion would not be appropriate for my study since it is unlikely teachers would assign the two question sets associated with the atoms facet cluster 20 times a year. Instead, I retained data for teachers who had more than 10 students in a class complete a question set. This decision removed cases where teachers or students were exploring Diagnoser (e.g., through "demo," "test," or "practice" accounts) or using Diagnoser to prepare for non-classroom related activities (e.g., a science bowl). Table 5.1 presents sample sizes for the two question sets and illustrates that about 90% of the original data remained for analysis after filtering the data using these two exclusion criteria.

Table 5.1.

Analytic Sample Sizes.

Analytic Sample	Frequency				% Original Cases Retained
	Original Cases	Duplicates Removed	Exploratory Users Removed	Final Cases	
Test A (Set 1)	5,659	233	363	5,063	89.5 %
Test B (Set 2)	2,626	76	163	2,387	90.9 %
Test C (Same Day)	-	-	-	1,105	-

The sample sizes in the "final cases" column of Table 5.1 comprise the analytic data I use in this dissertation. The analytic data consists of student response to items on three distinct

"tests" designed from the facet cluster for Atom. The first two analytic samples (Tests A and B) consist of student responses to the first ($n = 5{,}063$) and second ($n = 2{,}387$) question sets, respectively. There are no common items across these two tests. Most students did not respond to both question sets. Teachers assigned one question set (either the first or second) to students or assigned both. The third analytic sample consists of the subset of students that responded to the first question set and the second question set (Test C) on the same day ($n = 1{,}105$). Since students likely completed both question sets in succession, I treat the items on these two tests as comprising a "super" test that consists of items from both question sets.

I use the analytic data to explore whether there is evidence to support interpretations of test scores relative to an ordered facets hypothesis. Note that the sequence of facets in Figure 2.5 represents a testable hypothesis about how student understanding of the concept of atoms might develop during instruction. I score the items dichotomously to indicate whether or not a student answered the item correctly (i.e., 1 = correct and 0 = incorrect) and polytomously to convert the multiple-choice answer choices into nominal categories (i.e., 1 = a, 2 = b, 3 = c, etc.). I use NRM to model data for the first two analytic samples (Test A and Test B) to explore how partial credit scores could be assigned to the student response data. I use the third analytic sample (Test C) to explore whether there is evidence to support the creation of a scale using the PCM. In the latter analysis, I use information from the NRM to develop a partial credit scoring scheme.

5.2.2. Descriptive Statistics. One of the first steps in data analysis is to understand the characteristics of the analytic data using descriptive statistics. The left-hand side of Table 5.2 presents the proportion of students selecting each answer choice for Tests A and B. Diagnoser items have 3-7 response options, and all response options had some students select them. The

shaded boxes indicate the answer choices for the correct responses. Comparing these proportions across items reveals that the item difficulty varies within each question set. For example, 37% of students selected the correct response option for the most difficult set 1 item (item 1.5) compared to 87% of students selecting the correct response for the easiest item (item 1.7).

Table 5.2.

Response Option Frequencies.

Item	Percent of Students Selecting Response Option*							Pt. Biserial
	a	b	c	d	e	f	g	
1.1.	1.7	1.9	6.6	6.0	6.6	3.4	73.8	0.33
1.2.	14.1	75.4	3.6	6.9	-	-	-	0.35
1.3.	6.7	6.6	15.0	69.1	2.6	-	-	0.55
1.4.	7.4	6.8	21.2	58.8	5.9	-	-	0.51
1.5.	21.9	7.8	36.7	33.5	-	-	-	0.46
1.6.	9.4	17.3	51.0	19.6	2.8	-	-	0.57
1.7.	6.2	7.3	86.5	-	-	-	-	0.50
n = 5,063								
2.1.	7.0	1.3	2.3	4.7	-	2.9	81.8	0.46
2.2.	4.7	6.8	11.9	71.3	5.3	-	-	0.59
2.3.	11.2	80.1	4.4	4.3	-	-	-	0.47
2.4	14.4	9.1	53.1	23.4	-	-	-	0.55
2.5.	7.4	5.4	9.8	72.2	5.2	-	-	0.66
2.6.a.†	27.0	22.7	50.3	-	-	-	-	0.56
2.6.b.†	13.2	24.9	61.9	-	-	-	-	
2.7.	80.3	8.5	7.2	4.0	-	-	-	0.56
2.8.	8.2	8.8	83.0	-	-	-	-	0.53
n = 2,387								

* Response options associated with the scientifically accurate, or correct, idea are shaded in grey.
† These two multiple choice items are scored together in the dichotomous scoring scheme (see Appendix B).

The last column in Table 5.2 presents the point-biserial correlation between the response vector for an item and the total score on the test. This correlation provides information about how well an item distinguishes among students with different test scores. I calculated total scores

using dichotomized responses and removing the focal item from the calculation. The final column of Table 5.2 indicates that Diagnoser items are moderately discriminating with point-biserial correlations ranging from 0.33 (item 1.1) to 0.66 (item 2.5). There are no problematic items that have very low (< 0.2) or negative point-biserial correlations.

The total scores students obtained on the tests provide information about how students' understanding of the concept of atoms is distributed. To calculate total scores, I took the algebraic sum of the dichotomized scores for each student across all items on the test. Figure 5.1 plots each possible total score value on the x-axis and the frequency of students earning each possible total score on the y-axis. Ideally, students' understanding would have a normal distribution. However, Figure 5.1 illustrates all distributions have negative skew, perhaps because students have been exposed to some instruction prior to taking the assessment. The measures of central tendency (i.e., means and medians) vary across the tests, suggesting that teachers may have been using these tests in different ways. For example, students that answered both question sets on the same day (Test C) have greater total scores, on average, which might suggest that teachers were using the question sets as summative classroom assessments.

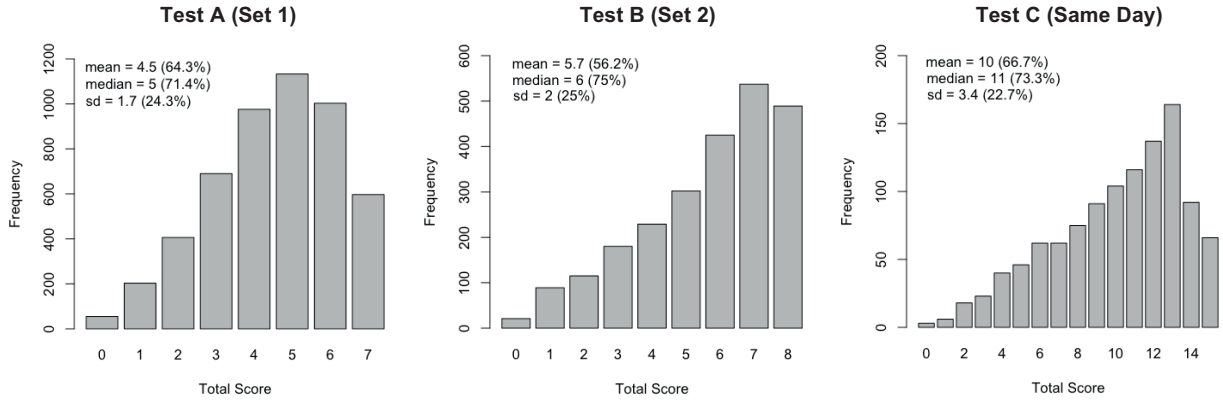


Figure 5.1. Total score distributions.

These descriptive statistics provide some preliminary information about the Diagnoser items and the students who responded to them. The items range in difficulty and discrimination both within and across question sets. The low response option frequencies for some items (see Table 5.2) indicate that some answer choices may not be attractive to students and could be candidates for removal or revision. In operational test development, multiple-choice distractors are considered "poor quality" if less than 5% of students select that distractor. However, low-frequency distractors are sometimes retained by content specialists if the items are considered to be very easy or if they are required to adequately cover the range of content assessed by the item (Gierl et al., 2017). The total score distributions indicate that, on average, students answered more than 50% of the questions on the test correctly. This may be due to teachers assigning question sets to students after exposing students to some instruction.

5.2.3. Test Characteristics. In any psychometric analysis, it is important to quantify the amount of measurement error in a test. If the item responses contain a lot of measurement error, then there is weaker evidence that the scored student response data is a good measure of the latent variable described by the cognitive model. One common estimate of reliability is Cronbach's (1951) α, which depends on the variance of individual items, the variance of the observed total scores, and the number of items on the test. It ranges from 0 to 1, with an estimate of 1 indicating perfect reliability and no measurement error. As the number of items on a test increases, α will also tend to increase. In general, the higher the estimate of α, the better the quality of the data. Very high values of α (e.g., greater than 0.90) are often required for tests with high-stakes uses, but moderately high values may be acceptable for the purposes of research or classroom assessment (Moss, 2003; Smith, 2003). Tests A, B, and C have estimated α values of

0.57, 0.73, and 0.79, respectively. Test A has the most measurement error, perhaps because teachers were using this question set with students in a variety of ways. The other two tests have less measurement error and indicate that the data associated with those two tests is better.

If the facet cluster used to design the Diagnoser items in Tests A-C is measuring a single psychological construct related to understanding the concept of atoms, there should be some empirical evidence that the items are measuring a unidimensional latent variable. Exploratory factor analysis (see Osborne, 2014) is a data reduction technique that examines all the pairwise relationships between individual variables (e.g., items on a test) and extracts latent factors from the measured variables using matrix algebra. I used the item correlation matrices for Tests A-C (see Appendix C) to conduct an exploratory factory analysis, extracting eigenvalues using principal axis factoring since the data have negative skew. One way to determine the number of latent factors the items on the test are measuring to plot the "scree" of successive eigenvalues. Sharp breaks suggest the appropriate number of factors. Figure 5.2 contains the scree plots produced from an exploratory factor analysis of the three tests.

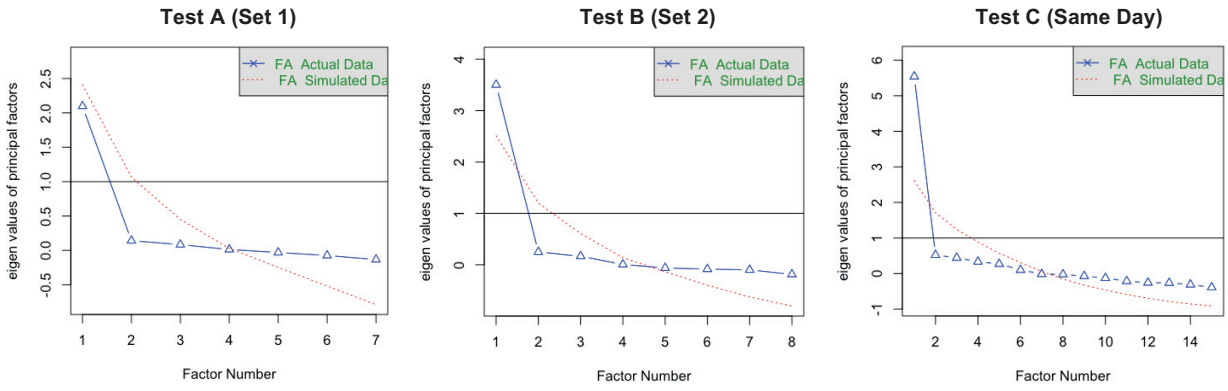


Figure 5.2. Scree plots.

The scree plots in Figure 5.2 all suggest that the tests are measuring a single dominant factor. Each scree plot has a sharp break between the first and second factors. However, the difference in extracted eigenvalues for Test A is smaller than the difference for Test B or Test C. The dotted red lines in Figure 5.2 present the results from parallel analysis, which compares the scree plot for observed data to the scree plot of a random data matrix of the same size as the original. Parallel analysis is sensitive to sample size and can suggest fewer factors for data with large sample sizes since the eigenvalues of the random data are close to 1 (Revelle, 2020). The results from parallel analysis support the conclusion that Tests B and C are measuring a single dominant factor. However, the parallel analysis results for Test A are inconclusive. The software (see Section 5.3) produces the following message after conducting parallel analysis with the data for Test A: "Parallel analysis suggests that the number of factors = 0 and the number of components = NA." Since Test A contains the most data, this result may be due to the sensitivity of parallel analysis to sample size.

The estimates of the reliability and dimensionality of Tests A-C suggest that the tests are of sufficient quality for the purpose of conducting research. Test A has the most measurement error, and it may pose the most challenges in IRT analyses. Consequently, it is important to examine whether or not the data, particularly for Test A, violate the assumptions of the NRM and PCM. The reliability estimates also indicate that the Diagnoser question sets may be acceptable for classroom uses, but they will likely need some improvement to reduce measurement error if these assessments are to be used for other purposes like the monitoring of student achievement. The exploratory factor analysis results suggest that each test is measuring a single dominant latent variable, likely related to students' understanding of the concept of atoms.

5.3. Software.

To conduct all statistical analyses for this dissertation, I use version 4.0.2 ("Taking Off Again") of the free R programming language (R Core Team, 2015). R is an open-source programming language built on the S programming language (Becker, Chambers, & Wilks, 1988). It has an active online community, offering many free tutorials and code examples. R also has free packages for psychometric analyses and strong data visualization functionality. To process, filter, and descriptively analyze the Diagnoser data, I use the packages contained in Wickham's (2019) tidyverse. I use the ltm package (Rizopoulos, 2018) to estimate Cronbach's α, the polycor package (Fox, 2019) to calculate polychoric item correlations (see Appendix C), and the psych package (Revelle, 2020) to conduct exploratory factor analysis.

To estimate the NRM and PCM item and person parameters and model fit statistics presented in the next chapter, I use the mirt package (Chalmers, 2012, 2020). As described in the package documentation, mirt fits a "maximum likelihood (or maximum a posteriori) factor analysis model to any mixture of dichotomous and polytomous data under the item response theory paradigm using either Cai's (2010) Metropolis-Hastings Robbins-Monro (MHRM) algorithm, with an EM algorithm approach outlined by Bock and Aitken (1981) using rectangular or quasi-Monte Carlo integration grids, or with the stochastic EM (i.e., the first two stages of the MHRM algorithm)" (Chalmers, 2020, p. 95). I use the default estimation method, which is the standard EM algorithm with fixed quadrature. The next chapter presents the results from fitting the NRM and PCM to the analytic data.

Chapter 6

Results

This chapter presents the results from fitting the NRM and PCM to the Diagnoser data introduced in the previous chapter. The goal is to illustrate how to use the method for LP validation with OMC items introduced in Chapter 4 (see Figure 4.6). The Diagnoser items have a similar structure to OMC items. There is a clear initial hypothesis for how the facets might be ordered, and this theory is reflected in the numerical codes assigned to the facets. The first section uses the NRM to explore how the empirical order of response options compares to the initial scoring hypothesis. The second section presents a revised version of the scoring hypothesis, including a transformation of the model of student thinking (i.e., the facet cluster in Figure 2.5) into an LP. The third section presents the results from refitting the NRM to the Diagnoser data using the revised scoring hypothesis. In this section, I evaluate the order of the NRM item parameters, explore the assumption of local independence, and examine whether there is evidence to support the property of parameter invariance. In the last section, I fit the PCM to the data and evaluate the quality of the resulting latent ability scale.

6.1. Fitting the NRM to Diagnoser Items to Explore an Initial Scoring Hypothesis.

The first step in analyzing OMC items is to develop an initial hypothesis for ordered scoring that can be explored using the NRM. All of the Diagnoser items I use were designed using the facet cluster for atoms (Figure 2.5). As I described in Chapter 2, facet clusters can be conceptualized as a hierarchical models of student thinking. Facets are grouped into related sets

of ideas, or "macrofacets," as indicated by the first digit of the facet code, and this digit can be used to produce an initial scoring hypothesis to reflect how the latent variable may be ordered. For example, item 1.3 has answer choices mapped to facet 02, 51, or 80, and the numerical order of the facet codes can be used to develop an initial hypothesis for how partial credit could be assigned to this item (i.e., 02 = 2, 51 = 1, and 80 = 0). Some Diagnoser items have options mapped to "unknown" facets, which means assigning partial credit to these responses is not straightforward. For this initial exploration, I assume that answer choices mapped to unknown facets are located somewhere in the messy middle (see Appendix B or Table 6.2 for the initial scoring hypothesis). Below, I evaluate the NRM slope parameters for the Diagnoser items using the data associated with Tests A and B, and then illustrate how analysis of NRM response option curves can be combined with content expertise to revise the initial scoring hypothesis.

6.1.1. Initial Evaluation of NRM Slope Parameters. Table 6.1 presents empirical estimates of the NRM item specific slopes (a_i) and scoring functions (ak_{ik-1}) for the Diagnoser items along with the standard errors (SEs) for the slope parameter estimates. To aid in interpretation, I use the Chalmers (2012, 2020) specification of the NRM (equation 4.3). Appendix D presents estimates for the response option intercepts (d_{ik-1}) and the Bock (1972) NRM slopes and intercepts (a_{ik}'s and c_{ik}'s). All the items are scored nominally and no categories are collapsed, but the scores are ordered such that the scientifically accurate response always has the highest score. The remaining answer choices are ordered using the initial scoring hypothesis derived from the facet codes (see Appendix B or Table 6.2). Recall, in the Chalmers' (2012, 2020) specification of the NRM, the response option slopes (i.e., scoring functions) for the lowest and highest categories are constrained for identification.

Table 6.1.

NRM Slope Parameters from Initial Calibration.

Item	NRM Slope Parameter*									
	a_i	ak_0	ak_1	ak_2	ak_3	ak_4	ak_5	ak_6	ak_7	ak_8
1.1.	0.36	0	1.06	1.48	1.48	3.80	3.58	6	-	-
	(0.03)	(-)	(0.63)	(0.46)	(0.46)	(0.39)	(0.33)	(-)	-	-
1.2.	0.22	0	-3.26	0.63	3	-	-	-	-	-
	(0.02)	(-)	(0.70)	(0.38)	(-)	-	-	-	-	-
1.3.	0.51	0	-0.02	-0.69	1.20	4	-	-	-	-
	(0.05)	(-)	(0.37)	(0.42)	(0.26)	(-)	-	-	-	-
1.4.	0.35	0	0.34	-0.62	2.59	4	-	-	-	-
	(0.03)	(-)	(0.32)	(0.37)	(0.17)	(-)	-	-	-	-
1.5.	0.37	0	-1.65	2.41	3	-	-	-	-	-
	(0.02)	(-)	(0.29)	(0.13)	(-)	-	-	-	-	-
1.6.	0.57	0	1.22	1.70	2.84	4	-	-	-	-
	(0.04)	(-)	(0.21)	(0.17)	(0.13)	(-)	-	-	-	-
1.7.	0.85	0	0.10	2	-	-	-	-	-	-
	(0.06)	(-)	(0.14)	(-)	-	-	-	-	-	-
n = 5,063										
2.1.	0.58	0	1.72	1.93	3.66	0.99	5	-	-	-
	(0.08)	(-)	(0.56)	(0.46)	(0.36)	(0.54)	(-)	-	-	-
2.2.	0.50	0	-0.69	-1.87	0.11	4	-	-	-	-
	(0.05)	(-)	(0.51)	(0.57)	(0.38)	(-)	-	-	-	-
2.3.	0.35	0	-1.75	-0.40	3	-	-	-	-	-
	(0.03)	(-)	(0.55)	(0.47)	(-)	-	-	-	-	-
2.4.	0.58	0	-0.12	1.88	3	-	-	-	-	-
	(0.04)	(-)	(0.23)	(0.13)	(-)	-	-	-	-	-
2.5.	0.57	0	-0.37	-0.79	0.32	4	-	-	-	-
	(0.05)	(-)	(0.40)	(0.45)	(0.34)	(-)	-	-	-	-
2.6.	0.13	0	-9.06	-3.02	-6.92	-5.12	-5.47	4.33	-0.55	8
	(0.01)	(-)	(2.24)	(2.23)	(1.81)	(1.47)	(1.76)	(0.84)	(1.35)	(-)
2.7.	0.60	0	1.45	0.66	3	-	-	-	-	-
	(0.05)	(-)	(0.26)	(0.24)	-	-	-	-	-	-
2.8.	0.68	0	-0.11	2	-	-	-	-	-	-
	(0.06)	(-)	(0.22)	(-)	-	-	-	-	-	-
n = 2,387										

* Standard errors are presented below the parameter estimates.

Table 6.1 illustrates that the number of response options is highly variable across items, ranging from a minimum of 3 choices (items 1.7 and 2.8) to a maximum of 7 or 9 (items 1.1 and

2.6). Items with the smallest number of possible answer choices tend to have smaller SEs for the estimated scoring functions compared to items with more response options. This result is due to the decreased number of respondents selecting response options for items with more possible choices (see Appendix A). Consider item 2.6. This item has large SEs and the most response options. The SE for ak_1 for item 2.6 is 2.24, which is the largest SE in Table 6.1, and this is due to just 36 respondents selecting choice 1. Despite challenges with precisely estimating the scoring functions (ak_{ik-1}'s), the item specific slopes (a_i's) have very small SEs. The estimated values of the item specific slopes vary considerably across items for each analytic sample.

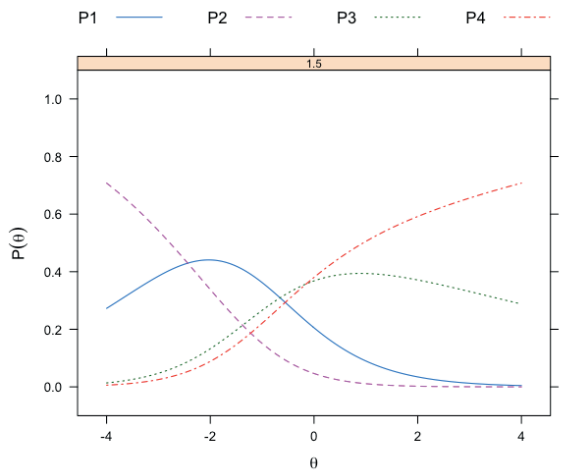
The order of the scoring functions in Table 6.1 reveals some preliminary information bout the quality of the initial scoring hypothesis. For all of the items, the scoring function for the scientifically accurate response always has the largest value, which is consistent with the decision to assign the highest nominal score to the correct response. The presence of large negative estimates for some scoring functions (e.g., items 1.2, 1.5, 2.2, 2.3, and 2.6) indicates that the response option assigned the lowest nominal score is indicative of higher ability than other answer choices hypothesized to be in the messy middle. There is also evidence to suggest that categories for some items should be collapsed. For example, the identical ak_2 and ak_3 estimates for item 1.1 indicate that these two categories can be combined. Items 1.7 and 2.8 only have 3 answer choices, and the estimated scoring functions (ak_1's) are not statistically different from the score for the lowest category. This result indicates that a dichotomous scoring scheme is more appropriate for these two items, which is consistent with the item design. Scoring functions for the remaining items are difficult to interpret using just statistical criteria because it is unclear whether the scores are sufficiently distinct to support partial credit scores relative to the model of student thinking, especially since the SEs for some scoring functions are large.

6.1.2. Using the NRM to Revise the Initial Scores for Items. In applied LP research, revisions to the initial hypothesis for how students' responses to the OMC items are scored should be interdisciplinary and involve both analysts and content experts. For this dissertation, however, I use my chemistry content knowledge and teaching experience in lieu of collaborating with other content experts. To begin, consider the relatively simplistic item 1.5 displayed in Figure 6.1. This item was designed to distinguish among four distinct ideas (facets 03, 40, 41, and 42). The common first digit in the facet code suggests that all of the distractors are related, but the facet statements suggest that there may be meaningful differences among the ideas. Facet 40 indicates that the student believes that atoms are created or destroyed through ordinary daily events, facet 41 specifies that the student believes atoms are destroyed or turned into a new form of energy when a substance is used or burned, and facet 42 indicates that a student believes atoms are created when a new substance is formed. Facet 40 relates to the misconception that atoms can be altered through ordinary events, facet 41 emphasizes an alternative idea related to the chemical process of burning, and facet 42 indicates another idea related to synthetic chemical reactions. These three facets also implicitly recognize that students have some difficulty accounting for the number of atoms in a chemical or physical change. The facets have a relatively low numerical code (40s), indicating that the alternative ideas are less problematic for understanding the learning goal. One might expect students of higher abilities to have some misconceptions associated with facet 40, 41, or 42.

Carbon, nitrogen, and hydrogen atoms are added together in a closed flask. After they are mixed, a chemical reaction happens and a thin film forms on the bottom of the flask.

Does the number of atoms in the flask increase, decrease, or remain the same?

(1) The number and types of atoms increase. [*Facet 42*]
(2) The number and types of atoms decrease. [*Facet 41*]
(3) The number of atoms remains the same, but the types of atoms change. [*Facet 40*]
(4) The number and types of atoms remain the same. [*Facet 03*]



Cat.	c_{ik}	a_{ik}	ak_{ik-1} [95% C.I.] ($a_i = 0.37$)	$b^*_{ikk'}$
(1)	0.07	-0.35	0	-
(2)	-1.41	-0.96	-1.65 [-2.22,-1.08]	-2.43
(3)	0.66	0.54	2.41 [2.16,2.67]	-1.38
(4)	0.69	0.76	3	-0.14

Figure 6.1. Response option curves for Diagnoser item 1.5.

The NRM response option curves and item parameters in Figure 6.1 offer guidance for how to revise the initial scoring hypothesis for item 1.5. Inspection of the response option curves indicates that students with very low abilities ($\theta < -2.5$) are most likely to choose the second response option, those of intermediate abilities ($-2.5 < \theta < 0$) are most likely to choose response option 1 or 3, and those with the highest ability ($\theta > 0$) are most likely to choose the correct answer. The order of the NRM response option slopes supports this interpretation ($a_{i2} = -0.96 < a_{i1} = -0.35 < a_{i3} = 0.54 < a_{i4} = 0.76$). The slope estimates for response options 3 and 4 are most similar in value, with response option 4 being most probable along a very small range of the latent ability scale ($-0.5 < \theta < 0$). The NRM intersection parameters are ordered for the initial scoring hypothesis ($b^*_{i21} = -2.43 < b^*_{i32} = -1.38 < b^*_{i43} = -0.14$), and they are also ordered if we recalculate them using the order suggested by NRM response option slopes

(i.e., $b_{i21}^{*} = -2.43 < b_{i31}^{*} = -0.66 < b_{i43}^{*} = -0.14$), but the distance between the latter two intersection parameters is quite small ($\Delta = 0.52$).

Analysis of the content of the item 1.5 can help to clarify whether these answer choices should be scored separately or if categories should be collapsed. Item 1.5 asks students what happens to the number of atoms in a closed flask after atoms of different elements are added together. The correct response requires that students understand the law of conservation of mass, which states that atoms cannot be created or destroyed by chemical or physical means. Students who select response option 2 have the lowest ability, and these students may have difficulty understanding mathematical concepts related to additivity or proportional reasoning (e.g., combining substances together results in a *decrease*). Students who select response options 1 or 4 have intermediate ability and may understand the mathematical concepts but have difficulty understanding the disciplinary idea that atoms are rearranged rather than created or destroyed in chemical reactions. The most recent science education framework (NRC, 2013) recognizes that mathematical and computational thinking is an important scientific practice and that the crosscutting concept of scale, proportion, and quantity is important across scientific disciplines. Although the facets were not developed to evaluate scientific practices or crosscutting concepts, this example illustrates how the order of facets may reveal information about how scientific practices, crosscutting concepts, and disciplinary ideas can be integrated.

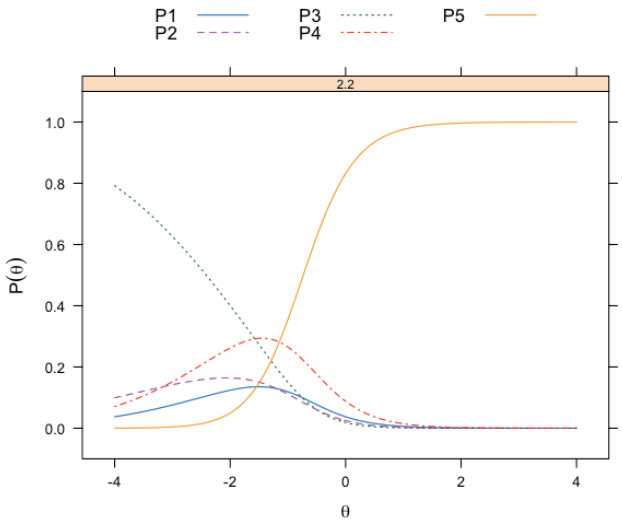
Five Diagnoser items were designed to distinguish among three macrofacets (items 1.3, 1.4, 1.6, 2.2, and 2.5) and have a design most similar to the OMC format promoted by Briggs and colleagues (2006). Figure 6.2 displays item 2.2 that asks students to accurately compare the size of an atom and a red blood cell. It was designed to distinguish among three distinct facets (facet 02, 52, and 82). Facet 52 indicates that a student compares the size of an atom to objects

visible with a microscope like bacteria, a virus, or a blood cell. Facet 82 specifies that the student thinks that living things are not made of atoms. Students who select the correct answer choice understand that atoms are very tiny and unable to be seen even using a light microscope (facet 02). The facet codes indicate that the answer choices associated with facet 82 are hypothesized to be most problematic for instruction, and we might expect students with the lowest abilities to have the greatest probability of selecting the answer choice associated with facet 82.

Select the answer that most accurately compares the size of the objects below.

I. atom of mercury
II. red blood cell

(1) An atom of mercury is much bigger than a red blood cell. [*Facet 82*]
(2) These two objects are the same size. [*Facet 52*]
(3) These two objects are close in size, but an atom of mercury is slightly larger. [*Facet 52*]
(4) These two objects are close in size, but a red blood cell is slightly larger. [*Facet 52*]
(5) A red blood cell is much bigger than an atom of mercury. [*Facet 02*]



Cat.	c_{ik}	a_{ik}	ak_{ik-1} [95% C.I.] ($a_i = 0.50$)	$b^*_{ikk'}$
(1)	-0.57	-0.16	0	-
(2)	-0.98	-0.50	-0.69 [-1.69,0.30]	-1.21
(3)	-1.27	-1.10	-1.87 [-2.98,-0.76]	-0.48
(4)	0.29	-0.10	0.11 [-0.63,0.85]	-1.56
(5)	2.54	1.86	4	-1.15

Figure 6.2. Response option curves for Diagnoser item 2.2.

The response option curves in Figure 6.2 offer guidance for how to revise the initial scoring hypothesis. Students with the lowest abilities ($\theta < -1.5$) are most likely to choose response option 3 and incorrectly believe that an atom of mercury is larger than a red blood cell.

The content of response option 3 is most similar to the content of response option 1 since that option also indicates that students incorrectly believe at an atom of mercury is bigger than a cell. Both answer choices suggest that students have difficulty understanding the relative *scale* of invisible particles, especially the size of particles from the physical sciences (e.g., atoms) compared to those from the life sciences (e.g., cells). Students with greater ability ($-1.5 < \theta < -1$) are most likely to select answer choice 4 and correctly believe that a cell is larger than an atom but fail to appreciate that the difference in size is substantial. Response option 2 indicates a similar misconception related to scale that the invisible objects have a similar size. Although options 3 and 4 also suggest that the objects have a similar size, these two options identify one object as larger than the other. There is a very small range where response option 3 is most probable, and there is no region along the latent ability scale where response options 1 or 5 are most probable. Students with the highest abilities ($\theta > -1$) are most likely to choose the correct response and accurately reason that cells are much larger than atoms. The NRM intersection parameters exhibit reversals, likely because some categories should be combined.

The results from items 1.5 and 2.2 suggest an emerging progression related to using the practice of mathematical thinking with the crosscutting concept of scale to understand the disciplinary idea of atoms. The results above also illustrate the difficulty of using the NRM slope parameters in isolation to revise the initial scoring hypothesis. The NRM response option slopes for item 2.2 suggest an order among the response options, but the slopes for the distractors all have a similar value ($a_{i2} = -0.82 < a_{i1} = -0.64 < a_{i5} = -0.26 < a_{i3} = -0.18$), especially when compared to the slope for the correct response ($a_{i4} = 1.90$). Without analyzing the content of the item, the slope estimates could be used as evidence to support a dichotomous scoring scheme. However, the scoring functions (ak_{ik-1}'s) are all imprecise estimates, as indicated by

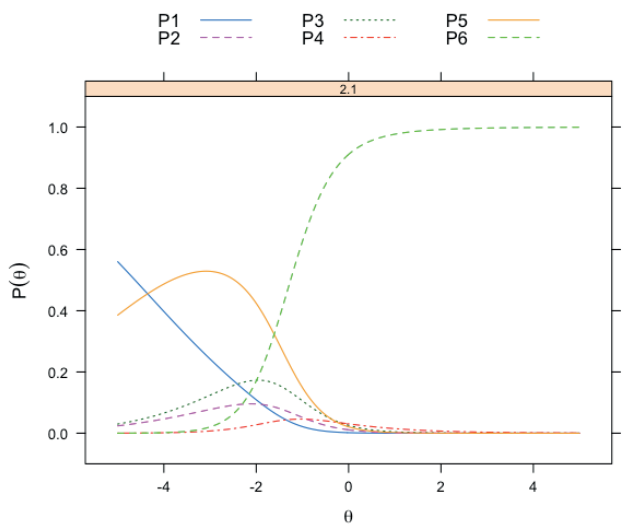
the large SEs (see Table 6.1) and wide confidence intervals. Few respondents selected the distractors for item 2.2, since this was an easy item (see Table 5.2). Although the imprecisely estimated NRM item parameters are used to derive the shapes of the response option curves, the approximate shape of these curves could also be recovered using simpler methods like those described in Chapter 3 (e.g., descriptive response option probability curves). Interpretation of the content that is most probable along different regions of the latent ability scale will likely not change even if other methods are used to create the response curves.

The most difficult Diagnoser items to analyze are those with answer choices corresponding to "unknown" facets (items, 1.1, 1.2, 2.1, 2.3, 2.6, and 2.7). Answer choices linked to unknown facets provide students with options that are not linked to ideas identified in the facet cluster. Minstrell (2000) recognized that students could have productive ideas that are not reflected in the facet cluster. Ongoing research could surface these ideas and potentially incorporate them into the model of student thinking. The inclusion of answer choices linked to unknown facets may also help minimize guessing since students who have ideas that are different from those in the facet cluster have an alternative choice to select. Consider item 2.1, which is displayed in Figure 6.3, and asks students to identify which of the following objects are made of atoms. This item evaluates four distinct facets (01, 80, 81, and 82) along with an unknown facet. The facet codes represented by the first digit indicate that the item only distinguishes among two macrofacets.

Which of the following are made of atoms?

I. Molecule
II. Virus
III. Paper

(1) II only [*Facet 80*]
(2) III only [*Facet 82*]
(3) I and II [*Facet 80*]
(4) II and III [*Facet 80*]
(5) I only [*Facet Unknown*]
(6) All of them [*Facet 01*]



Cat.	c_{ik}	a_{ik}	ak_{ik-1} [95% C.I.] ($a_i = 0.58$)	$b^*_{ikk'}$
(1)	-2.72	-1.30	0	-
(2)	-0.84	-0.29	1.72 [0.62,2.82]	-1.86
(3)	0.00	-0.17	1.93 [1.03,2.84]	-7.00
(4)	0.18	0.84	3.66 [2.95,4.37]	-0.18
(5)	-0.21	-0.72	0.99 [-0.07,2.05]	-0.25
(6)	3.58	1.63	5	-1.61

Figure 6.3. Response option curves for Diagnoser item 2.1.

The response option curves suggest that partial credit scoring with collapsed categories may be plausible for item 2.1. Students with very low ability ($\theta < -4$) are most likely to choose response 1, which indicates that only viruses are made of atoms. Similar to the lowest category for item 2.2 (see Figure 6.2), students who select response 1 may be conflating invisible particles from the life sciences (i.e., viruses) with those from the physical sciences (i.e., atoms). Facet 80 indicates that students think matter is comprised of things other than atoms. Students who have a more sophisticated understanding ($-4 < \theta < -2$) are more likely to indicate that only molecules are comprised of atoms (response 5). This answer choice is scientifically accurate. However, it is not the best choice since all of the objects are comprised of atoms. Students at this intermediate

level may have misconceptions related to the scale of atoms, specifically understanding that atoms are fundamental particles common to both invisible (viruses and molecules) and visible objects (paper). Students with the most sophisticated understanding ($\theta > -2$) recognize that atoms comprise all matter. The NRM slope parameters are less useful for understanding how to revise the initial scoring hypothesis because, like item 2.2 in Figure 6.2, the NRM parameter estimates are imprecise due to small amounts of students selecting the distractors.

The intersections between the response option curves further indicate that this is a very easy item and may not be useful for making distinctions among students near the middle of the ability scale. The intersection between the curves for options 1 and 5 occurs at the very bottom of the latent ability scale ($b_{51}^{*} = -4.33$). The intersection between the curves for options 5 and 6 ($b_{65}^{*} = -2.15$) also occurs at the lower end of the scale. Students have a relatively low probability of selecting answer choices associated with the remaining response option curves (for answer choices 2, 3 and 4). However, these choices are most probable along the region of the latent ability scale bounded by the intersection parameters b_{51}^{*} and b_{65}^{*}, suggesting that choices 2, 3 and 4 are most similar to response option 5. Although the response option curves for item 2.1 provide some evidence to support a partial credit scoring scheme, this item would only be useful if used with students of very low ability. The response option curves for items 1.2, 2.1, 2.3, and 2.7 display similar patterns to those depicted in Figure 6.4.

Item 2.6 is the most complex Diagnoser item used in this study. It is depicted in Figure 6.4, and it consists of two multiple-choice items each with three response options that are scored together. The first item is similar to item 1.5 (see Figure 6.1), and it evaluates students' understanding of the law of conservation of mass. The second item asks students to provide a rationale by connecting their response to properties of atoms. The two items are scored together

for the purpose of identifying students' facets. The response space consists of nine combinations of answer choices across the two items (e.g., selecting "a" for the first item and "b" for the second item). These combinations are numbered 1-9 in the right panel of Figure 6.4.

Ali mixed 50 ml of alcohol with 50 ml of water. No reaction occurred and neither of the liquids evaporated. She was surprised to notice that the fi-l volume of the alcohol-water solution was less than 100 ml.

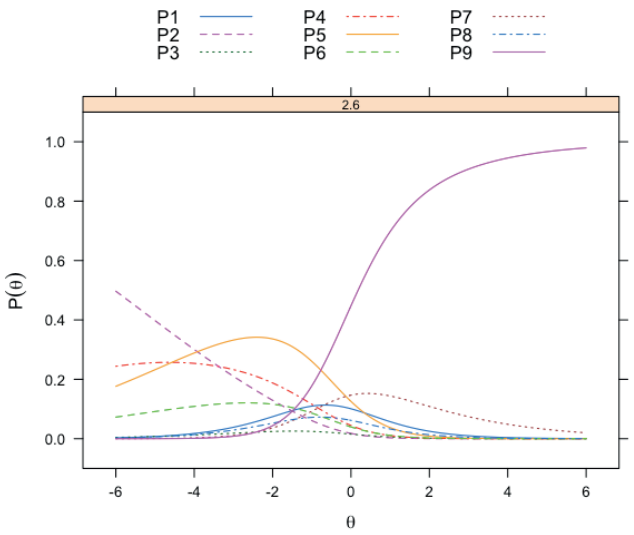
Suppose that Ali weighs the alcohol and the water before mixing and then weighs the alcohol-water solution after mixing.

How does the weight of the liquids compare before and after they are mixed?

(a) The alcohol-water solution after mixing weighs less. [*Next Item 6b*]
(b) The alcohol-water solution after mixing weighs more. [*Next Item 6b*]
(c) They weigh the same before and after mixing. [*Next Item 6b*]

Which of the following statements best matches your reasoning on the previous question?

(a) There are fewer atoms in the mixed solution compared to the number of atoms in the alcohol and water before mixing. [*Facet 41*]
(b) There are more atoms in this mixed solution compared to the number of atoms in the alcohol and water before mixing, but the atoms are just more tightly packed in the mixed solution. [*Facet 42*]
(c) There is the same number of atoms before and after mixing the alcohol and water, but the atoms are just more tightly packed in the mixed solution. [*Facet 00 if chose c for Item 6, otherwise unknown*]



Cat.	c_{ik}	a_{ik}	ak_{ik-1} [95% C.I.] ($a_i = 0.13$)	$b^*_{ikk'}$
(1 - aa)	0.42	0.25	0	-
(2 - ba)	-1.31	-0.89	-9.06 [-13.46,-4.67]	-1.52
(3 - ca)	-1.46	-0.13	-3.02 [-7.39,1.36]	0.20
(4 - ab)	-0.40	-0.62	-6.92 [-10.46,-3.37]	2.16
(5 - bb)	0.64	-0.40	-5.12 [-7.99,2.25]	-4.73
(6 - cb)	-0.51	-0.44	-5.47 [-8.91,-2.03]	-28.75
(7 - ac)	0.79	0.79	4.33 [2.68,5.97]	-1.06
(8 - bc)	-0.07	0.18	-0.55 [-3.19,2.09]	-1.40
(9 - cc)	1.91	1.26	8	-1.83

Figure 6.4. Response option curves for Diagnoser item 2.6.

The response option curves for item 2.6 have a similar pattern to item 2.1 displayed in Figure 6.3. Students with extremely low ability ($\theta < -4$) are most likely to choose option "b" for the first item and "a" for the second item. The content of these answer choices indicates some difficulty understanding the mathematical concepts of additivity or proportional reasoning (e.g., adding liquids together results in *fewer* atoms). Two of the three response combinations that indicate a similar inconsistency (4 - ab and 6 - cb) also have non-zero probabilities at very low values of the ability scale, and the third response combination (3 - ca) was relatively uncommon among respondents. These results suggest that these four combinations (ba, ab, cb, and ca) could be assigned the lowest score in a partial credit scoring scheme. Students with intermediate abilities (i.e., abilities bounded by the intersection parameters $b_{52}^{*} = -3.98$ and $b_{95}^{*} = -0.77$) were most likely to choose "b" for both items. Like the intermediate level for item 1.5, these students may understand additivity or proportional reasoning but have difficulty understanding the disciplinary idea that atoms are rearranged rather than created or destroyed in chemical reactions. Response combinations aa, bb, ac, and bc all indicate an intermediate level of understanding. Students who select the correct combination of responses (cc) provide evidence that they understand the law of conservation of mass and understand the relationship between this law and the concept of atoms. Like the two previously discussed example items, the NRM parameters for item 2.6 are estimated imprecisely and challenging to interpret in isolation.

6.2. Revising the Initial Scoring Hypothesis and Model of Student Thinking.

The results in the preceding section illustrate at the item-level how response option curves and the content of items can be interpreted together to revise an initial hypothesis about

how partial credit could be assigned to answer choices. These analyses also demonstrated that the NRM response option slopes and associated scoring functions can be difficult to interpret, especially when the number of students choosing a particular response option is low. The preceding discussion assumed that the items are of reasonable quality, recognizing that the items will be analyzed further in subsequent analyses. Analyses of the content of the items relative to the latent ability scale suggest that the facet cluster should be modified into an alternative developmental model of student thinking for the purpose of coherently mapping the model of student thinking to the latent ability scale.

Recall that facet clusters are sets of statements describing students' ideas about a physical situation or conceptual idea as they are heard in the classroom. As described in Chapter 2, the numerical codes assigned to facets signal an approximate order of development, but they are different from the levels of student ideas in LPs. Figure 6.5 presents the facet cluster for atoms in the left panel, and a three-level "facet progression" that describes a hypothesis for how these facets may develop during the course of instruction. I derived the facet progression by aggregating the exploratory insights from the preceding section across all items. Specifically, I identified the content features of the response options associated with each partial credit category and combined them into qualitative descriptions of reasoning at each level.

<table>
<tr><th>Facet Cluster</th><th>Facet Progression</th></tr>
<tr><td>00 All matter is made up of atoms, which are too small to be seen even with a powerful light microscope. Atoms cannot be created or destroyed by ordinary chemical or physical means.
01 The student understands that all matter is made up of atoms.
02 The student understands that atoms are tiny (too small to see even through a light microscope).
03 The student understands that atoms cannot be created or destroyed.</td><td>L3 Student understands the size of atoms relative to other objects and that atoms are distinct particles (i.e., the number and type of atoms are preserved in ordinary chemical or physical changes).</td></tr>
<tr><td>40 Student believes that atoms are created (or destroyed) through ordinary daily events.
41 When a substance is used or burned, atoms are destroyed, disappear or are turned into a form of energy.
42 When a new substance is created, atoms are created.</td><td rowspan="2">L2 Student understands that atoms are the smallest, distinct components of matter.
• Student does not adequately understand the scale of atoms relative to other invisible particles.
• Student attempts to apply mathematical concepts to understand the law of conservation of mass without fully understanding that atoms have distinct features that do not change (e.g., student may think that mixing two distinct types of atoms results in an increase of how many atoms are present).</td></tr>
<tr><td>50 The student does not have an accurate sense of the scale of an atom as compared with objects they can see.
51 Compares the size of the atom to objects visible to the naked eye like a grain of sand, speck of dust, tip of a pin, hair, etc. …
52 Compares the size of the atom with objects visible with a microscope like bacteria, a virus, blood cell, etc. …</td></tr>
<tr><td>80 The student thinks that not all matter is made up of atoms.
81 Matter is infinitely divisible (in theory) – if we kept cutting a substance, we would always get just a smaller amount of that same substance (until it is no longer strong enough to hold together).
82 Living things are not made of atoms.
83 Atoms do not exist because they cannot be seen.</td><td>L1 Student understands that objects consist of small invisible particles.
• Student conflates the size of invisible particles with those from the physical and life sciences.
• Student has difficulty understanding proportions (e.g., student may think mixing two types of atoms results in a decrease in the total amount of atoms present).</td></tr>
</table>

Figure 6.5. Hierarchical models of student thinking for the concept of atoms.

Both the facet cluster and the facet progression describe how students develop understanding of the concept of atoms. However, the facet cluster describes categories of related ideas that do not necessarily build on one another. The facet progression, on the other hand, represents a sequence of ideas that describe how student thinking builds from understanding that objects consists of small invisible particles (level 1), to understanding that these small particles are distinct and called atoms (level 2), and to finally understanding the relative size and distinctiveness of atoms (level 3). Even though the facet cluster for atoms was developed prior to the recent *Framework for Science Education* (NRC, 2013) and the NGSS, the facet progression derived from the facet cluster is an attempt to use prior research to begin weaving together disciplinary core ideas related to the structure and properties of matter (PS1.A) with the crosscutting concept of scale, proportion, and quantity and the scientific practice of using mathematics and computational thinking. The bullets associated with each level of the progression indicate some ways that students may be misapplying concepts related to scale or mathematical reasoning to science. Note that a single facet progression like that displayed in Figure 6.5 may be too fine-grained for the purpose of assessing the NGSS. Sets of progressions and their accompanying assessments, similar to Minstrell's strands of facet clusters, may be more useful for this purpose. The facet progression can be used to revise the initial hypothesis of how partial credit scores should be assigned to the Diagnoser items, as indicated in Table 6.2.

Table 6.2.

Revised Scoring Hypothesis for Diagnoser Items.

Item	Answer Choice*						
	a	**b**	**c**	**d**	**e**	**f**	**g**
1.1.	1 [0]	0 [0]	1 [1]	1 [0]	1 [0]	1 [0]	2 [2]
1.2.	1 [0]	2 [2]	0 [0]	1 [1]	-	-	-
1.3.	1 [1]	0 [1]	1 [1]	2 [2]	0 [0]	-	-
1.4.	1 [1]	0 [1]	1 [1]	2 [2]	0 [0]	-	-
1.5.	1 [0]	0 [0]	2 [1]	1 [0]	-	-	-
1.6.	1 [1]	1 [1]	2 [2]	1 [1]	0 [0]	-	-
1.7.	0 [0]	0 [0]	1 [1]	-	-	-	-
2.1.	1 [1]	0 [0]	1 [0]	1 [0]	-	1 [0]	2 [2]
2.2.	1 [1]	0 [1]	1 [1]	2 [2]	0 [0]	-	-
2.3.	1 [0]	2 [2]	0 [1]	1 [1]	-	-	-
2.4.	1 [0]	0 [0]	2 [1]	1 [0]	-	-	-
2.5.	1 [1]	0 [1]	1 [1]	2 [2]	0 [0]	-	-
2.6.	**aa** 1 [0]	**ba** 0 [0]	**ca** 1 [0]	-	-	-	-
	ab 0 [0]	**bb** 1 [0]	**cb** 1 [0]	-	-	-	-
	ac 1 [1]	**bc** 1 [1]	**cc** 2 [2]	-	-	-	-
2.7.	1 [2]	0 [0]	0 [1]	0 [0]	-	-	-
2.8.	0 [0]	0 [0]	1 [1]	-	-	-	-

* Initial scoring hypotheses are presented in square brackets.

In Table 6.2, response options hypothesized to be associated with the lowest level of the facet progression are assigned a score of 0, those with the intermediate levels are assigned a score of 1, and those associated with the highest level are assigned a score of 2. The initial scoring hypothesis derived from the macrofacet codes is presented in square brackets. Note that like the revised hypothesis, the initial hypothesis also contains three scoring categories. No Diagnoser items have answer choices mapped to more than three distinct facets. However, the numerical scores in the initial hypothesis have an inconsistent meaning across items. For example, an initial score of "1" for item 2.1 indicates an unknown facet, but an initial score of

"1" for item 2.2 indicates facet 52 (see Figures 6.2 and 6.3). In contrast, all of the partial credit scores in the revised scoring scheme are linked directly to the facet progression in Figure 6.5 and have a consistent interpretation relative to the facet progression. In the revised scheme, three items are scored dichotomously (items 1.7, 2.7, and 2.8) since the design of the items and the NRM response option curves do not provide evidence to support polytomous scoring. These items distinguish between levels 1 and 2 of the facet progression. The revised scheme uses scores to reflect variable performance relative to the latent variable described by the facet progression.

6.3. Fitting the NRM to Diagnoser Items to Evaluate a Revised Scoring Hypothesis.

In this section, I use the NRM to confirm that the revised scoring scheme is working as intended. The main difference between the prior application of the NRM and this analysis is that I collapse response categories together using the revised scoring scheme (see Table 6.2) rather than analyzing all response options individually. As I show below, collapsing categories greatly improves the precision of the NRM item parameters and improves the interpretation of the NRM parameters relative to the facet progression. I also evaluate the appropriateness of using the NRM with the Diagnoser data by evaluating the IRT assumption of local independence and exploring whether there is evidence to support the property of parameter invariance.

6.3.1. Interpreting Slope and Intersection Parameters.

Table 6.3 presents the NRM slope parameters (a_i's and ak_{ik-1}'s) with accompanying SEs and 95% confidence intervals for all the Diagnoser items after applying the revised scoring scheme. Appendix D contains the estimated intercept parameters (d_{ik-1}) and the Bock (1972) NRM slopes and intercepts (a_{ik}'s

and c_{ik}'s) after applying the revised scores. The right-hand side of Table 6.3 also presents the NRM category intersection parameters, which I calculate from the Bock slopes and intercepts using equation 4.5. Since the Chalmers (2012, 2020) specification of the NRM imposes constraints on the scoring functions for the lowest and highest categories, the only NRM slope parameters that are estimated are a_i and ak_1. Appendix D contains the NRM item parameters for the analytic sample of students who answered both question sets on the same day (Test C) using the revised scoring scheme. The results for Test C are similar to those presented in this section.

Table 6.3.

NRM Slope Parameters Using the Revised Scoring Hypothesis.

Item	NRM Parameter Estimates									
	a_i	(SE)	[95% C.I.]	ak_0	ak_1	(SE)	[95% C.I.]	ak_2	b^*_{21}	b^*_{32}
1.1.	0.85	(0.09)	[0.67,1.02]	0	0.45	(0.16)	[0.15,0.76]	2	-7.55	-1.09
1.2.	0.71	(0.06)	[0.59,0.83]	0	1.17	(0.09)	[0.99,1.34]	2	-2.84	-2.29
1.3.	1.07	(0.06)	[0.94,1.20]	0	0.58	(0.08)	[0.43,0.73]	2	-2.26	-0.94
1.4.	0.76	(0.04)	[0.68,0.85]	0	1.05	(0.07)	[0.91,1.18]	2	-1.62	-0.97
1.5.	0.79	(0.05)	[0.70,0.89]	0	1.31	(0.06)	[1.20,1.43]	2	-2.40	0.89
1.6.	1.05	(0.08)	[0.90,1.20]	0	1.04	(0.07)	[0.90,1.18]	2	-3.39	-0.09
1.7.	1.69	(0.10)	[1.49,1.88]	0	1	-	-	-	-1.60	-
n = 5,063										
2.1.	1.31	(0.18)	[0.96,1.66]	0	0.66	(0.18)	[0.32,1.01]	2	-4.18	-1.31
2.2.	1.20	(0.09)	[1.03,1.38]	0	0.31	(0.10)	[0.11,0.51]	2	-1.76	-0.98
2.3.	0.80	(0.08)	[0.64,0.96]	0	0.66	(0.15)	[0.38,0.95]	2	-3.23	-1.79
2.4.	0.86	(0.06)	[0.74,0.99]	0	0.78	(0.09)	[0.60,0.96]	2	-2.76	-0.3
2.5.	1.30	(0.10)	[1.10,1.49]	0	0.23	(0.11)	[0.03,0.44]	2	-2.55	-0.93
2.6.	0.91	(0.07)	[0.78,1.04]	0	0.82	(0.08)	[0.65,0.98]	2	-2.82	0.15
2.7.	1.44	(0.10)	[1.24,1.64]	0	1	-	-	-	-1.32	-
2.8.	1.38	(0.10)	[1.18,1.58]	0	1	-	-	-	-1.52	-
n = 2,387										

The NRM parameters in Table 6.3 support the ordered hypothesis in the revised scoring scheme. All of the scoring functions are ordered as expected (i.e., $ak_0 < ak_1 < ak_2$). The intersection parameters (b_{21}^* and b_{32}^*) exhibit no reversals and divide the latent ability scale into distinct regions where responses in each category are most likely. Comparing the location of the intersection parameters across items indicates that the scale is not evenly divided across items (e.g., $b_{21}^* = -7.55$ and $b_{32}^* = -1.09$ for item 1 while $b_{21}^* = -2.84$ and $b_{32}^* = -2.29$ for item 2). The 95% confidence interval for just four items (1.2, 1.4, 1.6, and 2.1) contains the expected score of 1, which is reasonably strong statistical evidence to suggest the middle category is distinct from the other two and can be constrained to the expected score of 1. The remaining items (1.1, 1.3, 1.5, 2.2, 2.3, 2.4, 2.5, and 2.6) have confidence intervals that do not contain the expected score of 1 *nor* the expected score for the lowest or highest category (i.e., 0 or 2). For item 2.5, $ak_1 = 0.23$, but the 95% confidence interval does not contain the score of 0 suggesting a distinct score. Similarly, $ak_1 = 1.31$ for item 1.5, and the 95% confidence interval also does not contain the score for the highest category. These results suggest the presence of a distinct intermediate level but signal problems with constraining the score of the middle category to be 1.

The item specific slope parameters (a_i) indicate that the items vary considerably in discrimination. In general, the SEs in Table 6.3 are much smaller than the SEs in Table 6.1. This suggests the NRM parameters estimated using the revised scoring scheme are more precisely estimated than the parameters for the initial scoring scheme, likely because the revised scoring scheme contains fewer categories per item (see Appendix A for a fuller discussion of the impact of sample size on NRM parameter estimation). In both tables, the SEs for the response option slopes tend to be larger than the SEs for the item slopes. These results also identify some potential challenges with constraining response option slope parameters (a_i's) to constant values.

6.3.2. Evaluating the Assumption of Local Independence. A fundamental assumption of all IRT models is that the only relationship that exists among the item responses on a test is through the latent variable (θ). Violations of the assumption of local independence indicate problematic items that should be revised to remove dependencies. Yen's (1984, 1993) Q_3 statistic can be used to identify potential dependencies among items. Items with estimates of Q_3 greater than 0.20 in absolute value are often flagged for revision. Tables 6.4 and 6.5 present estimates of Q_3 calculated using residuals from the NRM fit to Diagnoser data scored using the revised scoring scheme.

Table 6.4.

Yen's Q_3 Fit Statistics for Test A (Question Set 1).

	1.1.	1.2.	1.3.	1.4.	1.5.	1.6.	1.7.
1.1.	1						
1.2.	-0.08	1					
1.3.	-0.15	-0.06	1				
1.4.	-0.10	-0.08	-0.16	1			
1.5.	-0.08	-0.06	-0.10	-0.09	1		
1.6.	-0.16	-0.08	-0.16	-0.09	-0.02	1	
1.7.	-0.10	-0.07	-0.18	-0.10	-0.10	-0.10	1

Table 6.5.

Yen's Q_3 Fit Statistics for Test B (Question Set 2).

	2.1.	2.2.	2.3.	2.4.	2.5.	2.6.	2.7.	2.8
2.1.	1							
2.2.	**-0.21**	1						
2.3.	-0.05	-0.13	1					
2.4.	-0.09	-0.14	-0.03	1				
2.5.	**-0.22**	-0.07	-0.15	-0.17	1			
2.6.	-0.07	-0.16	-0.04	-0.04	-0.10	1		
2.7.	0.00	-0.19	-0.06	-0.05	-0.18	-0.06	1	
2.8.	-0.03	-0.11	-0.07	-0.09	-0.17	-0.11	-0.01	1

Note: Estimates greater in absolute value than 0.20 are bolded in red.

There are no items on Test A that have local dependencies, but there is some evidence of item dependencies in Test B. Items 2.1, 2.2, and 2.5 have Q_3 statistics slightly greater than 0.20 in absolute magnitude. These items all have a similar format. They all require students to compare two or more objects that are distinguished by roman numerals (e.g., "I", "II", etc.). Students who are familiar with solving items with this kind of format may be using strategies that are unrelated to what they know about the topic of atoms (e.g., eliminating answer choices). These items can be modified by presenting scenarios that compare objects in slightly different contexts. Item 2.7 provides an example of a different design that asks students to compare statements from two different students. Interestingly, local dependencies only appear for Test B, even though Test A also has similar comparison items. This discrepancy may be due to differences in the composition of the analytic samples for Tests A and B.

6.3.3. Evaluating the Property of Parameter Invariance. If an IRT model fits the data, item characteristics are not dependent on the test or the group of respondents used to calibrate the parameters. Investigating whether parameter invariance holds provides insight into how well the NRM fits the empirical data scored using the revised scoring scheme. To evaluate whether or not there is evidence to support parameter invariance, I randomly split the analytic data associated with each test into two subsamples and recalibrate the NRM item parameters for each subsample. I then correlate the item parameters across the two subsamples (e.g., correlating the ak_1 parameters from the first subsample with the ak_1 parameters from the second subsample). I repeat this process 1,000 times, using a different random split each time.

The summary statistics in Table 6.6 indicate moderately strong evidence to support parameter invariance. The correlations for the intercept parameters (d_1 and d_2) are high (> 0.95)

across both analytic samples (i.e., for Test A and Test B). The correlations for the slope parameters (a_i and ak_1) are high for Test A (0.91 and 0.84), but noticeably lower for Test B (0.75 and 0.79). The weaker correlations are likely the result of imprecisely estimating the response option slopes due to the reduced sample size that results from splitting the empirical data in half. Note that I present correlations only for the NRM item-level slopes, scoring functions, and response option intercepts since these are the parameters estimated by mirt.

Table 6.6.

Summary Statistics for NRM Item Parameter Correlations Across Random Splits.

NRM Parameter	Mean	SD	Med.	Min.	Max.
Test A (n = 5,063)					
a_i	0.91	0.08	0.93	0.61	1.00
ak_1	0.84	0.15	0.89	0.39	1.00
d_1	0.95	0.03	0.96	0.82	1.00
d_2	0.97	0.03	0.97	0.82	1.00
Test B (n = 2,387)					
a_i	0.75	0.14	0.78	0.11	0.99
ak_1	0.79	0.14	0.82	0.16	0.99
d_1	0.94	0.04	0.95	0.68	0.99
d_2	0.98	0.02	0.98	0.88	0.99

An alternative test of parameter invariance is to split the sample using another variable. In operational testing contexts, these variables could include a students' gender, race or ethnicity, or free or reduced-price lunch status. I do not have access to these variables. Instead, I split the sample by the date the student completed the question set. I have student response data from the years 2010-2017. In 2013, a new framework (NRC, 2013) and standards for science education (NGSS Lead States, 2013) were released that advocated for a three-dimensional approach that weaves together scientific practices, crosscutting concepts, and disciplinary core ideas. If science

instruction in 2010-2013 was substantively different from instruction in 2014-2017, then we might expect to see differences in students' abilities and consequently differences in the estimated NRM item parameters. For the a_i, ak_1, d_1, and d_2 NRM parameters, the Test A correlations are, respectively, 0.96, 0.93, 0.94, and 0.93, and the Test B correlations are 0.68, 0.63, 0.92, and 0.97, respectively. Parameter invariance appears to hold for Test A but weaken for Test B. The weaker correlations for Test B are likely driven by imprecision in the item parameter estimates caused by reduced sample size that results from the purposeful split; just 859 students responded to Test B items in 2014-2017 compared to 1,528 students from 2010-2013.

6.4. Using the PCM to Evaluate the Quality of the Latent Ability Scale.

When the facets are reorganized into a progression and the Diagnoser items are scored using this progression, the NRM produces reasonably strong evidence to support partial credit scoring (see Table 6.3). However, there is considerably variability in the NRM slope estimates, suggesting potential challenges if these slopes are constrained to constant values. In this section, I fit the PCM to the Diagnoser data for students who completed both question sets on the same day. The goal is to evaluate whether or not a latent scale can be constructed that would permit independent comparisons among the items or students. First, I analyze PCM category intersection parameters to ensure that there are no reversals that would signal problems with the scoring or design of the items. Then, I use an item-person map to evaluate whether the latent scale can be divided into distinct regions corresponding to the facet progression.

6.4.1. Interpreting PCM Intersection Parameters and Fit Statistics. Table 6.7 contains PCM category intersection parameters and infit and outfit statistics for Diagnoser items scored using the revised scoring scheme. The PCM intersection parameters for both the dichotomous and polytomous items are defined relative to intersections among adjacent levels of the facet progression displayed. More specifically, b_1 is the intersection between a response option curve mapped to the lowest level of the LP and the curve for responses mapped to the intermediate level, and b_2 is the intersection between the curve associated with the intermediate level and the curve for the highest level. All of the intersection parameters are ordered except for those associated with items 2.2 and 2.5.

Table 6.7.

PCM Item Parameters and Fit Statistics Using the Revised Scoring Hypothesis.

Item	Intersection Parameters		Infit Statistics		Outfit Statistics	
	b_1	b_2	MSQ	t-statistic	MSQ	t-statistic
1.1	-3.20	-1.23	0.99	-0.23	0.93	-1.07
1.2	-2.25	-1.62	1.13	2.18	1.10	1.12
1.3	-1.58	-1.05	0.9	-2.38	0.78	-3.72
1.4	-1.61	-0.80	1.03	0.62	0.95	-0.93
1.5	-2.56	0.37	1.06	1.61	1.04	0.96
1.6	-3.31	-0.07	0.96	-1.01	0.93	-1.81
1.7	-2.11	-	0.89	-2.00	0.68	-3.84
2.1	-3.48	-1.78	0.86	-2.42	0.63	-4.95
2.2	***-0.92***	***-1.41***	0.91	-1.93	0.80	-2.79
2.3	-2.05	-1.78	0.97	-0.40	0.83	-1.89
2.4	-1.86	-0.58	0.97	-0.73	0.88	-2.49
2.5	***-1.22***	***-1.39***	0.87	-2.97	0.69	-4.49
2.6	-2.16	0.27	0.94	-1.63	0.92	-2.30
2.7	-1.55	-	0.90	-2.50	0.79	-3.39
2.8	-1.66	-	0.91	-1.92	0.81	-2.74

Notes: Bold and italicized category intersection parameters represent category intersection reversals. The cells shaded in grey represent values for misfitting items (i.e., t-statistics greater than 2 in absolute value).

The infit and outfit mean square statistics in Table 6.7 indicate that there is widespread item misfit, likely caused by constraining the response option slope parameters. Of the 15 items, 10 have infit or outfit statistics that exceed the expected values. Although estimated using different analytic data, the NRM item parameters in Table 6.3 indicate that just 4 Diagnoser items support constraining the scoring functions to the expected partial credit scores (items 1.2, 1.4, 1.6, and 2.1). Like the variable point-biserial correlations in Table 5.1, the NRM slope estimates in Table 6.3 range from 0.71 to 1.69, suggesting that the Diagnoser items have varying discrimination. Consequently, using Rasch models that impose constraints on the slope parameters may be inappropriate for these data.

Tables 6.8 and 6.9 present results from investigating whether there is evidence for local dependence and whether parameter invariance may be present when using the revised scoring scheme. The Q_3 statistic identifies dependencies between items 1.5 and 2.2, items 1.4 and 2.4, and items 1.5 and 2.5. Items 2.2 and 2.4 also have reversals, suggesting these items should be revised. Removing items 2.2 and 2.4 and recalibrating the PCM parameters results in ordered category intersection parameters across all items, and the Q_3 statistics for all items falls below 0.20. Removing these two items mitigates both reversals and local dependencies. Table 6.9 presents the average results after 1,000 iterations of randomly splitting the analytic data for Test C into two sets, calibrating the PCM parameters separately for each set, and then correlating them. The random splits suggest that the PCM intercept parameters are highly correlated across random splits. Purposefully splitting the sample based on the year of completion also produces strong correlations (0.89 for d_1 and 0.96 for d_2). These analyses provide preliminary evidence to support the assumptions and properties underlying the use of IRT models with the Diagnoser data even though there is evidence for widespread misfit.

Table 6.8.

Yen's Q_3 Fit Statistics for Test C.

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8
1.1	1														
1.2	-0.06	1													
1.3	0.01	-0.04	1												
1.4	0.00	-0.03	-0.05	1											
1.5	0.03	-0.08	-0.04	-0.06	1										
1.6	-0.05	-0.09	-0.04	-0.07	0.00	1									
1.7	0.01	-0.06	-0.06	-0.06	-0.03	-0.01	1								
2.1	0.12	-0.09	-0.07	-0.18	-0.14	-0.11	0.14	1							
2.2	-0.12	-0.17	-0.03	-0.03	**-0.20**	-0.18	-0.09	-0.03	1						
2.3	-0.13	0.07	-0.08	-0.13	-0.12	-0.04	-0.05	0.04	-0.1	1					
2.4	-0.13	-0.12	-0.18	**-0.20**	-0.09	0.03	-0.05	0.00	-0.05	-0.04	1				
2.5	-0.19	-0.13	-0.02	-0.06	**-0.21**	-0.11	-0.05	-0.01	0.19	-0.07	-0.06	1			
2.6	-0.12	-0.10	-0.08	-0.15	-0.06	-0.07	-0.01	-0.02	-0.10	-0.05	-0.03	-0.04	1		
2.7	-0.03	-0.08	-0.08	-0.11	-0.08	-0.03	0.10	0.08	-0.10	0.02	0.00	-0.02	-0.03	1	
2.8	-0.04	-0.09	-0.08	-0.07	-0.06	-0.12	0.06	0.08	0.01	-0.02	-0.01	-0.04	-0.05	0.06	1

Table 6.9.

Summary Statistics for PCM Item Parameter Correlations Across Random Splits.

PCM Parameter	Mean	SD	Med.	Min.	Max.
d_1	0.94	0.03	0.94	0.80	0.99
d_2	0.96	0.02	0.97	0.85	1.00

Note: The d parameters are the PCM response option intercepts estimated by mirt (see Equation 4.3).

The results from scoring the Diagnoser using the revised hypothesis reveal that the PCM constraints on the data are quite strong. The presence of intersection parameter reversals for items 2.2 and 2.5 and the prevalence of misfit suggest that there is too much variability in the response option slopes and intercepts to constrain them for PCM applications. Although the PCM intersection parameter reversals and local dependence disappear when items 2.2 and 2.5 are removed from the data, the infit and outfit statistics continue to flag the majority of items on the test for misfit. These results indicate that independent comparisons among the items should

only be made after they are revised to fit the model better. Despite this limitation, I ignore misfit in the next section and illustrate how one might analyze the items together for the purpose of dividing the latent ability scale into regions that correspond to the facet progression.

6.4.2. Using an Item-Person Map to Develop a Criterion-Referenced Scale. If OMC items fit a Rasch model, the strongest preliminary evidence for the order of levels in an LP would emerge if all items have ordered category intersection parameters *and* the parameters corresponding to transitions between levels are located along distinct regions of the latent ability scale. Figure 6.6 presents an item-person map for the Diagnoser items scored using the facet progression after removing the problematic items 2.2 and 2.4. This display plots students' estimated abilities and PCM category intersection parameters along the same latent ability scale. The left-most portion of the display displays a histogram of students' estimated abilities, and the right-hand side plots the estimated category intersection parameters for each item.

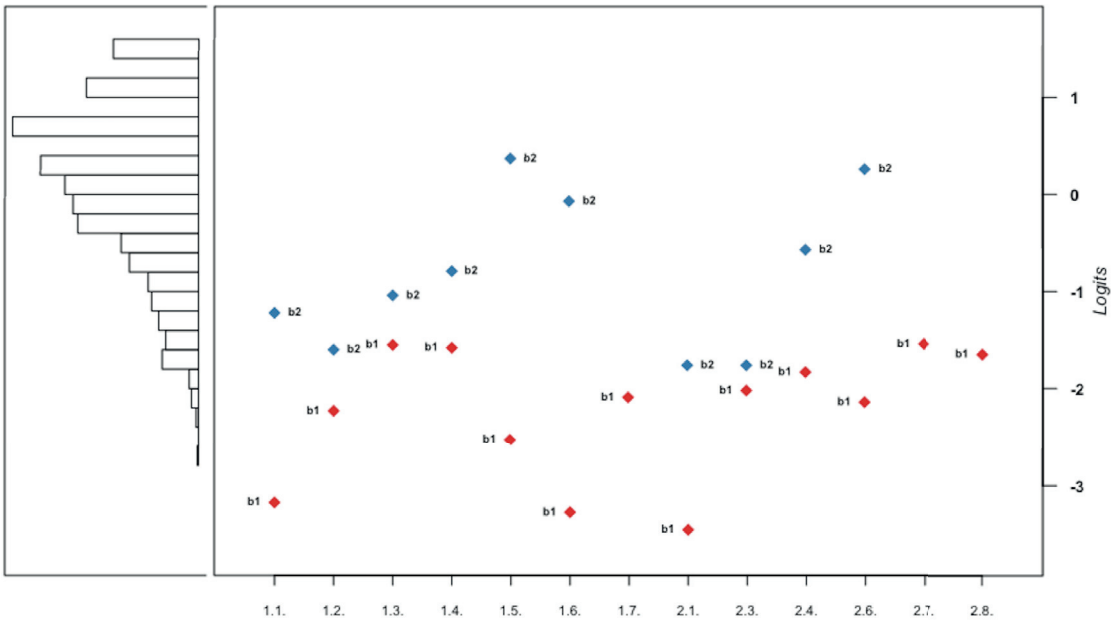


Figure 6.6. Item-person map for Diagnoser items.

Initial inspection of the item-person map reveals some challenges with the test that can be difficult to detect only from inspection of the values of the item parameters. First, the test was quite easy for most students. Most of the students have estimated abilities greater than -1 logit, but the majority of item parameters are located below -1 logit. In the ideal scenario, most item parameters would be located along the same region of the latent ability scale as the respondents. Second, the intersection parameters do not segment the scale into clear regions. Consider items 1.2 and 1.3. The intersection parameter for item 1.2 that quantifies the probability of scoring in the highest level compared to the intermediate level ($b_2 = -1.60$) is located at approximately the same location of the scale as the parameter for item 1.3 associated with the probability of scoring in the intermediate level compared to the lowest level ($b_1 = -1.55$). Overlapping intersection parameters make it challenging to divide the scale into regions corresponding to levels of the LP. Third, the distance between the estimated parameters varies considerably across items. Intersection parameters for some items are very close to one another (e.g., $b_1 = -2.02$ and $b_2 = -1.76$ for item 2.3), while other items have them far apart (e.g., $b_1 = -2.53$ and $b_2 = 0.37$ for item 1.5). These results suggest that revisions to the design of the items may permit better measurement of the latent variable described by the progression. Note this should not be surprising since the items were designed to measure facets in a cluster *not* a progression.

Classifying students into levels of the progression may still be possible, especially if the goal of assessment is to provide information to teachers for the purpose of supporting classroom instruction. The three items where the intersection parameters overlap the most (items 1.2, 2.1, and 2.3) all have one or more answer choices mapped to unknown facets. If we remove these items from the test and recalibrate the PCM parameters, the b_1 parameters across all items are always smaller in magnitude compared to the b_2 parameters. The mean of the intersection

parameters ($\overline{b_1} = -2.14$ and $\overline{b_2} = -0.44$) can then be used to slice the latent ability scale into three distinct regions that can be mapped to the levels of the progression. This method was used by Brown and Wilson (2011) to divide the latent ability scale into regions corresponding to the levels of an LP. Students located at the lowest level of the progression would have estimated abilities less than -2.14 logits, those in the intermediate level would have abilities between -2.14 and -0.44 logits, and those at the highest level would have estimated abilities greater than -0.44 logits. Applying this procedure results in 0.05% of students ($n = 6$) being located at the lowest level of the LP, 30% ($n = 327$) located in the intermediate level, and 70% of students ($n = 772$) being located at the highest level of the facet progression. This application of the PCM offers some evidence to support criterion-referenced interpretations of the scale, even though the results also indicate that some Diagnoser items should be revised. In the next chapter, I discuss these results relative to the literature on LP validation, identify limitations with the method developed in this dissertation, and identify some potential areas for additional research.

Chapter 7

Discussion

The purpose of this dissertation was to develop an IRT-based method to interpret the order of LP levels using response data collected from items with answer choices mapped to levels of the LP. In the previous chapter, I demonstrated how the analysis of response option curves and item parameter estimates from the NRM and PCM can be used to evaluate and revise a hypothesis for how the scores on Diagnoser items connect to a model of student thinking. In this final chapter, I discuss the main research contribution of this study in the first section by focusing on the strengths of this method for LP validation studies. Then, I identify some challenges and limitations of using the NRM and PCM together in LP validation studies. I conclude this dissertation by offering some suggestions for future research on the psychometric modeling of items with responses mapped to levels of a hierarchical model of student thinking.

7.1. Summary of Methodological Contribution.

The OMC item type (Briggs, et al., 2006) is an innovative approach to assessment design that attempts to improve the quality of information tests can provide about what students can know and do by coupling the selected-response item format with a developmental model of student thinking. As discussed in Chapters 2 and 3, the OMC item design is one example of how items can be designed relative to hierarchical models of student thinking, which are models that describe how student thinking becomes more sophisticated over time. The items that comprise the Diagnoser assessment system (Thissen-Roe, et al., 2014) can be viewed as extensions of the

OMC item design, as described in Chapter 2, since Diagnoser items all have one response option associated with a correct idea but vary how the remaining answer choices are mapped to the levels of the model of student thinking. Research that interprets student responses to multiple-choice items that are designed using a model of student thinking has focused on either descriptive analysis of response option curves for individual items or on Rasch analysis using item-person maps. Item-level response option curve analysis can help to clarify how scores could be assigned to responses to reflect increasing amounts of the latent variable being measured, and item-person maps can reveal whether a quantitative scale can be constructed that maps onto the LP hypothesis. As detailed in Chapters 4, 5, and 6, this dissertation connects together these separate strands of psychometric research (i.e., response option curve analysis and analysis of item-person maps) for the purpose of improving test score interpretations relative to LPs.

In this dissertation, I have shown how the interpretation of item parameters from the NRM and PCM can inform both the partial credit scoring of items and the revision of a model of student thinking. The NRM can be used to evaluate and revise an initial hypothesis of how partial credit should be assigned to items. Analyzing patterns among response option curves relative to the content of the items can reveal insight into how the latent variable might be structured and how partial credit could be assigned to the answer choices. Aggregating these findings across items led to a revision of the initial scoring hypothesis and accompanying model of student thinking. This resulted in a re-specification of the hierarchal model of student thinking from a facet cluster to a facet progression that more clearly maps onto the assumed quantitative structure of the latent variable. Fitting the NRM and PCM to the data scored using the facet progression recovered the expected order but revealed some problems with the assessment.

These challenges resulted from violations of the PCM model assumptions and made it difficult to construct a latent ability scale that permits independent comparisons among students or items.

7.2. Limitations.

The IRT-based method of analyzing items using the NRM and PCM that I developed in this dissertation is most appropriate to use when there are two preconditions. First, there should be some theoretical evidence, at a minimum, that the latent variable has a quantitative structure with clearly, defined ordered levels. As Michell (1997) has argued, this assumption is seldom made explicit or evaluated in empirical studies. But it is important to clarify a clear hypothesis for how we expect the latent variable to be structured if we seek to conduct LP validation studies that map hierarchical models of student thinking onto quantitative latent ability scales. Second, there should be enough student response data to permit adequate estimation of NRM item and person parameters (see Appendix A). The NRM is a relatively complex IRT model with slope and intercept parameters for every response option within an item. For lengthy tests, a very large amount of student data is necessary to permit precise estimation of the NRM item parameters that are the focus of interpretation in this dissertation.

As I discussed in Chapter 2, there are a variety of models of student thinking that can be coupled with educational tests and these models have different assumptions about how learning occurs. Partially ordered models of student thinking hypothesize that mastery develops through a relatively sudden process and consist of taxonomies of alternative and disciplinary correct ideas. These kinds of models of student thinking may be most appropriate to use with psychometric models that consider the latent variable to consist of categorical attributes rather than a quantity

(e.g., diagnostic classification models or latent class analysis). Hierarchical models of student thinking, like LPs, are more appropriate to use for designing OMC or Diagnoser items and for IRT analyses. These models impose a hypothetical order among sets of ideas. This order is a minimum requirement for a latent variable that has a quantitative structure. Once order has been established more stringent tests can be performed to explore whether the latent variable is quantitative. It can be difficult to know in advance whether the latent variable has a categorical or quantitative structure, but a model of student thinking can make an initial hypothesis explicit.

The goal of using the NRM and PCM to analyze item responses is to clarify how the numerical scores assigned to observations about students connect to both levels of a hierarchical model of student thinking and to locations along a latent ability continuum. The NRM is particularly useful for this purpose because it permits estimation of item parameters for each multiple-choice response option. However, as I illustrated in Chapter 6 (see Table 6.1), NRM item parameters can be imprecisely estimated when a test item contains numerous response options or when very few students selected an answer choice. The sample size requirements are especially problematic for LP validation studies because they are typically conducted using relatively small samples of students. Alternative methods of analyzing response option curves presented in Chapter 3 may be a useful alternative in these contexts (e.g., descriptive analysis of response option curves or Rasch-based analysis of item option curves), but additional research should explore the extent to which similar conclusions result from the NRM and the application these alternative methods.

7.3. Directions for Future Research.

LP validation studies are a new area of research that uses insight from psychometrics to provide evidence to support interpretations of test scores relative to hierarchical models of student thinking like LPs. Because of the novelty of this research program, there are open questions about the "right" psychometric model to use with LPs. Different models may reveal different information about what students know and can do relative to a model of student thinking, but studies comparing and contrasting information produced by multiple modeling approaches are relatively rare in the research literature. Comparative studies are necessary to understand whether model specification (i.e., specification of the latent variable described by a model of student thinking as categorical or quantitative) leads to substantively different information about what students know and can do.

7.3.1. LP Validation Research. The psychometric characteristics of OMC or OMC-like items are also not well understood. This study has shown that the Diagnoser items used in this study have variable item-level discrimination (Tables 5.2 and 6.3) and that the scoring functions for the items vary considerably across items (Table 6.3). These results suggest that analyzing these items using psychometric models that impose constraints on the slopes of response option curves may not be appropriate because more restrictive models may not fit the empirical data. Additional research is needed to explore the extent to which the results from the Diagnoser items generalize across permutations of the OMC item format. Specifically, a key question is whether or not the OMC item format can generate data that can be scored relative to the levels of an LP and analyzed using Rasch models to produce scales that meet the criteria of specific objectivity.

Although the idea behind the OMC item design is a relatively intuitive one, this item format may be difficult to use with more complex models of student thinking. If a hierarchical model of student thinking consisted of five levels, it would be difficult to write distinct items where all the answer choices mapped to distinct levels of the LP. Additional research is needed to explore how to analyze item responses that deviate from the ideal OMC item design (e.g., items where the responses only map to the three lowest levels of a five-level LP). The NRM may still provide useful information about how partial credit might be assigned to the response options, and it can clarify how to interpret intersection parameters relative to the model of student thinking. For example, I illustrated in Chapter 6 how the intersection parameters for the dichotomously scored Diagnoser items could be interpreted relative to transitions between the lower two levels of the facet progression.

More research is also needed to distinguish among the competing interpretative perspectives on how to define the order of the latent variable relative to model parameter estimates. Recall, Bock's (1972, 1997) original conception of the order of the latent variable emphasized interpretation of the empirical order of NRM slope parameters (see Figures 4.2 and 6.2). This perspective suggests an approach for assigning partial credit to answer choices that have non-zero probabilities along the latent ability continuum. The competing perspective offered by Masters (1982) and Andrich (2013, 2015) is more stringent and emphasizes interpretation of the order of category intersection parameters to signal meaningful transitions between performance categories. The latter perspective is useful in LP validation studies where the categories are defined by levels of the LP, but the former perspective may be appropriate in more exploratory studies where the properties of the resulting latent variable scale are less important than understanding the characteristics of the assessment.

Finally, there remain substantive questions about whether the OMC item format alone would support strong claims relative to the more recent, complex educational standards like the NGSS. OMC items are selected-response items. As such, they are limited in their ability to provide information about how well students are able to produce evidence to support claims about what students can do, especially with respect to disciplinary practices. It is not well understood how information about what students know relative to a model of student thinking collected from OMC items compares to information from more authentic performance tasks. These comparability studies may be more advantageous for large-scale uses than for the purposes of locating students along a continuum for instructional purposes.

7.3.2. Leveraging Facet Clusters to Design Next Generation Science Assessments.

Research that describes how student thinking develops over time is not a novel endeavor, and modern science education reform continues to emphasize the importance of articulating "progressions" of how student thinking evolves. Both *A Framework for K-12 Science Education* (NRC, 2013) and the *Next Generation Science Standards* (NGSS Lead States, 2013) provide rough sketches of how disciplinary core ideas, crosscutting concepts, and scientific and engineering practices may develop across the K-12 grade span. These dimensions were never envisioned as independent, and the *Framework* states: "in order to facilitate students' learning, the dimensions must be woven together in standards, curricula, instruction, and assessment" (p. 29-30). A key question that remains for the field is whether or not the dimensions must intentionally be woven together to create new resources as suggested by the *Framework* or whether the interdependence among the strands means that the three dimensions are always present in science education resources but perhaps unnoticed. I adopt the latter perspective in this

dissertation, and I have shown how older science education resources (a facet cluster) could be adapted and transformed to promote the development of conceptual understanding and critical thinking rather than the mastery of discrete pieces of information.

If crosscutting concepts and practices cannot be separated from disciplinary core ideas, then progressions of core ideas that may appear at first glance to be one-dimensional can be a useful starting point for the design of new resources to support science education reform. The advantage of this approach is leveraging the wealth of conceptual change research on models of student thinking. The key challenge, though, is finding authentic opportunities to layer and develop more fully the crosscutting concepts and practices that may be latent in the older one-dimensional models. For example, older models of student thinking like facet clusters and the assessments designed from them may be insufficient for students to deeply engage with the science and engineering practices. Critical questions remain about how these models could be used to create alternative learning opportunities more consistent with the ideas articulated in the *Framework*. Building on older, disciplinary-idea-focused, models of student thinking has the potential of being a more fruitful route to designing resources consistent with the *Framework* than attempting to create new materials from scratch.

7.4. Conclusion.

Producing criterion-referenced information about what students know and can do relative to a model of a student thinking represents a new expert consensus for the vision of educational assessment that was elevated in *Knowing What Students Know* (NRC, 2001). This shift has implications for the quality of information that can be provided to support teachers' instructional

practices. Rather than just classifying students into a group that understands the learning goal and a group that does not (e.g., proficient or not), coupling educational assessments with explicit little learning theories has the potential of providing substantive information about what students who have not yet mastered the learning goal might know and be able to do. Targeted instruction that builds on what students already know relative to a model of student thinking may help students attain the learning goal represented by the highest level of the LP more efficiently.

Despite this potential, analyses of educational assessments designed using models of student thinking is a new, interdisciplinary area of research that blends content expertise with psychometrics. This dissertation contributes to this research literature by developing an IRT-based method of analyzing responses to OMC items for the purpose of exploring hypotheses about the order of the latent variable described by the model of student thinking. I illustrate how to apply this method using data from the Diagnoser assessment system. As LP validation studies continue to be conducted to evaluate new models of student thinking using OMC items, this study can help researchers identify and apply methods that produce evidence for the structure of the latent variable. The use of the NRM and PCM I advance in this dissertation provides new psychometric information that can be used for the validation of tests designed using LPs.

References

Alonzo, A. C., & Steedle, J. T. (2008). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389-421. doi:10.1002/sce.20303

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573. doi:10.1007/BF02293814

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*, 7-16. doi:10.1097/01.mlr.0000103528.48582.7c

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "threshold disorder controversy." *Educational and Psychological Measurement*, *73*(1), 78-124. doi:10.1177/0013164412450877

Andrich, D. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, *34*(2), 8-14. doi:10.1111/emip.12074

Andrich, D., & Styles, I. (2011). Distractors with information in multiple choice items: A rationale based on the Rasch model. *Journal of Applied Measurement*, *12*(1), 67-95.

Battisti, B. T., Hanegan, N., Sudweeks, R., & Cates, R. (2010). Using item response theory to conduct a distractor analysis on conceptual inventory of natural selection. *International Journal of Science and Mathematics Education*, *8*(5), 845-868. doi:10.1007/s10763-009-9189-4

Becker, R. A., Chambers, J., & Wilks, A. R. (1988). *The new S language*. London: Chapman & Hall.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley Publishing.

Bock, R. D. (1972). Estimating item parameters and latent abilities when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. doi:10.1007/BF02291411

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 33-50). New York, NY: Springer-Verlag.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403-425. doi:10.1007/S11336-013-9350-4

Briggs, D.C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*(2), 204-226. doi:10.1111/jedm.12011

Briggs. D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch model. *Measurement*, *146*, 961-971. doi:10.1016/j.measurement.2019.07.035

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 293-316). Boston, MA: Sense Publishers.

Briggs, D. C., & Circi, R. (2017). Challenges to the use of artificial neural networks for diagnostic classifications with student test data. *International Journal of Testing*, *17*(4), 302-321. doi:10.1080/15305058.2017.1297816

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33-63. doi:10.1207/s15326977ea1101_2

Briggs, D. C., & Peck, F. A. (2015). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research and Perspectives*, *13*(2), 75-99. doi:10.1080/15366367.2015.1042814

Brookhart, S. M. (2004). Classroom assessment: Tension and intersections in theory and practice. *Teachers College Record*, *106*(3), 429-458. doi:10.1111/j.1467-9620.2004.00346.x

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review*, *23*(2), 221-234. doi:10.1007/s10648-011-9161-z

Castle, C. (2018). *Measuring multidimensional science learning: Item design, scoring, and psychometric considerations* (Unpublished doctoral dissertation). Boston College, Boston, CO.

Cai, L. (2010). Metropolis-Hastings-Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307-335. doi:10.3102/1076998609353115

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. doi:10.18637/jss.v048.i06

Chalmers, R. P. (2020). *Multidimensional item response theory*. Retrieved from https://cran.r-project.org/web/packages/mirt/mirt.pdf

Chatterji, M. (2003). *Designing and using tools for educational assessment*. New York, NY: Allyn & Bacon.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.

Chen, J., Gotwals, A. W., Anderson, C. W., & Reckase, M. D. (2016). The influence of item formats when locating a student on a learning progression in science. *International Journal of Assessment Tools in Education*, *3*(2), 101-122. doi:10.21449/ijate.245196

Chen, F., Yan, Y., & Xin, T. (2017). Developing a learning progression for number sense based on the rule space model in China. *Educational Psychology*, *37*(2), 128-144. doi:10.1080/01443410.2016.1239817

Chi, M. T. H., & Glaser, R. (1985). Problem solving ability. In R. J. Sternberg (Ed.), *Human abilities: An information-processing approach* (pp. 227-250). New York, NY: W. H. Freeman & Company.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7-76). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Circi, R. (2015). *The marginal edge of learning progressions and modeling: Investigating diagnostic inferences from learning progressions assessment* (Unpublished doctoral dissertation). University of Colorado Boulder, Boulder, CO.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the Learning Sciences*, *13*(1), 15-42. doi:10.1207/s15327809jls1301_2

Confrey, J. (2012). Better measurement of higher cognitive processes through learning trajectories and diagnostic assessments in mathematics: The challenge in adolescence. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The adolescent brain: Learning, reasoning, and decision making* (p. 155-182). Washington, DC: American Psychological Association.

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). Learning progressions in science: An evidence-based approach to reform. *Consortium for Policy Research in Education Research Reports* (RR-63). Retrieved from http://www.cpre.org/images/stories/cpre_pdfs/lp_science_rr63.pdf

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. doi:10.1007/BF02310555

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement*, *23*(1), 3-19. doi:10.1177/01466219922031130

DeBarger, A. H., Ayala, C., Minstrell, J., Kraus, P., & Stanford, T. (2009, April). Facet-based progressions of student understanding in chemistry. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models.* New York, NY: Springer.

DeMars, C. E. (2003). Sample size and the recovery of Nominal Response Model item parameters. *Applied Psychological Measurement*, *27*(4), 275-288. doi:10.1177/0146621603027004003

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. B. Pufall (Eds.), *Constructivism in the computer age* (pp. 35-60). Hillsdale, NJ: Erlbaum.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*(2-3), 105-225. doi:10.1080/07370008.1985.9649008

diSessa, A. A. (2008). A bird's-eye view of the "pieces" vs "coherence" controversy (from the "pieces" side of the fence). In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 453-478). New York, NY: Routledge.

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1-19. doi:10.1007/s11336-013-9342-4

Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. *Studies in Science Education*, *5*(1), 61-84. doi:10.1080/03057267808559857

Duckor, B., Draney, K., & Wilson, M. (2017). Assessing assessment literacy: An item response modeling approach for teacher educators. *Pensamiento Educativo*, *54*(2), 1-25. doi:10.7764/PEL.54.2.2017.5

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, *46*(6), 606-609. doi:10.1002/tea.20316

Duncan, R. G., Choi, J., Castro-Faix, M., & Cavera, V. L. (2017). A study of two instructional sequences informed by alternative learning progressions in genetics. *Science & Education*, *26*, 1115-1141. doi:10.1007/s11191-017-9932-0

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A 10.1080/03057267.2011.604476zreview and analysis. *Studies in Science Education*, *47*(2), 123-182. doi:10.1080/03057267.2011.604476

Fischer, G. H., & Molenaar, I. W. (Eds.) *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer-Verlag.

Fox, J. (2019). *Polychoric and polyserial correlations.* Retrieved from https://cran.r-project.org/web/packages/polycor/polycor.pdf

Frohbieter, G., Greenwald, E., Stecher, B., & Schwartz, H. (2011). *Knowing and doing: What teachers learn from formative assessments and how they use the information.* (CSE Technical Report). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Fulmer, G. W. (2015). Validating proposed learning progressions on force and motion using the force concept inventory: Findings from Singapore secondary schools. *International Journal of Science and Mathematics Education*, *13*(6), 1235-1254. doi:10.1007/s10763-014-9553-x

Fulmer, G. W., Liang, L. L., & Liu, X. (2014). Applying a force and motion learning progression over an extended time span using the force concept inventory. *International Journal of Science Education*, *36*(17), 2918-2936. doi:10.1080/09500693.2014.939120

Furtak, E. M., Circi, R., & Heredia, S. C. (2018). Exploring alignment among learning progressions, teacher-designed formative assessment tasks, and student growth: Results from a four-year study. *Applied Measurement in Education*, *31*(2), 143-156. doi:10.1080/08957347.2017.1408624

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*(6), 1082-1116. doi:10.3102/0034654317726529

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21-35. doi:10.1111/j.1745-3992.2006.00076.x

Gotwals, A. W., & Alonzo, A. C. (2012). Introduction: Leaning into learning progressions in science. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions*. (pp. 3-12). Boston, MA: Sense Publishers.

Gotwals, A. W., & Songer, N. B. (2009). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*, *94*(2), 259-281. doi:10.1002/sce.20368

Greeno, J. G., Pearson, P. D., and Schoenfeld, A. H. (1996). *Implications for NAEP of research on learning and cognition*. Report commissioned by the National Academy of Education Panel on the NAEP Trial State Assessment. Menlo Park, CA: Institute for Research on Learning.

Hadenfeldt, J. C., Bernholt, S., Liu, X., Neumann, K., & Parchmann, I. (2013). Using ordered multiple-choice items to assess students' understanding of the structure and composition of matter. *Journal of Chemical Education*, *90*(12), 1602-1608. doi:10.1021/ed3006192

Hadenfeldt, J. C., Neumann, K., Bernholt, S., Liu, X., & Parchmann, I. (2016). Students' progression in understanding the matter concept. *Journal of Research in Science Teaching*, *53*(5), 683-708. doi:10.1002/tea.21312

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hammer, D., & Elby, A. (2003). Tapping epistemological resources for learning physics. *The Journal of the Learning Sciences*, *12*(1), 53-90. doi:10.1207/S15327809JLS1201_3

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Paper prepared for the Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessments and Student Standards (SCASS). Washington, DC: Council of Chief State School Officers. Retrieved from https://www.csai-online.org/sites/default/files/Learning_Progressions_Supporting_2008.pdf

Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, *12*(2), 184-192. doi:10.1039/C1RP90023D

Herrmann-Abell, & DeBoer, G. E. (2014). Developing and using distractor-driven multiple-choice assessments aligned to ideas about energy forms, transformation, transfer, and conservation. In R. F. Chen, A. Eisenkraftt, D. Fortus, J. Krajcik, K. Neumann, J. Nordine, & A. Scheff (Eds.), *Teaching and learning of energy in K-12 education* (pp. 103-133).

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141-158. doi:10.1119/1.2343497

Ishimoto, M., Davenport, G., Wittman, M. C. (2017). Use of item response curves of the Force and Motion Conceptual Evaluation to compare Japanese and American students' views on force and motion. *Physical Review Physics and Education Research*, *13*(2), 1-15. doi:10.1103/PhysRevPhysEducRes.13.020135

Jorion, N., Gane, B. D., James, K., Schroder, L., DiBello, L., & Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education*, *104*(4), 454-494. doi:10.1002/jee.20104

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527-535. doi:10.1037/0033-2909.112.3.527

Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2016). Developing and validating cognitive models in assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 75-101). New York, NY: John Wiley & Sons, Inc.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Kulkarni, A. V., Aziz, B., Shams, I., & Bussee, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *Journal of the American Medical Association*, *302*(10), 1092-1096. doi:10.1001/jama.2009.1307

Kuo, C.-Y., Wu, H.-K., Jen. T.H., & Hsu, Y.-S. (2015). Development and validation of a multimedia-based assessment of science inquiry abilities. *International Journal of Science Education*, *37*(14), 2326-2357. doi:10.1080/09500693.2015.1078521

Laliyo, L. A. R., Botutihe, D., N., & Panigoro, C. (2019). The development of two-tier instrument based on distractor to assess conceptual understanding level and student misconceptions in explaining redox reactions. *International Journal of Learning, Teaching and Educational Research*, *18*(9), 216-237. doi:10.26803/ijlter.18.9.12

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lee, H.-S., & Liu, O. L. (2009). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*(4), 665-688. doi:10.1002/sce.20382

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, *24*(2), 115-136. doi:10.1080/08957347.2011.554604

Lehrer, R., & Schauble, L. (2015). Learning progressions: The whole world is NOT a stage. *Science Education*, *99*(3), 432–437. doi:10.1002/sce.2015.99.issue-3

Linacre, J. M. (2020). A user's guide to WINSTEPS MINISTEP: Rasch-model computer programs. Retrieved from https://www.winsteps.com/a/Winsteps-Manual.pdf

Liu, X., & Lesniak, K. (2005). Students' progression of understanding the matter concept from elementary to high school. *Science Education*, *89*(3), 433-450. doi:10.1002/sce.20056

Liu, O. L., Lee, H.-S., & Linn, M. C. (2011a). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment*, *16*(3), 164-184. doi:10.1080/10627197.2011.611702

Liu, O. L., Lee, H.-S., & Linn, M. C. (2011b). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching*, *48*(9), 1079-1107. doi:10.1002/tea.20441

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.

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

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1-27. doi:10.1016/0022-2496(64)90015-X

MacPherson, A. (2015). *Beyond answer choice B: The development of an assessment of argumentation in ecology* (Unpublished doctoral dissertation). Stanford University, Palo Alto, CA.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491-511. doi:10.1177/0013164414539162

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/BF02296272

Masters, G. N., & Forster, M. (1996). *Progress maps: Assessment resource kit*. Melbourne, Australia: The Australian Council for Educational Research Ltd.

Michell, J. (1997). Quantitative science the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355-383. doi:10.1111/j.2044-8295.1997.tb02641.x

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In National Research Council, *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 44-73). Washington, DC: National Academies Press.

Minstrell, J., Anderson, R., & Li, M. (2016). Diagnostic instruction: Toward an integrated system for classroom assessment. In R. Duschl & A. S. Bismack (Eds.), *Reconceptualizing STEM education: The central role of practices* (pp. 49-67). New York, NY: Routledge.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*. (ETS Research Report RR-03-16). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to developing a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8). 1024-1048. doi:10.1002/tea.21397

Morris, G. A., Branum-Martin, L., Harshman, N., Baker, S. D., Mazur, E., Dutta, S., Mzoughi, T., & McCauley, V. (2006). Testing the test: Item response curves and test quality. *American Journal of Physics*, *74*(5), 449-453. doi:10.1119/1.2174053

Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). *American Journal of Physics*, *80*(9), 825-831. doi:10.1119/1.4731618

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, *22*(4), 13-25. doi:10.1111/j.1745-3992.2003.tb00140.x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176. doi:10.1177/014662169201600206

National Academies of Science, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures.* Washington, DC: The National Academies Press.

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

National Research Council. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: The National Academies Press.

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

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

National Research Council. (2013). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.

Newton, P., & Shaw, S. (2015). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in education: Principles, policy, & practice*, *23*(2), 178-197. doi:10.1080/0969594X.2015.1037241

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* Washington, DC: The National Academies Press.

Osborne, J. W. (2014). *Best practices in exploratory factor analysis*. CreateSpace Independent Publishing Platform.

Penuel, W. R. (2015). Learning progressions as evolving tools in joint enterprises for educational improvement. *Measurement: Interdisciplinary research and perspectives*, *13*(2), 123-127. doi:10.1080/15366367.2015.1055145

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (pp. 787-850). Washington, DC: American Educational Research Association.

Piaget, J. (1936). *Origins of intelligence in the child.* London: Routledge & Kegan Paul.

Plummer, J. D., & Maynard, L. (2014). Building a learning progression for celestial motion: An exploration of students' reasoning about the seasons. *Journal of Research in Science Teaching*, *51*(7), 902-929. doi:10.1002/tea.21151

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211-227. doi:10.1002/sce.3730660207

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

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

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Revelle, W. (2020). *Procedures for psychological, psychometric, and personality research*. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in earth science. *Journal of Research in Science Teaching*, *49*(6), 713-743. doi:10.1002/tea.21029

Rizopoulos, D. (2018). *Latent trait models under IRT*. Retrieved from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford 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*(3), 265-296. doi:10.1002/(SICI)1098-2736(199803)35:3<265::AID-TEA3>3.0.CO;2-P

Sadler, P. M. (2005). The relevance of multiple-choice tests in assessing science understanding. In J. Mintzes, J. Wandersee, & J. Novak (Eds.), *Assessing science understanding* (pp. 249-278). Cambridge, MA: Academic Press.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 18.

Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning. *Applied Measurement in Education*, *26*(3), 191-204. doi:10.1080/08957347.2013.793185

Shavelson, R. J., & Kurpius, A. (2012). Reflections on learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 345-355). Boston, MA: Sense Publishers.

Shear, B. R., & Roussos, L. A. (2017). Validating a distractor-driven geometry test using a generalized diagnostic classification model. In B. D. Zumbo & A. M. Hubley, *Understanding and investigating response processes in validation research* (pp. 277-304). Cham, Switzerland: Springer.

Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in education: Principles, policy, & practice*, *23*(2), 268-280. doi:10.1080/0969594X.2016.1141168

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, *31*(2), 165-172. doi:10.1080/08957347.2017.1408628

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21-34. doi:10.1111/emip.12189

Sinatra, G. M., & Pintrich, P. R. (Eds.) (2003). *Intentional conceptual change*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, *22*(4). doi: 10.1111/j.1745-3992.2003.tb00141.x

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, *51*(3), 541-565. doi: 10.1177/0013164491513003

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3*(2), 115-163. doi:10.1207/s15327809jls0302_1

Steedle, J. T. (2008). *Latent class analysis of diagnostic science assessment data using Bayesian networks* (Unpublished doctoral dissertation). Stanford University, Palo Alto, CA.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, *46*(6), 699-715. doi:10.1002/tea.20308

Stevens, S. Y., Delgado, C., & Krajcik, J. S. (2010). Developing a hypothetical multi-dimensional learning progression for the nature of matter. *Journal of Research in Science Teaching*, *47*(6), 687-715. doi:10.1002/tea.20324

Stiggins, R. J. (1999). Assessment, student confidence, and school success. *Phi Delta Kappan*, *81*, 191-198.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). New York, NY: Pearson.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345-354.

Testa, I., Capasso, G., Colantonio, A., Galano, S., Marzoli, I., Di Uccio, U. S., Trani, F., & Zappia, A. (2019). Development and validation of a university students' progression in learning quantum mechanics through exploratory factor analysis and Rasch analysis. *International Journal of Science Education*, *41*(3), 388-417. doi:10.1080/09500693.2018.1556414

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501-519. doi:10.1007/BF02302588

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567-577. doi:10.1007/BF02295596

Thissen-Roe, A., Hunt, E., & Minstrell, J. (2004). The DIAGNOSER project: Combining assessment and learning. *Behavior Research Methods, Instruments, & Computers, 36*(2), 234-240. doi:10.3758/BF03195568

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

Todd, A., & Kenyon, L. (2016). Empirical refinements of a molecular genetics learning progression: Molecular constructs. *Journal of Research in Science Teaching*, *53*(9), 1385-1418. doi:10.1002/tea.21262

van der Linden, W. J., & Hambleton, R. K. (1997). The nominal categories model. *Handbook of modern item response theory*. New York, NY: Springer-Verlag.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, *4*(1), 45-69. doi:10.1016/0959-4752(94)90018-3

Vosniadou, S. (2007). Conceptual change and education. *Human Development*, *50*(1), 47-54. doi:10.1159/000097684

Vosniadou, S., Vamvakoussi, X., & Skopeliti, I. (2008). The framework theory approach to the problem of conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 3-34). New York, NY: Routledge.

Vygotsky, L. S. (1978). *Mind in society: The development of the higher psychological processes.* Cambridge, MA: The Harvard University Press.

Wallace, C. S. (2011). *An investigation into introductory astronomy students' difficulties with cosmogony, and the development, validation, and efficacy of a new suite of cosmology lecture-tutorials* (Unpublished doctoral dissertation). University of Colorado Boulder, Boulder, CO.

Weinberg, P. J. (2012). *Assessing causal mechanistic reasoning: Promoting system thinking* (Unpublished doctoral dissertation). Vanderbilt University, Nashville, TN.

Wickham, H. (2019). *Easy install and load the 'Tidyverse'*. Retrieved from https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf

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

Wind, S. A., &. Gale, J. D. (2015). Diagnostic opportunities using Rasch measurement tin the context of a misconceptions-based physical science assessment. *Science Education*, *99*(4), 721-741. doi:10.1002/sce.21172

Wind, S. A., Alemdar, M., Lingle, J. A., Moore, R., & Asilkalkan, A. (2019). Exploring student understanding of the engineerring design process using distractor analysis. *International Journal of STEM Education*, *6*(4), 1-18. doi:10.1186/s40594-018-0156-x

Wren, D., & Barbera, J. (2014). Psychometric analysis of the thermochemistry concept inventory. *Chemistry Education Research and Practice*, *15*(3), 380-390. doi:10.1039/C3RP00170A

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

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*(4), 339-355.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments. Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213. doi:10.1111/j.1745-3984.1993.tb00423.x

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). Lanham, MD: Rowman & Littlefield.

Appendix A

Exploring the Impact of Sample Size on the Estimation of NRM Slope Parameters

This appendix presents the results of a simulation study to explore the impact of sample size on the estimation of NRM slope parameters. Prior research by De Ayala and Sava-Bolesta (1999) found that the ratio of the sample size to the total number of item parameters was important in NRM parameter estimation. DeMars (2004) extended De Ayala and Sava-Bolesta's research by conducting a simulation study to explore the effect of sample size ratio (i.e., the ratio of sample size to the total number of item parameters) on the recovery of NRM parameters. In the NRM, the total number of item parameters is the product of the number of items and the number of categories per item. DeMars explored two sample sizes (2400 and 600), two conditions of total item parameters (120 and 240), and two conditions of parameters per item (6 parameters per item and 12 parameters per item). She found that if one wanted an average root mean square error of no more than 0.10, this could be obtained for most conditions with 2400 cases. DeMars did not propose a heuristic rule of what sample size to category ratio is adequate.

Interpreting NRM response option curves and the order of slope parameters would only be sensible if there was evidence to suggest that there was enough data to precisely estimate the slope parameters. The goal of the simulation presented in this section is to identify ranges of sample sizes where NRM slope parameters would be estimated with imprecision because of inadequate overall sample size. In simulation studies involving the NRM, there are a great deal of independent variables that can be manipulated (e.g., ability distribution, number of items, number of categories per item, overall sample size, number of students choosing a response option, etc.). For this simulation, I use the simdata() function from the mirt package (Chalmers,

2012, 2019) with fixed inputs to simulate data using the NRM. The independent variable is sample size, and I vary it from 100 to 7000 in increments of 100. I constrain the total number of NRM parameters to the number of parameters for Test A (7 items, 26 total item parameters) and Test B (8 items, 26 total item parameters) and use the estimates in Table 6.3 as "true" values for the parameters. For Test C which combines all items into a super-test, I combine the parameter estimates from Tests A and B to identify "true" parameter inputs. Using these inputs, I simulate data using simdata() for each sample size (i.e., 100, 200, etc.). I then fit the NRM to the simulated data and store these estimates, repeating this procedure 100 times.

To analyze the NRM slope parameters, I plot the average root mean square error for each value of the independent variable to identify the sample sizes where error for each parameter is smallest. The root mean square error ($RMSE$) is a popular statistic in IRT simulation studies because it combines information about the average distance between the estimated parameter ($\hat{\phi}_j$) and true parameter (ϕ_{True}) across n replications (i.e., $Bias$) and the deviation of the estimated parameter across replications (i.e., the standard error or SE of the parameter):

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(\hat{\phi}_j - \phi_{True})^2}{n-1}} = \sqrt{Bias^2 + SE^2} \quad \text{(A.1)}$$

In this simulation, I first compute the RMSE for each NRM slope parameter by comparing the "true" slope to the recovered slope parameter for each value of the independent variable. Then, for ease of interpretation, I average the $RMSE$ estimates within each unique sample size. The results for Test A, B, and C are plotted in the following figure.

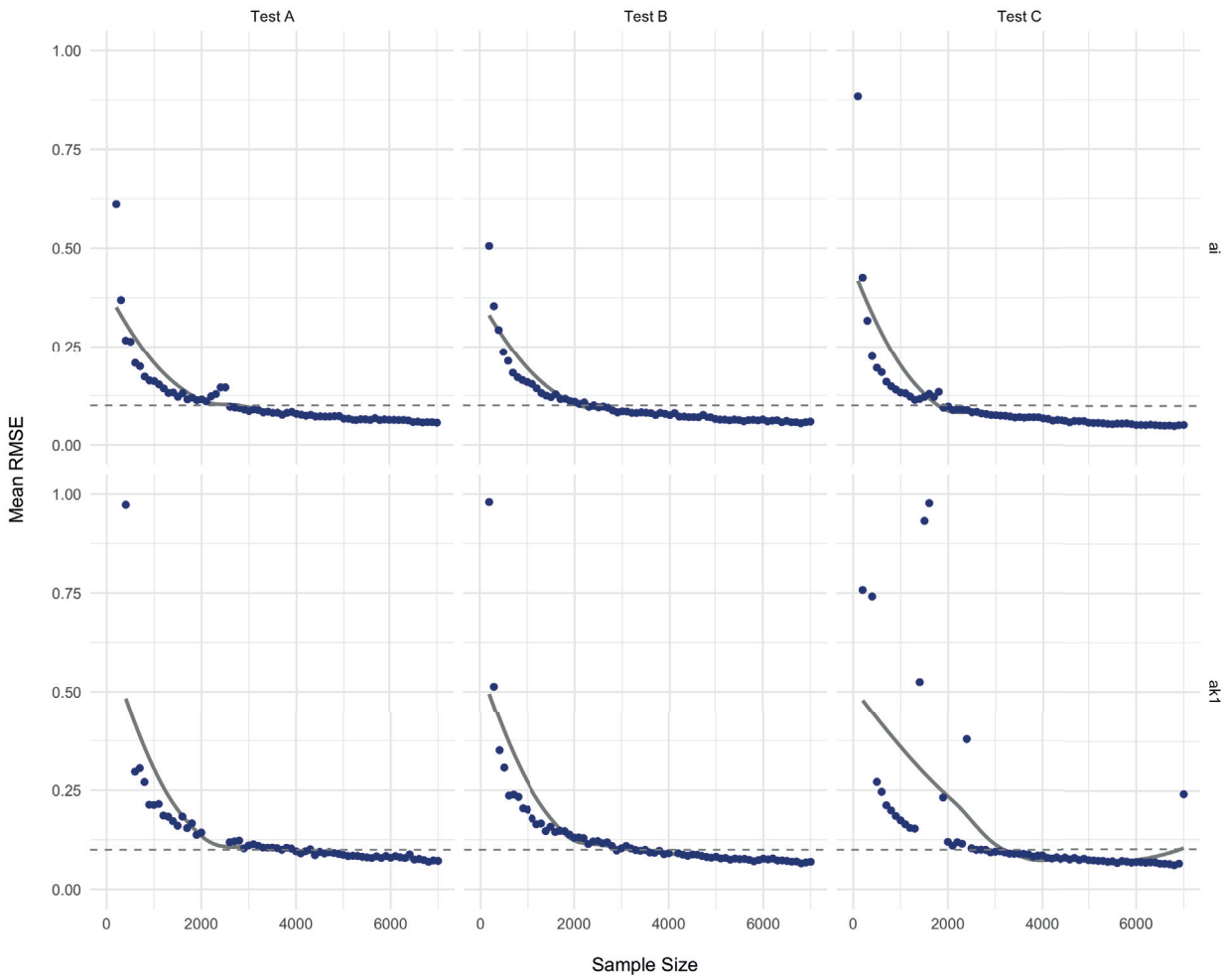


Figure A.1. Average RMSE for Tests A, B, and C.

The above plots illustrate that approximately 2000 observations are required to estimate NRM item specific slope parameters (a_i's) for the Diagnoser tests with an RMSE of about 0.10 or smaller and about 3000 observations are required to estimate the response option slopes (ak_1's) with an RMSE of about 0.10 or smaller. There are no noticeable differences across tests. I choose the RMSE threshold of 0.10 since that was the value chosen by DeMars. The empirical NRM slope estimates presented in Table 6.3 are generally lower than 0.10, but there are cases where the SE exceeds this value. For example, item 2.1 has an SE for a_i and ak_1 equal to 0.18

(see Table 6.3). These elevated SEs are driven by uneven numbers of students selecting responses associated with the three categories. Of the 2387 students that answered item 2.1, just 30 selected a response associated with the lowest category, 405 selected a response associated with the middle category, and 1952 students chose the correct answer.

As indicated in Table 5.1, the empirical sample sizes for Tests A, B, and C are 5063, 2387, and 1105, respectively. In general, the NRM item parameters in Tables 6.1 and 6.3 for Test A are more precisely estimated than the parameters for Test B because Test A has a larger overall sample size than Test B. If used with data for Test C, the NRM item parameter estimates would contain a sizable amount of error, and this can be seen by the relatively large standard errors for some item parameters in Table D.5 (see Appendix D). Using the NRM to estimate item parameters for the Diagnoser tests with fewer than 1,000 cases will likely result in even more sampling error.

Appendix B

Diagnoser Items

This appendix presents the Diagnoser items that were written to evaluate student thinking relative to the atoms facet cluster (Figure 2.5). The following instructions are provided to teachers on www.diagnoser.com regarding how to use the items contained in the question sets:

> *Diagnoser is a computer program that poses questions to your students via the internet. The questions are grouped into Diagnoser sets of questions on particular topics (e.g., Effects of Pushes and Pulls). Diagnoser questions are designed to elicit different facets of student. Thinking. The program provides feedback to students to assist them in monitoring their own learning. After students have completed a set of questions, teachers are able to view the facets diagnosed for each question for each student.*
>
> *Assign the first or early sets about mid-way through your unit shortly after the ideas have been developed, but before the student has begun the application lessons or related projects. This way the results of the assessment can inform subsequent instruction on the topic. Additional Diagnoser sets can be assigned (if needed) later in the unit as a way for students to check on their own learning. Some teachers may also assign sets at the beginning of a review unit on topics covered in previous courses to see if students are ready for a next level of instruction.*

These instructions represent the intended use of Diagnoser. Teachers can use the items as they see fit during the course of instruction. This appendix presents the full text of the items, the nominal score assigned to each response option during the exploration of the initial scoring hypothesis in curly brackets, and the mapping of facets to answer choices in square brackets.

Atoms Question Set 1

1.1 Which of the following are made of atoms?

I. Bacteria
II. Firewood
III. Molecule

a. {0} I only [*Facet 80*]
b. {1} II only [*Facet 80*]
c. {2} III only [*Facet Unknown*]
d. {5} I and II [*Facet 80*]
e. {3} I and III [*Facet 80*]
f. {4} II and III [*Facet 82*]
g. {6} All of them [*Facet 01*]

1.2. Assume you have the technology to cut a piece of gold into the smallest piece possible.

What would you be left with?

a. {0} An incredibly tiny piece of gold. [*Facet 81*]
b. {3} A single atom of gold that is too small to be seen. [*Facet 01*]
c. {1} Nothing, because there is nothing there when you get that small. [*Facet 83*]
d. {2} None of the above. [*Facet Unknown*]

1.3. Select the answer that most accurately compares the size of the objects below.

I. atom of carbon
II. speck of dust

a. {1} These two objects are about the same size. [*Facet 51*]
b. {2} These two objects are close in size, but an atom of carbon is larger. [*Facet 51*]
c. {3} These two objects are close in size, but a speck of dust is larger. [*Facet 51*]
d. {4} A speck of dust is much bigger than an atom of carbon. [*Facet 02*]
e. {0} An atom of carbon is much bigger than a speck of dust. [*Facet 80*]

1.4. Select the answer that most accurately compares the size of the objects below.

I. atom of iron
II. cold virus

a. {1} These two objects are the same size. [*Facet 52*]
b. {2} These two objects are close in size, but an atom of iron is slightly larger. [*Facet 52*]
c. {3} These two objects are close in size, but a cold virus is slightly larger. [*Facet 52*]
d. {4} A cold virus is much bigger than an atom of iron. [*Facet 02*]
e. {0} An atom of iron is much bigger than a cold virus. [*Facet 82*]

1.5. Carbon, nitrogen, and hydrogen atoms are added together in a closed flask. After they are mixed, a chemical reaction happens and a thin film forms on the bottom of the flask.

Does the number of atoms in the flask increase, decrease, or remain the same?

a. {0} The number and types of atoms increase. [*Facet 42*]
b. {1} The number and types of atoms decrease. [*Facet 41*]
c. {3} The number and types of atoms remain the same. [*Facet 03*]
d. {2} The number of atoms remains the same, but the types of atoms change. [*Facet 40*]

1.6. A piece of paper burns in a closed flask. As it burns, does the number of atoms in the flask increase, decrease, or remain the same?

a. {1} The number and type of atoms increase. [*Facet 42*]
b. {2} The number and type of atoms decrease. [*Facet 41*]
c. {4} The number and type of atoms remain the same. [*Facet 03*]
d. {3} The number of atoms remains the same but the types of atoms change. [*Facet 40*]
e. {0} None of the above. The paper is not made up of atoms. [*Facet 80*]

1.7. Three students were discussing atoms and cells during biology class.

Brett: "*Non-living things are made of atoms. Living things are made of cells, not atoms.*"
James: "*Only some non-living things are made of atoms.*"
Steve: "*All living and non-living things are made of atoms.*"

With which student do you agree?

a. {0} Brett [*Facet 82*]
b. {1} James [*Facet 80*]
c. {2} Steve [*Facet 01*]

Atoms Question Set 2

2.1. Which of the following are made of atoms?

I. Molecule
II. Virus
III. Paper

a. {4} I only [*Facet Unknown*]
b. {0} II only [*Facet 80*]
c. {1} III only [*Facet 82*]
d. {2} I and II [*Facet 80*]
f. {3} II and III [*Facet 80*]
g. {5} All of them [*Facet 01*]

2.2. Select the answer that most accurately compares the size of the objects below.

I. atom of mercury
II. red blood cell

a. {1} These two objects are the same size. [*Facet 52*]
b. {2} These two objects are close in size, but an atom of mercury is slightly larger. [*Facet 52*]
c. {3} These two objects are close in size, but a red blood cell is slightly larger. [*Facet 52*]
d. {4} A red blood cell is much bigger than an atom of mercury. [*Facet 02*]
e. {0} An atom of mercury is much bigger than a red blood cell. [*Facet 82*]

2.3. Assume you have the technology to cut a piece of aluminum foil into the smallest piece of aluminum possible.

What would you be left with? Choose the best response.

a. {0} An incredibly tiny piece of aluminum foil. [*Facet 81*]
b. {3} A single atom of aluminum that is too small to be seen. [*Facet 01*]
c. {1} Nothing, because there is nothing to see when you get that small. [*Facet 83*]
d. {2} None of the above. [*Facet Unknown*]

2.4. A flask is filled with yellow-green chlorine gas. A piece of sodium metal is lowered into the flask. Then water is dropped on the sodium metal and the flask is closed. The metal bursts into flames and the flask is filled with a white cloudy looking substance. This substance is salt.

In this closed system, how does the number of chlorine atoms in the flask before the salt forms compare to the number of chlorine atoms in the flask after the salt forms?

a. {0} The number of atoms increases. [*Facet 42*]
b. {1} The number of atoms decreases. [*Facet 41*]
c. {3} The number of atoms remains the same. [*Facet 03*]
d. {2} The number of atoms remains the same, but the types of atoms change. [*Facet 40*]

2.5. Select the answer that most accurately compares the size of the objects below.

I. particle of dust in the air
II. atom of silver

a. {1} These two objects are about the same size. [*Facet 51*]
b. {2} These two objects are close in size, but an atom of silver is larger. [*Facet 51*]
c. {3} These two objects are close in size, but the particle of dust in the air is larger. [*Facet 51*]
d. {4} The particle of dust in the air is much bigger than an atom of silver. [*Facet 02*]
e. {0} An atom of silver is much bigger than the particle of dust in the air. [*Facet 80*]

2.6.a. Ali mixed 50 ml of alcohol with 50 ml of water. No reaction occurred and neither of the liquids evaporated. She was surprised to notice that the final volume of the alcohol-water solution was less than 100 ml.

Suppose that Ali weighs the alcohol and the water before mixing and then weighs the alcohol-water solution after mixing.

How does the weight of the liquids compare before and after they are mixed?

a. The alcohol-water solution after mixing weighs less. [*Next Item 6b*]
b. The alcohol-water solution after mixing weighs more. [*Next Item 6b*]
c. They weigh the same before and after mixing. [*Next Item 6b*]

2.6.b. Which of the following statements best matches your reasoning on the previous question?

a. There are fewer atoms in the mixed solution compared to the number of atoms in the alcohol and water before mixing. [*Facet 41*]
b. There are more atoms in this mixed solution compared to the number of atoms in the alcohol and water before mixing, but the atoms are just more tightly packed in the mixed solution. [*Facet 42*]
c. There is the same number of atoms before and after mixing the alcohol and water, but the atoms are just more tightly packed in the mixed solution. [*Facet 00 if chose c for Item 6, otherwise Facet Unknown*]

Item 2.6.a.	Item 2.6.b.	Initial Score
a	a	{0}
b	a	{1}
c	a	{2}
a	b	{3}
b	b	{4}
c	b	{5}
a	c	{6}
b	b	{7}
c	c	{8}

2.7. Two students were discussing their science class today.

Rafael said, "*All matter like air, water, rocks, and trees is made of atoms.*"

Frank said, "*All matter like air, water, and rocks is made of atoms, but living things like trees are made of cells, not atoms.*"

With which student do you agree?

a. {3} Rafael [*Facet 01*]
b. {0} Frank [*Facet 82*]
c. {2} Both [*Facet Unknown*]
d. {1} Neither [*Facet 80*]

2.8. From the list below, select the smallest pieces in a living organism.

a. {0} Cells [*Facet 82*]
b. {1} DNA [*Facet 82*]
c. {2} Atoms [*Facet 01*]

Appendix C

Empirical Item Correlation Matrices

This appendix contains the empirical item correlation matrices for the three analytic samples introduced in Chapter 5. The cells consist of polychoric correlations, which assume that two ordinal variables (i.e., responses to two items on the test) dissect continuous latent variables that have a bivariate normal distribution. To estimate the polychoric correlations, I used dichotomously scored Diagnoser data with the polychor function in the R package polycor (Fox, 2019). The function uses a maximum likelihood to estimate the correlations.

Table C.1.

Item Correlation Matrix for Test 1.

Item	1.1.	1.2.	1.3.	1.4.	1.5.	1.6.	1.7.
1.1.	1.00						
1.2.	0.21	1.00					
1.3.	0.43	0.31	1.00				
1.4.	0.32	0.18	0.30	1.00			
1.5.	0.21	0.12	0.26	0.17	1.00		
1.6.	0.28	0.19	0.36	0.29	0.28	1.00	
1.7.	0.44	0.25	0.46	0.35	0.15	0.34	1.00

Table C.2.

Item Correlation Matrix for Test 2.

Item	2.1.	2.2.	2.3.	2.4.	2.5.	2.6.	2.7.	2.8.
2.1.	1.00							
2.2.	0.50	1.00						
2.3.	0.44	0.40	1.00					
2.4.	0.39	0.41	0.36	1.00				
2.5.	0.52	0.74	0.42	0.42	1.00			
2.6.	0.43	0.39	0.36	0.41	0.45	1.00		
2.7.	0.54	0.42	0.34	0.38	0.46	0.39	1.00	
2.8.	0.51	0.49	0.33	0.29	0.45	0.31	0.47	1.00

Table C.3.

Item Correlation Matrix for Test 3.

Item	1.1.	1.2.	1.3.	1.4.	1.5.	1.6.	1.7.	2.1.	2.2.	2.3.	2.4.	2.5.	2.6.	2.7.	2.8.
1.1.	1.00														
1.2.	0.21	1.00													
1.3.	0.46	0.34	1.00												
1.4.	0.33	0.21	0.38	1.00											
1.5.	0.29	0.07	0.29	0.19	1.00										
1.6.	0.27	0.12	0.36	0.21	0.29	1.00									
1.7.	0.43	0.24	0.43	0.35	0.21	0.36	1.00								
2.1.	0.56	0.23	0.49	0.20	0.15	0.31	0.59	1.00							
2.2.	0.31	0.18	0.52	0.45	0.06	0.25	0.47	0.53	1.00						
2.3.	0.28	0.41	0.43	0.23	0.06	0.28	0.38	0.55	0.45	1.00					
2.4.	0.20	0.16	0.28	0.15	0.28	0.44	0.34	0.50	0.43	0.42	1.00				
2.5.	0.33	0.20	0.63	0.47	0.13	0.35	0.50	0.57	0.77	0.49	0.50	1.00			
2.6.	0.27	0.20	0.37	0.21	0.21	0.31	0.49	0.47	0.32	0.36	0.43	0.44	1.00		
2.7.	0.36	0.19	0.39	0.24	0.14	0.29	0.54	0.53	0.38	0.41	0.44	0.51	0.39	1.00	
2.8.	0.33	0.16	0.37	0.29	0.12	0.21	0.48	0.53	0.51	0.37	0.38	0.47	0.36	0.46	1.00

Appendix D

NRM Parameter Estimates

This appendix presents NRM item parameter estimates obtained by fitting the NRM to the Diagnoser data. I use the function mirt() to estimate NRM item parameters using the model specification presented in equation 4.3. From a programming perspective, the functional form of the NRM given by equation 4.3 is more flexible than Bock's original specification (equation 4.2) because other IRT models can be derived from equation 4.3 by relaxing or constraining the model parameters. Multidimensional models can be obtained by allowing the item-specific slopes (a_i) and person abilities (θ_p) to vary by dimension. Constraining the response option slope parameters (ak_{ik-1}) to the expected ordinal values for the scores while permitting the item-specific slopes (a_i) to be freely estimated results in the generalized partial credit model (Muraki, 1992). Constraining the ak_{ik-1} parameters to the expected scores *and* constraining the a_i parameters across items to be constant produces the PCM (Masters, 1982).

The following tables present the mirt NRM intercepts (Chalmers, 2012, 2020) and the Bock (1972) response option slopes and intercepts for all three analytic samples. The mirt NRM slopes are presented and interpreted in Chapter 6. Standard errors are included for all of the mirt item parameters. No response categories were collapsed in the initial calibration, but categories are collapsed in the revised calibration as described in Chapter 6. Chapter 6 illustrates how the item parameters for select items can be interpreted together with the theory used to design the items to revise the initial scoring hypothesis. NRM category intersection parameters for the initial scoring hypothesis can be calculated using the Bock item parameters (Tables D.2 and D.3) and equation 4.5, but they are omitted from the following tables for clarity.

Table D.1.

Chalmers NRM Intercept Parameters from Initial Calibration.

Item	Chalmers NRM Intercept Parameter*								
	d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
1.1.	0	0.53	1.92	1.91	1.77	2.32	4.90	-	-
	(-)	(0.33)	(0.28)	(0.28)	(0.28)	(0.27)	(0.26)	-	-
1.2.	0	-1.87	-0.67	1.76	-	-	-	-	-
	(-)	(0.14)	(0.07)	(0.05)	-	-	-	-	-
1.3.	0	0.92	0.55	2.20	3.92	-	-	-	-
	(-)	(0.20)	(0.22)	(0.18)	(0.18)	-	-	-	-
1.4.	0	0.32	-0.04	1.70	2.66	-	-	-	-
	(-)	(0.12)	(0.13)	(0.10)	(0.10)	-	-	-	-
1.5.	0	-1.48	0.58	0.62	-	-	-	-	-
	(-)	(0.10)	(0.05)	(0.05)	-	-	-	-	-
1.6.	0	1.90	2.69	3.02	3.87	-	-	-	-
	(-)	(0.20)	(0.20)	(0.20)	(0.20)	-	-	-	-
1.7.	0	0.25	3.47	-	-	-	-	-	-
	(-)	(0.15)	(0.13)	-	-	-	-	-	-
n = 5,063									
2.1.	0	1.88	2.72	2.90	2.51	6.29	-	-	-
	(-)	(0.63)	(0.60)	(0.60)	(0.60)	(0.59)	-	-	-
2.2.	0	-0.42	-0.71	0.85	3.10	-	-	-	-
	(-)	(0.25)	(0.27)	(0.19)	(0.17)	-	-	-	-
2.3.	0	-1.48	-1.07	2.22	-	-	-	-	-
	(-)	(0.2)	(0.17)	(0.09)	-	-	-	-	-
2.4.	0	-0.52	0.99	1.70	-	-	-	-	-
	(-)	(0.14)	(0.10)	(0.10)	-	-	-	-	-
2.5.	0	0.15	-0.42	0.79	3.35	-	-	-	-
	(-)	(0.24)	(0.27)	(0.21)	(0.19)	-	-	-	-
2.6.	0	-1.73	-1.88	-0.82	0.22	-0.93	0.38	-0.49	1.49
	(-)	(0.22)	(0.22)	(0.15)	(0.11)	(0.16)	(0.10)	(0.12)	(0.08)
2.7.	0	-0.06	0.22	3.09	-	-	-	-	-
	(-)	(0.19)	(0.18)	(0.15)	-	-	-	-	-
2.8.	0	0.01	2.81	-	-	-	-	-	-
	(-)	(0.16)	(0.12)	-	-	-	-	-	-
n = 2,387									

* Standard errors are presented below the parameter estimates.

Table D.2.

Bock NRM Slope Parameters from Initial Calibration.

Item	Bock NRM Slope Parameter								
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
1.1.	-0.90	-0.52	-0.36	-0.36	0.48	0.40	1.27	-	-
1.2.	-0.02	-0.73	0.12	0.63	-	-	-	-	-
1.3.	-0.46	-0.47	-0.81	0.15	1.58	-	-	-	-
1.4.	-0.45	-0.33	-0.67	0.47	0.97	-	-	-	-
1.5.	-0.35	-0.96	0.54	0.76	-	-	-	-	-
1.6.	-1.11	-0.42	-0.14	0.51	1.17	-	-	-	-
1.7.	-0.60	-0.51	1.11	-	-	-	-	-	-
n = 5,063									
2.1.	-1.30	-0.29	-0.17	0.84	-0.72	1.63	-	-	-
2.2.	-0.16	-0.50	-1.10	-0.10	1.86	-	-	-	-
2.3.	-0.08	-0.69	-0.22	0.99	-	-	-	-	-
2.4.	-0.69	-0.76	0.40	1.06	-	-	-	-	-
2.5.	-0.36	-0.57	-0.80	-0.18	1.91	-	-	-	-
2.6.	0.25	-0.89	-0.13	-0.62	-0.40	-0.44	0.79	0.18	1.26
2.7.	-0.77	0.11	-0.37	1.03	-	-	-	-	-
2.8.	-0.43	-0.50	0.93	-	-	-	-	-	-
n = 2,387									

Table D.3.

Bock NRM Intercept Parameters from Initial Calibration.

Item	Bock NRM Intercept Parameter								
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
1.1.	-1.91	-1.38	0.01	0.00	-0.13	0.41	2.99	-	-
1.2.	0.20	-1.68	-0.48	1.96	-	-	-	-	-
1.3.	-1.52	-0.60	-0.97	0.69	2.40	-	-	-	-
1.4.	-0.93	-0.61	-0.97	0.77	1.74	-	-	-	-
1.5.	0.07	-1.41	0.66	0.69	-	-	-	-	-
1.6.	-2.30	-0.40	0.39	0.73	1.57	-	-	-	-
1.7.	-1.24	-0.99	2.23	-	-	-	-	-	-
n = 5,063									
2.1.	-2.72	-0.84	0.00	0.18	-0.21	3.58	-	-	-
2.2.	-0.57	-0.98	-1.27	0.29	2.54	-	-	-	-
2.3.	0.08	-1.40	-0.99	2.30	-	-	-	-	-
2.4.	-0.54	-1.06	0.45	1.16	-	-	-	-	-
2.5.	-0.77	-0.62	-1.19	0.01	2.57	-	-	-	-
2.6.	0.42	-1.31	-1.46	-0.40	0.64	-0.51	0.79	-0.07	1.91
2.7.	-0.81	-0.87	-0.59	2.27	-	-	-	-	-
2.8.	-0.94	-0.93	1.87	-	-	-	-	-	-
n = 2,387									

Table D.4.

NRM Parameters from Revised Calibration.

Item	Chalmers NRM Intercepts					Bock Slopes and Intercepts					
	d_0	d_1	(SE)	d_2	(SE)	a_1	a_2	a_3	c_1	c_2	c_3
1.1.	0	2.87	(0.20)	4.31	(0.20)	-0.69	-0.31	1.00	-2.39	0.48	1.91
1.2.	0	2.36	(0.14)	3.71	(0.14)	-0.75	0.08	0.67	-2.02	0.34	1.69
1.3.	0	1.39	(0.12)	2.83	(0.11)	-0.92	-0.30	1.22	-1.41	-0.01	1.42
1.4.	0	1.28	(0.08)	1.98	(0.07)	-0.77	0.02	0.75	-1.09	0.19	0.90
1.5.	0	2.52	(0.09)	2.04	(0.10)	-0.88	0.17	0.71	-1.52	1.00	0.52
1.6.	0	3.69	(0.19)	3.78	(0.19)	-1.06	0.03	1.04	-2.49	1.20	1.29
1.7.	0	2.67	(0.09)	-	-	-0.84	0.84	-	-1.34	1.34	-
n = 5,063											
2.1.	0	3.68	(0.52)	5.97	(0.53)	-1.17	-0.29	1.46	-3.22	0.46	2.75
2.2.	0	0.67	(0.15)	2.66	(0.15)	-0.93	-0.55	1.48	-1.11	-0.44	1.55
2.3.	0	1.72	(0.19)	3.63	(0.19)	-0.71	-0.18	0.89	-1.78	-0.07	1.85
2.4.	0	1.85	(0.12)	2.17	(0.12)	-0.80	-0.13	0.93	-1.34	0.51	0.83
2.5.	0	0.78	(0.17)	2.92	(0.17)	-0.97	-0.66	1.63	-1.24	-0.45	1.69
2.6.	0	2.09	(0.12)	1.92	(0.12)	-0.85	-0.11	0.96	-1.34	0.75	0.59
2.7.	0	1.91	(0.09)	-	-	-0.72	0.72	-	-0.95	0.95	-
2.8.	0	2.10	(0.09)	-	-	-0.69	0.69	-	-1.05	1.05	-
n = 2,387											

Table D.5.

Chalmers NRM Parameters Using Revised Scoring Hypothesis and Same Day Sample.

Item	Chalmers NRM Item Parameter Estimate										
	a_i	(SE)	ak_0	ak_1	(SE)	ak_2	d_0	d_1	(SE)	d_2	(SE)
1.1.	0.66	(0.13)	0	0.42	(0.32)	2	0	2.46	(0.28)	3.71	(0.28)
1.2.	0.69	(0.10)	0	1.35	(0.16)	2	0	1.93	(0.24)	3.37	(0.24)
1.3.	0.97	(0.09)	0	0.46	(0.14)	2	0	1.15	(0.18)	2.35	(0.17)
1.4.	0.78	(0.08)	0	1.18	(0.12)	2	0	1.43	(0.16)	2.20	(0.16)
1.5.	0.71	(0.09)	0	1.44	(0.11)	2	0	2.43	(0.18)	2.15	(0.18)
1.6.	0.88	(0.12)	0	1.05	(0.14)	2	0	3.18	(0.28)	3.25	(0.28)
1.7.	1.85	(0.18)	0	1	-	-	0	2.70	(0.18)	-	-
2.1.	1.80	(0.30)	0	0.84	(0.19)	2	0	4.50	(0.94)	6.92	(0.95)
2.2.	1.02	(0.09)	0	0.36	(0.15)	2	0	0.48	(0.18)	2.14	(0.16)
2.3.	0.90	(0.11)	0	0.55	(0.19)	2	0	1.56	(0.27)	3.49	(0.26)
2.4.	0.83	(0.09)	0	0.62	(0.14)	2	0	1.46	(0.16)	2.08	(0.16)
2.5.	1.27	(0.12)	0	0.18	(0.14)	2	0	0.64	(0.21)	2.57	(0.21)
2.6.	0.85	(0.09)	0	0.77	(0.12)	2	0	1.89	(0.15)	1.64	(0.16)
2.7.	1.50	(0.14)	0	1	-	-	0	1.80	(0.12)	-	-
2.8.	1.41	(0.13)	0	1	-	-	0	1.87	(0.12)	-	-

n = 1,105

Table D.6.

Bock NRM Parameters Using Revised Scoring Hypothesis and Same Day Sample.

Item	NRM Item Parameter Estimate							
	a_1	a_2	a_3	c_1	c_2	c_3	b^*_{21}	b^*_{32}
1.1.	-0.53	-0.26	0.79	-2.06	0.40	1.66	-9.11	-1.20
1.2.	-0.77	0.16	0.60	-1.77	0.16	1.61	-2.08	-3.30
1.3.	-0.79	-0.34	1.14	-1.17	-0.02	1.18	-2.56	-0.81
1.4.	-0.83	0.09	0.74	-1.21	0.22	0.99	-1.55	-1.18
1.5.	-0.81	0.21	0.60	-1.53	0.90	0.62	-2.38	0.72
1.6.	-0.89	0.03	0.86	-2.14	1.03	1.11	-3.45	-0.10
1.7.	-0.93	0.93	-	-1.35	1.35	-	-1.45	-
2.1.	-1.71	-0.19	1.90	-3.81	0.69	3.11	-2.96	-1.16
2.2.	-0.81	-0.43	1.24	-0.87	-0.40	1.27	-1.24	-1.00
2.3.	-0.77	-0.27	1.04	-1.68	-0.12	1.81	-3.12	-1.47
2.4.	-0.72	-0.21	0.93	-1.18	0.28	0.90	-2.86	-0.54
2.5.	-0.92	-0.69	1.61	-1.07	-0.43	1.50	-2.78	-0.84
2.6.	-0.79	-0.13	0.92	-1.18	0.71	0.46	-2.86	0.24
2.7.	-0.75	0.75	-	-0.90	0.90	-	-1.20	-
2.8.	-0.71	0.71	-	-0.94	0.94	-	-1.32	-

n = 1,105