APPLYING ITEM RESPONSE THEORY METHODS TO DESIGN A LEARNING PROGRESSION-BASED SCIENCE ASSESSMENT

By

Jing Chen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods
Curriculum, Teaching, and Educational Policy

2012

ABSTRACT

APPLYING ITEM RESPONSE THEORY METHODS TO DESIGN A LEARNING PROGRESSION-BASED SCIENCE ASSESSMENT

By

Jing Chen

Learning progressions are used to describe how students' understanding of a topic progresses over time and to classify the progress of students into steps or levels. This study applies Item Response Theory (IRT) based methods to investigate how to design learning progression-based science assessments. The research questions of this study are: 1) how to use items in different formats to classify students into levels on the learning progression, 2) how to design a test to give good information about students' progress through the learning progression of a particular construct and 3) what characteristics of test items support their use for assessing students' levels.

Data used for this study were collected from 1500 elementary and secondary school students during 2009-2010. The written assessment was developed in several formats such as the Constructed Response (CR) items, Ordered Multiple Choice (OMC) and Multiple True or False (MTF) items. The followings are the main findings from this study.

The OMC, MTF and CR items might measure different components of the construct. A single construct explained most of the variance in students' performances. However, additional dimensions in terms of item format can explain certain amount of the variance in student performance. So additional dimensions need to be considered when we want to capture the differences in students' performances on different types of items targeting the understanding of the same underlying progression. Items in each item format need to be improved in certain ways to classify students more accurately into the learning progression levels.

This study establishes some general steps that can be followed to design other learning progression-based tests as well. For example, first, the boundaries between levels on the IRT scale can be defined by using the means of the item thresholds across a set of good items. Second, items in multiple formats can be selected to achieve the information criterion at all the defined boundaries. This ensures the accuracy of the classification. Third, when item threshold parameters vary a bit, the scoring rubrics and the items need to be reviewed to make the threshold parameters similar across items. This is because one important design criterion of the learning progression-based items is that ideally, a student should be at the same level across items, which means that the item threshold parameters (d_1, d_2 and d_3) should be similar across items.

To design a learning progression-based science assessment, we need to understand whether the assessment measures a single construct or several constructs and how items are associated with the constructs being measured. Results from dimension analyses indicate that items of different carbon transforming processes measure different aspects of the carbon cycle construct. However, items of different practices assess the same construct. In general, there are high correlations among different processes or practices. It is not clear whether the strong correlations are due to the inherent links among these process/practice dimensions or due to the fact that the student sample does not show much variation in these process/practice dimensions. Future data are needed to examine the dimensionalities in terms of process/practice in detail.

Finally, based on item characteristics analysis, recommendations are made to write more discriminative CR items and better OMC, MTF options. Item writers can follow these recommendations to write better learning progression-based items.

To My Father

ACKNOWLEDGEMENTS

The current work is a result of several years of effort, challenges and hard work. It would not be possible without the help and support of my advisors, colleagues, friends and family.

I am especially grateful to my two advisors, Professor Charles Anderson and Professor Mark Reckase. They guided me navigating through the complexity of educational research in both qualitative and quantitative dimensions. I joined in Andy's Environmental Literacy project since I started my graduate program five years ago. I am deeply influenced by his devotion and enthusiasm to teaching and research. His knowledge and expertise greatly broaden my vision in learning progression as well as the science education in general. I started working with Mark since I began my dual degree program in the Measurement and Quantitative Methods. I am the real beneficiary of Mark's teaching style, broad knowledge and the down-to-earth attitude towards research.

I would like to express my sincere appreciation to Prof. Christina Schwarz, Prof. Amelia Gotwals and Prof. Edward Roeber for their great comments and suggestions for my dissertation. I thank my colleagues in the Environmental Literacy project. We have been working together to code the 1,500 test papers. All the data used in this dissertation is the results of their hard work.

I want to thank my parents for their long time support and encouragement. Finally, I am own my deepest appreciation to my husband, Jiangang. His love and encouragement are indispensible for me to go through the challenges in my research and life.

PREFACE

First, this research is supported in part by grants from the National Science Foundation: Learning Progression on Carbon-Transforming Processes in Socio-Ecological Systems (NSF 0815993), and Targeted Partnership: Culturally Relevant Ecology, Learning Progressions and Environmental Literacy (NSF-0832173), and CCE: A Learning Progression-based System for Promoting Understanding of Carbon-transforming Processes (DRL 1020187). Additional support comes from the Great Lakes Bioenergy Research Center. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the United States Department of Energy.

Second, the focus of this study is assessment design. When the analyses indicate that an item is not well aligned with the learning progression framework, it is difficult to tell whether the problem is with the item, the rubric, or with the learning progression framework itself. The learning progression framework itself may not capture some alternate trajectories students take to achieve environmental literacy. Thus further investigating the validity of the learning progression framework is an important topic. However, this dissertation focuses on the problems with the assessment and the items rather than the learning progression framework itself. So I consider the problems from the assessment perspective and suggest improvements from the assessment perspective only.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Designing assessments is a complex process, involving numerous interdependent components. How to design high-quality science assessments is a tough question to answer. By examining the fundamental components of the carbon cycle assessment developed by the Environmental Literacy project in the past several years and the interplay among these components, this study aims at exploring the ways to design high-quality science assessments. In this chapter, I will firstly introduce the environmental and social background that shape the Environmental Literacy project, and then introduce the assessment developed by the project. Finally, I will elaborate the specific research purposes of this study.

1.1 Environmental and social background

The global climate is changing. The global surface temperature has increased 1.4°F since the beginning of the 20th century, with about 1.1ºF of the increase occurring in the past 30 years. The temperature will likely rise at least another 2°F, and possibly more than 11°F, in the next 100 years (IPCC, 2007). The predicted consequences of this increase include widespread melting of snow and ice, rising global average sea level, increasing frequency and severity of storms, and other effects on natural ecosystems and human agriculture (Crowley, 2000; Falkowski et al., 2000; Keeling & Whorf, 2005).

Most scientists agree that the warming in recent decades has been caused primarily by the increasing concentrations of greenhouse gases such as carbon dioxide and water vapor in the atmosphere. The climate changes because sunlight can pass through the atmosphere and warm up the planet, but the greenhouse gases hinder the escape of heat from the Earth to outer space. The increasing concentration of greenhouse gases in atmosphere primarily results from human

activities such as the burning of fossil fuels and deforestation.

Human activities destroy the balance between the amount of carbon emitted into the atmosphere and the amount of carbon absorbed by plants and other “sinks” on the surface of the Earth. Currently, combustion of fossil fuels releases about 7 billion tons of carbon to the atmosphere every year (Hotinski, 2007). Only about half of this excess carbon dioxide is absorbed by the ocean, plants, and trees, while the rest accumulates in the atmosphere. So human activities have influenced the ecological carbon cycle, causing more and more carbon goes from forest and fossil fuel to atmosphere. This enhances the natural greenhouse effect and leads to the increase of the global surface temperature.

The burning of fossil fuels usually occurs in automobiles, in factories, and in power plants that provide energy for people. It is a result of both the individual and collective human activities. Hence, slowing down the rate of carbon dioxide emission and global warming requires both individual and collective efforts. In order to make responsible and knowledgeable decisions about the urgent climate change, people need to understand the complex carbon cycling processes and to be environmentally literate. So the topic of “carbon cycling” has its unique scientific and practical importance. It is especially important for K-12 students to understand the carbon cycling when they are in schools so that they can become more environmentally responsible citizens in the future.

However, many studies showed that Americans are by and large uninformed or misinformed about environmental science (KACEE News, 2005). The Ninth Annual National Report Card shows that the environmental "illiteracy" remains widespread among American adults, though 95 percent of them endorse environmental education in the schools (NEETF & Roper, 2001). Less than half of the American public realizes that driving cars and using electrical

appliances in their homes contribute to global climate change. Among the general public, only 45% people can correctly identified emissions from autos, homes, and industries as the main cause of global climate change (NEETF & Roper, 2001). Only 12% Americans can pass a basic quiz on the awareness of energy topics (Coyle, 2005).

Both global climate change and the widespread environmental "illiteracy" among Americans make it imperative to improve the environmental science education. The *Environmental Literacy in American* report indicates that environmentally knowledgeable people are: 10% more likely to save energy in their homes; 50% more likely to recycle; 10% more likely to purchase environmentally safe products and 50% more likely to avoid using chemicals in yard care (Coyle, 2005). Environmental knowledge will make a difference in the decisions that people make about environmental issues. So science education needs to prepare students with the environmental knowledge so that they are more likely to make environmentally friendly decisions. The investigation of students' understanding of carbon cycling and how their understanding progresses over time is a necessary first step to find out ways to improve our current science education to help more students become environmentally literate.

1.2 Environmental literacy project and learning progression[1]-based assessments

The goal of the Environmental Literacy project at Michigan State University (MSU) is to improve students' environmental literacy by the time they are graduating from high school or college. Environmental literacy involves an understanding of the underlying environmental principles and applying these principles in everyday life to make informed decisions. The Environmental Literacy project has several research strands and carbon cycle is one of the

[1] Learning progressions have been referred to by many different names, including progress variables, learning trajectories, progressions of developmental competence, and profile strands.

strands. The carbon cycle strand aims to investigate students' learning progression in understanding the carbon cycle.

The carbon cycle is a key to understanding environmental systems. All living organisms are made of carbon compounds. Plants are the producers that generate organic carbon and harness light energy into chemical potential energy. All living organisms transform carbon compounds in order to grow and oxidize carbon compounds to obtain energy. In human systems, the combustion of organic carbon supplies energy to run vehicles, electrical appliances and etc. Thus, the key biogeochemical processes include (a) organic carbon generation (photosynthesis), (b) organic carbon transformation (biosynthesis, digestion, food webs, and carbon sequestration), and (c) organic carbon oxidization (cellular respiration, combustion). Because these processes are the means by which living organisms and human systems acquire energy and the means by which environmental systems regulate levels of atmospheric CO_2, these processes are used to describe the environmental systems. These carbon-transforming processes are essential for students to understand the environmental systems.

To explore students' understanding of these carbon-transforming processes and how their understanding progresses over time, the Environmental Literacy research team developed learning progressions to describe students' progress over time. The idea of a learning progression implies that "science learning is not simply a process of acquiring more knowledge and skills, but rather a process of progressing toward greater levels of competence as new knowledge is linked to existing knowledge, and as new understandings build on and replace earlier, naïve conceptions" (Wilson & Bertenthal, 2005, p.114). In the Environmental Literacy project, the learning progression is developed in an iterative process that moves back and forth between three major elements: a framework, assessments and scoring rubrics.

The framework describes learning of a specific concept over long periods of time. First, researchers develop an initial framework. Under the guidance of the framework, assessment tasks are designed to assess students' understanding of the concept. Then based on the framework and the patterns in the assessment data, scoring rubrics are developed to grade students' responses to the assessment tasks. Then researchers use the results to revise the framework. After the framework has been revised, they revise existing items and develop new items according to the revised framework. This is an iterative process, in which the results from the assessments lead to revisions in the framework and the other way around. Over the past five years, the Environmental Literacy project has gone through three iterative development cycles. Each iterative cycle represents an effort to strengthen the linkage and coherence among the elements. The assessments include both written assessments and clinical interviews.

The learning progression hypothesis suggests that there are general patterns in the development of students' knowledge and practice that are both conceptually coherent and empirically verifiable (Anderson, 2010). Through an iterative process of design-based research, moving back and forth between the development of frameworks and empirical studies of students' reasoning and learning, researchers can develop research-based frameworks, assessments and scoring rubrics that are both conceptually coherent and empirically verifiable. In the Environmental Literacy project, the researchers implicitly make claims such as:

1) The learning progression framework represents empirically-verified levels of students' achievement in developing accounts of carbon-transforming process.
2) The learning progression framework and the carbon cycle assessment are unidimensional. The learning progression levels, which represent increasing understanding and complexity, can be ordered along a continuum. Though the assessment includes items

assessing different processes, students use the same ability to answer these items.

3) The assessment can accurately locate individual students' understanding within the framework.

There are multiple threats to the validity of these claims. In a previous study (Mohan, Chen, Baek, Choi, Lee, & Anderson, 2009), the carbon cycle research team investigated the validity of the first and second claims. The study showed that the carbon cycle learning progression framework represented empirically-verified levels of students' achievement. Another conclusion was that students have a similar level of reasoning on different carbon transforming processes. While there were unique patterns for each item, the overall trend did not suggest major difference in reasoning among the process dimensions. So the assessment and the learning progression framework were essentially unidimensional. However, this conclusion was not based on a statistical analysis. Continually monitoring the patterns in the dimensions is needed when revising the assessments and the framework.

The focus of this dissertation is to investigate how to design learning progression-based science assessments to accurately classify students among the achievement levels. This is related to both the second and the third claims mentioned above. The dimensionality analysis can inform the assessment design. Whether students use the same ability or different abilities to answer the items has different implications for the assessment design. Meanwhile, designing assessments to accurately classify students among the learning progression achievement levels ensures the validity of the third claim above. The research focus of this dissertation is specified as three research questions listed in the following section.

1.3 Purpose of this study

The general purpose of this study is to investigate how to design learning progression-

based science assessments to accurately classify students' understanding into learning progression levels. There are three specific research questions:

1) How can tests be designed that use items in different formats (constructed response, ordered multiple-choice, multiple True or False) to accurately classify students' understanding into levels on the learning progression?
2) The carbon cycle learning progression framework includes students' understanding of different carbon transforming processes and different scientific practices. Whether students' understanding of these processes and practices is associated with a single construct or different constructs? If items of different process/practice measure different constructs, how should a test be designed to estimate students' proficiency for a particular process or practice?
3) What characteristics of test items support their use for assessing students' levels in a learning progression? How can these characteristics be used as design criteria for test items?

The first, second and third research questions are investigated in Chapter 5, 6 and 7 respectively. Both the first and second questions are about designing a test that is sufficient to accurately classify students' understanding into levels on the learning progression in science. More specifically, accurately classifying students means the test has small measurement error for students over a range of abilities. The psychometric information reflects the amount of measurement error of persons' ability. The larger the information, the smaller the measurement error of persons' ability (more details about information can be found in Section 3.2.4).

A test with high test information is desirable since it can measure students' ability more precisely. In practice, test information around 10 can be considered as good for detecting the

difference between individual students. Test information around 5 can be considered as sufficient to detect the difference between two groups of about 30 students each (see section 3.3 for a detailed discussion about why choosing information 10 or 5 as a rule of thumb). If the test information is 5, then the test can detect a certain difference between two groups of students at certain significance and power levels. For example, if each group has 30 students and assumes the observed variance of the ability estimates of each group is 0.5, the test can detect a difference of 0.58 (this is on the latent trait scale rather than raw score scale) between the mean ability estimates of the two groups at the significance level of .05 and the power level of 0.8.

The first research question investigates how to design a test composed of items in different formats to classify students' understanding into levels. The items developed by the Environmental Literacy project are in three formats: 1) Constructed Response (CR) items, 2) Ordered Multiple-Choice (OMC) (Briggs, Alonzo, Schwab, & Wilson, 2006) plus CR items and 3) Multiple True or False (MTF) plus CR items. Each item format is explained below.

The CR items require examinees to create their own responses rather than choosing a response from a set of options. There are two most common types of CR items: short-answer items and essay items. The CR items developed by the research group are short-answer items, each of which requires about five minutes for students to answer. Each CR question is scored according to a scoring rubric that gives varying degrees of credit according to the learning progression achievement levels.

The OMC +CR items are two-tier items that have two parts. The first part is an OMC question that requires students to choose a response from a list of options. A unique feature of the OMC items in comparison to the traditional multiple-choice (MC) items is that each option is linked to a particular developmental level of students' understanding of the target concept.

Students will get partial credit if they select a response that represents lower level understanding. The second part is a CR question that asks students to explain the choice they made in the OMC part.

The MTF + CR items are also two-tier items that have two parts. In the MTF part, a set of true or false questions ask students to judge what are the matter or energy source(s) for events such as tree growth or human growth from a list of options. Based on their responses, we pinpoint their achievement levels. The CR part then asks students to explain the choices they made in the MTF part.

Each item format has its advantages and disadvantages. OMC and MTF are effective item formats, which require relatively shorter administration time and less scoring effort than CR items. However, guessing can be involved when students answer items in these formats, especially for students at low ability levels who are more likely to guess. So the items may not measure low-level students precisely. CR items require longer administration time and more scoring effort, but they are more appropriate for measuring students' high order thinking. An important question is how to design a test composed of items in different formats to utilize the advantages of each format and measure the target construct is the first question.

The second research question investigates whether students' understandings of different processes/practices are associated with a single latent construct or several constructs. The extent to which students' understandings of these processes/practices are related to each other will be investigated. The results have implications for the test design and item selection process. For example, if students' understandings of different processes/practices are very distinct from each other, then only items of each particular process/practice should be selected to assess students' understandings of that process/practice. The research reported here investigated whether the

processes/practices define an unique ability scale in the item response dimensional space and whether the different processes/practices are highly intercorrelated.

The third research question is what characteristics of test items support their use for assessing students' levels in a learning progression and how can these characteristics be used as criteria for designing test items. There are many guidelines and tips about how to write good items that can be found in literature (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Thorndike, 2005). The research purpose of this dissertation is to relate item characteristics to information provided by the item about students' learning progression levels. Learning progression-based items not only assess whether or not students master the concept, but also assess the trajectories of students' learning of the measured concept and where are they along the trajectory. Results from the item characteristics analysis can tell us what kind of OMC/MTF options can accurately differentiate students among levels and what kind of CR item stem can elicit detailed responses from students at different ability levels. These results can provide guidelines and clear targets to item writers to write good learning progression-based items in addition to the test specifications.

Chapter 2 Literature Review

The research questions proposed in the first chapter are centered on how to design learning progression-based science assessments. Before investigating this, we need to have some fundamental understanding of the assessment and the learning progression. The first section (2.1) of this chapter starts by reviewing assessments from the theoretical perspective and illustrates the assessment triangle that underlies every assessment. Then, each assessment triangle vertex is explained in the following three sections. Section 2.2 reviews a model of cognition and learning. It introduces the learning progression idea and carbon cycle learning progression framework that the assessment is based on. Section 2.3 summarizes literatures about the advantages and disadvantages of items in different formats to provide guidelines for evaluating the effectiveness of items in multiple formats. Section 2.4 reviews the psychometric models that are used for the data analysis.

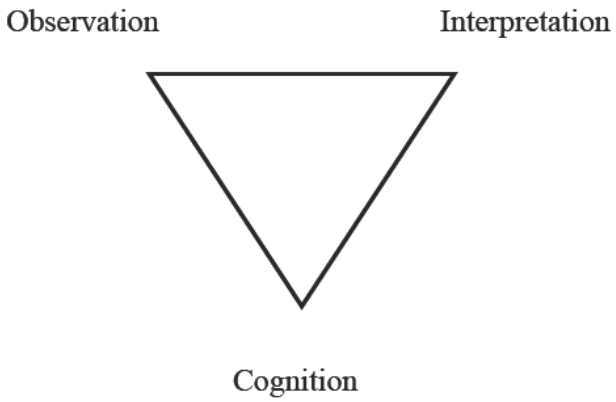
2.1 Assessment triangle

Assessments are designed based on a coherent argument to suit the assessment's purpose (Mislevy, Steinberg & Almond, 2003). It is a process of reasoning that starts from the observation of students' performances on the assessment tasks to inferences about their knowledge or skills of the measured concept. Mislevy and his colleagues pointed out that good assessment tasks cannot be developed in isolation. Development must start from the intended inferences, then the observations and performances that are needed to support those inferences, the assessment tasks that will elicit these performances, and reasoning to connect each component to make coherent arguments for assessment design (Almond, Steinberg & Mislevy, 2002; Mislevy, Steinberg & Almond, 2003).

The National Research Council (NRC) portrayed assessment as a triangle with three

corners—cognition, observation, and interpretation (Pellegrino, Chudowsky, & Glaser, 2001). This triangle underlies all assessment and it presents assessment as an integration of the three fundamental elements (See Figure 2.1 below). Pellegrino, et al, claimed in the NRC report: "These three elements—cognition, observation, and interpretation—must be explicitly connected and designed as a coordinated whole. If not, the meaningfulness of inferences drawn from the assessment will be compromised" (p. 2).

Figure 2.1 The assessment triangle



The *Cognition* vertex plays the central role in assessment design. It refers to the theories and beliefs about what people know, how they know it, what are the knowledge and skills that are important to measure. In other words, it is not only which knowledge and skills are to be assessed but also how the knowledge and skills develop. The design of assessment needs to be consistent with best available understanding of how students learn. In measurement terminology, the targeted knowledge and skill to be assessed is referred as the "construct". An assessment should start from an explicit and clearly conceptualized cognitive model of learning with a well-defined construct that is considered most important to assess.

The *Observation* vertex refers to a set of assessment tasks that are used to elicit responses

from examinees. They are based on theories and beliefs of the kinds of tasks that will prompt students to provide valid and rich responses. The assessment tasks developed for observation need to serve the purpose of the assessment. They must be carefully designed to elicit the knowledge and cognitive processes that the model of learning suggests are most important for competence in the domain. Meanwhile, observations need to support the inferences and decisions that will be made based on the assessment results.

The *Interpretation* vertex includes all the methods and tools that are used to reason from observations to inferences of students' learning. Interpretation methods or tools such as statistical models or qualitative models are used to characterize and summarize the patterns of the data collected through assessment tasks. The interpretation model needs to fit the model of cognition and learning to characterize the knowledge and skills that cognitive theories suggest are important to pursue. Meanwhile, the interpretation model depends on the type of data collected through observation. Through interpretation, the observations of students' performances are synthesized into inferences of their knowledge, skills and other attributes being assessed.

The three vertices of the assessment triangle need to be connected with each other to lead to an effective assessment and sound inferences (Pellegrino, 2009). Assessment developers should not only focus on the observation corner, but also pay explicit attention to all three elements of the assessment triangle (cognition, observation, and interpretation) and their coordination. They need to use the assessment triangle as a foundation to develop systematic approaches to design assessments. These systematic approaches will be different from the common approaches that merely focus on the development of "good items" in isolation from all other important facets of design (NRC, 2006).

Similar to the assessment triangle, the Berkeley Evaluation and Assessment Research

(BEAR) center developed four building blocks for constructing quality assessments that map to the NRC Assessment triangle: construct maps, items design, the outcome space, and the measurement model (Kennedy, 2005; Wilson, 2005). The difference between BEAR's building blocks and the assessment triangle is that the building blocks emphasize that assessments need to be based on a developmental perspective on student learning. It also emphasizes the alignment between what is taught and what is assessed in constructing assessments. The BEAR center supported the development of the carbon cycle assessment following these building blocks.

The design of high-quality assessment is a complex process. The analyses proposed in this study investigate the connections between the cognition, the observation and the interpretation vertices.

2.2 Cognition—Learning progression of carbon cycle

2.2.1. Learning progression and progress variables

The NRC emphasizes the central role of a model of cognition and learning in assessment design (NRC, 2001). Good science assessments need to be based on a modern understanding of students' science learning. Contemporary theories of learning emphasize that learning is a process of constructing understanding that involves ongoing revision and reorganization of current thinking as new knowledge is acquired (NRC, 2007). This suggests one should take a developmental approach for science assessment design.

A developmental approach for assessment is the process of monitoring students' progress in a domain of learning over time. This will help to find out the best ways to facilitate their further learning. A developmental approach involves knowing what students know now, and what they need to know in order to progress. The developmental approach can be applied to develop large-scale assessments at national and state levels, as well as classroom assessments.

This approach uses a learning progression or some other continuum to design assessments to monitor students' progress over time.

Learning progressions 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 (e.g., six to eight years)" (Duschl, Schweingruber, & Shouse, 2007, chapter 8). Learning progressions are anchored on one end by what we know about the reasoning of students about specific concepts entering school (i.e., lower anchors). On the other end, learning progressions are anchored by societal expectations (e.g., science standards) about what we want high school students to understand about science when they graduate (i.e., upper anchors). Learning progressions describe the intermediate understandings between these anchor points (Mohan, Chen & Anderson, 2009).

Progress variables are used to track students' increasingly sophisticated understanding of a given concept. Progress variables mediate between big ideas and specific concepts and skills being learned during instruction (Wilson, 2005). Progress variables are aspects of knowledge and practice that are present at the achievement levels of the learning progression. The development of progress variables can be traced across levels. The difference between learning progressions and progress variable is summarized in Merritt and Krajcik's paper (2009): "Learning progressions are a means for determining how to support student learning of the big ideas of science. They are big picture, research-based and provide opportunities to think about how to engage students in long-term learning of both content and practice skills. Progress variables serve as a means for tracking students' progress during instruction. Just like learning progressions, they are also research-based. Moreover, the development of these progress

variables is important because they can form the basis for tracking students' understanding of the particle nature of matter. " (p.2)

The National Research Council recommended using learning progressions to inform science assessment (NRC, 2006, 2007). However, learning progression-grounded items pose development challenges. It is difficult to write items that provide opportunities for students to respond at multiple levels of a learning progression (Anderson, Alonzo, Smith, & Wilson, 2007). CR items must be carefully designed to elicit complete responses while not telling students what should be included. Multiple-choice items must be written in such a way that the highest level is not indicated by the use of "science-y" terminology not present in lower level options (Alonzo & Steedle, 2008). So this study is aiming to investigate the development of learning progression-grounded items.

2.2.2. Carbon cycle learning progression

The carbon cycle assessment is developed based on the achievement levels the Environmental Literacy project identified in previous studies (Mohen, Chen & Anderson, 2009; Jin & Anderson, 2010). Over the past several years, the Environmental Literacy project research team has taken an iterative approach that moves back and forth between the development of the learning progression framework and the development of assessments. The learning progression framework was refined based on the empirical data gathered from assessments. Then the assessments were modified according to the refined framework. The current assessments were developed based on four achievement levels:

Level 1: Simple force-dynamic accounts.

Level 2: Elaborated force-dynamic accounts (e.g., different functions for different organs)

Level 3: Attempts to trace matter and energy, but with errors (e.g., matter-energy confusion,

failure to fully account for mass of gases).

Level 4: Correct qualitative tracing of matter and energy through processes at multiple scales (e.g. macroscopic scale, microscopic scale and large scale).

In the following paragraphs, I will provide more detailed descriptions of the learning progression levels and provide example responses from one item (item label ENERPEOP) to illustrate each level. This item asks students "what are the energy sources for people to live and grow?" Students are required to choose Yes or No for a list of five things: water, food, nutrients, sunlight, oxygen and then explain their answers.

At the lowest level--Level 1--students describe the world in terms of objects and events rather than chemically-connected processes. Their understandings are confined to the macroscopic scale without recognizing the underlying chemical changes or energy transformations of events. Students describe macroscopic processes in terms of the action-result chain. They think the actors use enablers to accomplish their goals and the interactions between actors and enablers do not involve any change of matter/energy. A typical level 1 response from the item is: "People need water when they were exercising, so we can feel energized. You need food so you won't feel weak during the day. You need nutrients to help keep your body healthy and strong. When you exercise, it also helps your body stay strong. You need sunlight to feel refreshed."

At Level 2, students continue to attribute events to the purposes and natural tendencies of actors, but they also recognize that macroscopic changes result from "internal" or "barely visible" parts and mechanisms that involving changes of materials and energy in general. A typical level 2 response to the item is "we drink water and eat food. Food has nutrients and vitamins that are converted into energy and pumped to your muscles. We do exercise to burn fat.

Sunlight does not give us energy, plant use it but not us." Student began to pay attention to components of food such as "nutrients" and "vitamins".

At Level 3, students can reason about macroscopic or large-scale phenomena but because of limited understanding at the atomic-molecular scale, they cannot trace matter and energy separately and consistently through those phenomena. Level 3 students link macroscopic changes to chemical changes and describe chemical changes as changes involving atoms, organic molecules, and energy forms, but do not successfully conserve matter and energy, for example, they might think organic molecules convert into energy. A level 3 response to the item is "All living organisms need water; food contains glucose needed to make ATP for energy. Nutrients are a food source. Exercise controls the size of lipids. Sunlight is the energy source of all life."

Only at the highest level--Level 4--students can use atomic-molecular models to trace matter/energy systematically through multiple processes connecting multiple scales. They use constrained principles (conservation of atoms and mass, energy conservation and degradation), codified representations (e.g. chemical equations, flow diagrams) to explain chemical changes. A level 4 response is "Humans break down carbohydrates as well as fats. The body gets these from food. H_2O, nutrients, exercise, CO_2, and O_2 are all necessary for life but are not energy sources. Humans don't undergo photosynthesis so sunlight isn't an energy source". See Table A.1 in the Appendix for detailed descriptions of these four achievement levels.

This learning trajectory is a typical path followed by American students (Anderson, 2010). It gives clues about the types of assessment tasks that will elicit evidence to support inferences about student achievement at different points along the progression. By considering the ways in which students learn science, the science assessments and tasks can be created to gather information on how well and to what degree students are progressing over time toward

more expert understanding.

2.2.3. Carbon cycle progress variables: process and practice

There are two identified progress variables that present at all the achievement levels of the carbon cycle learning progression: process and practice. The Environmental Literacy project identified the process and practice progress variables because these are used to mediate between big ideas (carbon cycle) and specific concepts and skills being learned in classrooms. They are used to track students' increasingly sophisticated understanding of the carbon cycle.

Understanding carbon cycling in socio-ecological systems is challenging for most students. Many studies showed that students did not fully understand carbon cycling in socio-ecological systems. Kempton, Boster, and Hartley (1995) found that many students confused global warming with ozone depletion. Research found that it took time for students to understand the mechanism of global warming over the course of secondary education (Boyes & Stanisstreet, 1993). Some other studies (e.g., Anderson, Sheldon, & Dubay, 1990; Songer & Mintzes, 1994; Fisher, et al., 1984) documented a wide range of students' difficulties to understand carbon-transforming processes such as photosynthesis and cellular respiration.

To explain the carbon cycle in complex coupled human and natural system, students need to see the key processes that tie systems together and perform certain practices (e.g. tracing energy, tracing matter) to explain those processes. Thus, the Environmental Literacy project identified the key carbon-transforming processes and practices that can be used as conceptual tools for students to reason about carbon cycling in complex systems. There are six key carbon-transforming processes and five scientific practices identified. The six carbon-transforming processes are described below:

- Photosynthesis — the chemical process that plants convert carbon dioxide into organic compounds using the energy from sunlight. The items about plant growth assess students' understanding of photosynthesis.
- Biosynthesis/digestion — organic carbon is transformed during these processes. Biosynthesis a cellular process by which substrates are converted to more complex products under the catalysis of enzymes. Digestion is the mechanical and chemical breakdown of food into smaller components that are more easily absorbed into a blood stream. The items about animal growth assess students' understanding of biosynthesis and/or digestion.
- Cellular respiration — organic carbon is oxidized during cellular respiration (including decomposition) and combustion. Cellular respiration is a set of the metabolic reactions and processes by which the chemical energy of organic molecules (e.g. glucose, carbohydrates, fats, proteins) is released and partially captured in the form of ATP. The animal function items address students' understanding of cellular respiration.
- Decomposition — the process by which organic material is broken down into simpler forms of matter. Decomposition is one type of cellular respiration. The items such as APPLEROT and TREEDECAY (APPLEROT and TREEDECAY are item labels. APPLEROT stands for "apple rot" and TREEDECAY stands for "tree decay") measure students' understanding of decomposition.
- Combustion — the burning of a fuel and oxidant to produce energy and release carbon dioxide. The items about burning fossil fuels, burning candles or matches assess students' understanding of combustion.

- Cross-process events— a set of related carbon transforming processes. Students' understanding of cross-process events is measured by items that require students to connect their understanding of different carbon transforming processes to reason about a phenomena such as global warming.

The five practices identified are specified below:

- Macroscopic practice—students' general account of material kinds and what is happening to materials and forms of energy (or actors, actions, and enablers) at the macroscopic scale. Items that assess macroscopic practice are questions such as "Where does the object come from?", "Where does it go?" and "How does it change?".
- Mass/Gases/Amount practice—quantitatively accounting changes in mass (or size/amount) of materials. Items that assess this practice are questions such as "Does air have weight/mass?" and "Does the stuff contribute to weight gain/loss?".
- Energy/Causes practice—accounting specifically for energy or closely related terms (power, light, heat). Items of this practice are "What are the things that cause changes?", "Where does the energy come from?" and "Where does the energy go?".
- Microscopic practice—using structures and functions of subsystems to account for macroscopic observations. Questions such as the following address this practice: "What are the smaller/invisible parts?", "Are there invisible changes behind the macroscopic phenomena?" and "How are they related to macroscopic phenomena?".
- Large-scale practice—association and tracing among macroscopic processes (using structure and function of large-scale systems). This practice focuses on questions such as "How are changes/events similar or different?" and "How are changes/events connected?".

Though students have learned some fundamental principles to trace matter or trace energy in their science classes, they seldom apply them to environmental issues. Numerous studies have found that students intuitively focus on visible aspects of systems and do not use atomic-molecular accounts to explain macroscopic or large-scale events (Hmelo-Silver, Marathe, & Liu, 2007; Lin & Hu, 2003). A study conducted in the Environmental Literacy project indicated that pre-service science teachers did not trace matter and energy separately to explain chemical changes. Instead, they thought fat was "burned up" or "used for energy" when people lost weight (Wilson, Anderson, Heidemann, Merrill, Merritt, Richmond, Sibley & Parker, 2006). The identified key practices can help students to reason about the carbon-transforming processes.

The assessment items are developed to assess these practices and processes. Each item assesses students' understanding of a single process or their understanding of cross-process events (e.g. global warming). Each item also addresses one scientific practice such as tracing energy. The assessment consists of items focusing on six processes (plant growth, animal growth, animal functioning, decomposition, combustion, cross-process) and five practices (macro, mass, energy, micro, large-scale). One goal of the assessment is to measure students' ability of each practice/process that the assessment is designed to measure. So the "cognition" vertex of the assessment are levels of achievement with respect to the processes/practices and the "observation" tasks need to be designed to measure these processes/practices precisely.

2.3 Observation—Items in different formats

When designing a test, the selected item format(s) should be useful for eliciting evidence of students' understanding of the measured construct. The three item formats used in this assessment, two-tier items consisting of OMC plus CR parts, two-tier items consisting of MTF plus CR and CR only formats, are used to tap into students' learning progression of carbon

cycling. The groups of items in these formats are assembled so the scores that they give can shed light on the full range of the science content knowledge, understandings, and skills included in the construct as elaborated by the related learning performances. Knowing the advantages and disadvantages of each item format will help us to select the appropriate format(s) to achieve the goal of our assessment.

It is widely recognized that the MC item format is an effective means for determining how well students have acquired basic content knowledge. However, the limitations of MC items are also well recognized such as the guessing effect and not showing students' original thoughts. Researchers pointed out that MC items might not be able to measure high order thinking (Delandshere & Petrosky, 1998; Kennedy, 1999; Lane, 2004) and might encourage teachers to drill students on isolated facts and formulas (Frederiksen, 1984; Shepard, 2000). However, some well-designed MC items can be used to measure complex cognitive processes. For example, the Force Concept Inventory (Hestenes, Wells & Swackhamer, 1992) was an assessment that used MC items but tapped higher-level cognitive processes.

MTF items are similar to MC items. The difference is, rather than selecting one best answer from several alternatives, students respond to each of the several alternatives as separate True or False questions. This item format is especially good for assessing students' commitments to fundamental principles. Since MTF items allow students to select multiple answers, to answer the item correctly, students need to not only identify all the correct answer(s) but also exclude all the incorrect answer(s). This requires students to have deep understanding of the principles being assessed and apply those principles consistently. For example, students need to identify sunlight as the energy source for tree growth, and recognize though trees need nutrients, water and air to grow, these are not energy sources for trees.

Some articles addressed the advantages and disadvantages of the MTF format. Frisbie (1992) gave a comprehensive review of the literature and synthesized the following merits of MTF items: "(a) They are a highly efficient format for gathering achievement data, (b) they tend to yield more reliable scores than MC and other objective formats, (c) they measure the same skills and abilities as content-parallel MC, (d) they are a bit harder than MC for examinees, and (e) they are perceived by examinees as harder but more efficient than MC" (p. 25). There are also some shortcomings with MTF format. Usually, answering MTF item involves a lot of guessing, especially for examinees with the least knowledge who will guess most. MTF may not be reliable at the low ability range. Grosse and Wright (1985) found that the examinees' response style (guess "T" more often or guess "F" more often) would determine whether the true score or the false score was more reliable. Dunham (2007) found that students' responses to the MTF item were influenced by an "optimal number correct" response set. For example, examinees tended to endorse three or four of the six MTF options more frequently than would be expected by chance alone. These results suggest that MTF item can be used as an alternative to MC items but when designing and analyzing MTF items, attention needs to be paid to the reliability of the items, the guessing involved in the responses, and the response style factor.

The major advantage of CR items is that they are more appropriate for measuring students' abilities to organize, integrate and synthesize their knowledge and their abilities to solve novel problems. CR items can be used to demonstrate students' original thoughts and they allow students to show the process of their reasoning. Hence, CR items can serve as useful assessment tool for teachers (McNeill & Krajcik, 2007; Champagne, Kouba & Gentiluomo, 2008). CR items also have disadvantages such as the difficulty in administrating, scoring, inconsistencies among raters, and not always showing students' thinking.

Some previous studies were conducted to compare the use of OMC items with the use of other types of items when assessing the same concepts. Briggs, Alonzo, Schwab, and Wilson (2006) used OMC items to assess students' levels on a learning progression of the earth and solar system. The results indicated that test scores based on OMC items compared favorably with scores based on traditional MC items in terms of their reliability. There was a weak to moderate positive correlation between students' scores on OMC items and their scores on comparable tests consisting of traditional MC items. Alonzo & Steedle found that compared to CR items, OMC items "appear to provide more precise diagnoses of students' learning progression levels and to be more valid, eliciting students' conceptions more similarly to cognitive interviews compared to open-ended items" (Alonzo & Steedle, 2008, p.1).

Other researchers found inconsistency in students' responses to items in different formats but addressing the same underlying principles. Steedle (2006) found that students performed differently on MC and short-answer items targeting the understanding of the same underlying progression. Lee, Liu & Linn (2011) found that compared to MC items, CR items discriminated between high and low knowledge integration ability students much more effectively, measured a wider range of knowledge integration levels, and were more sensitive to knowledge integration instruction.

Many studies that compared MC and CR items across a range of outcomes suggested that these items might measure a different aspect of the construct, especially at the extremes of the distribution (Lee, Liu & Linn, 2011). Ercikan, Schwarz, Julian, Burket, Weber, and Link (1998) used an IRT model to calibrate both item types on a single scale and discovered that, when combined to produce a single scale, the overall measurement accuracy improved because the CR items could tap very-low and very-high ability groups. Wilson and Wang (1995) reported that

"performance-based items provided more information than multiple-choice items and also provided greater precision for higher levels of the latent variable" (p. 51).

The different results from these studies point to the need to better understand the affordances of different item types for assessing students' learning progression levels. Meanwhile, the analysis of the two-tier items consisting of OMC and CR parts or consisting of MTF and CR parts helps to explore new forms of assessment to be used in classroom and large-scale contexts. This study examines how effective are items in these formats in terms of differentiating students among levels. Then the study investigates ways to make good use of items in these formats to form a test that can both effectively and accurately diagnose students' learning progression levels.

2.4 Interpretation— The psychometric models

In large-scale assessment programs, the measurement model used for the data analysis will be either based on classical test theory (CTT) (Novick, 1966; Lord & Novick 1968; Allen & Yen, 2002) or item response theory (IRT) (brief introduction of CTT and IRT can be found in Chapter 3). All models are incorrect to some extent. They are an oversimplification of reality. However, the model does not have to be assumed to be absolutely correct to be useful. The decision about which measurement model should be used is generally based on the inferences one wants to support with test results. The idea of learning progressions is that students develop successively more sophisticated ways of thinking about a topic over time. So their abilities are assumed to lie on a latent continuum, a scale along which individuals can be ordered. Applying IRT models to estimate students' proficiency along the latent continuum is appropriate.

A variety of measurement models are available depending on the type of item and the assumptions that are made. The measurement model used to interpret the data can be evaluated

by two criteria. First, it should be tightly connected with the "cognition" part to formalize the relationships posited in the model of cognition and learning. A measurement model grounded in substantive theory is more likely to lead to solid inferences (Rao & Sinharay, 2007). Second, the measurement model needs to fit the data adequately. The fit of a measurement model is evaluated in terms of the extent to which observed data deviate from predictions of the model.

In this study, models will be selected based on their fit with both the substantive theory and the empirical data. The measurement models used in this study are unidimensional and multidimensional partial credit models. These models are reviewed in Chapter 3 and compared in terms of assumptions and properties. Classical test statistics and IRT models are used to analyze the quality of the assessment. The dimension analysis is used to find out the latent constructs that the assessment assesses. After knowing more about the latent constructs and how precise different composites of ability are being assessed, this study explores ways to design the test that can give good information about students' progress through the learning progression of the construct being measured.

Thus this study gives specific consideration to all three components of the assessment triangle and integrates the three components as a whole. For the carbon cycle assessment, the cognition to be measured is students' levels of performance on the progress variables: process and practice. The observations are the items in multiple formats. These need to be selected based on what would constitute evidence of student competencies and what are the effective ways to collect evidences. The interpretation should fit with the data and be supported by the theoretical underpinning.

Chapter 3 Psychometric Theories and Related Terminologies

The analyses in this study are mainly based on the IRT. Classical test statistics are also involved to evaluate the assessment. In this chapter, a brief introduction of the general framework of IRT and classical test theory is included. The relevant IRT models are reviewed and relevant terminologies are explained. These can provide some background knowledge for people who are not familiar with the measurement theories and terminologies to make sense of the data analyses. In addition, this chapter also includes a discussion of the relationships between test information and hypothesis testing and the rational to set the amount of desired information at 5 as a rule of thumb.

3.1 Classical test theory

Classical test theory (CTT) (Novick, 1966; Lord & Novick 1968; Allen & Yen, 2002) can be regarded as roughly synonymous with true score theory. CTT is based on the understanding that a given test score can be thought of as consisting of two parts. One is the error of measurement and the other is the actual individual score on the studied attribute, which is of interest. This latter part is called true score.

Up to the 1980's, interpreting test scores was largely based on the CTT. In CTT, an observed score (X) is equal to the true score (T) plus error (E).

$$X = T + E \tag{3.1}$$

CTT assumes that each person has a true score, T, which is not directly observable. The true score will be equal to the observed score if there are no errors in measurement. A person's true score is defined as the expected number-correct score over an infinite number of

independent administrations of the test under the exact same condition. The observed score, X, is equal to the true score plus measurement error, which is not correlated with the true score by the definition of CTT.

The most important concept in CTT is reliability. It describes the relations between the observed score, the true score and the measurement error. Reliability is defined as the proportion of variance of observed score that is attributable to the true score rather than to the error:

$$\rho_X = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \tag{3.2}$$

That is, reliability (ρ_X) is the complement to 1 of the ratio of error variance to observed variance.

The item discrimination index based on the CTT is analyzed in this study. It is the correlation between the students' scores of the item and their total scores. It represents the discrimination ability of an item and is an indicator of the item quality. The higher the average item discrimination ability, the higher the reliability of the test (Ebel, 1979). Ebel proposed ranges for evaluating discrimination indices for items on classroom tests. The ranges are summarized below:

- ≥ 0.40 Very good
- 0.30 ~ 0.39 Reasonably good, consider improving
- 0.20 ~ 0.29 Marginal, needs improvement
- ≤0.19 Poor, reject or revise

3.2 Item response theory

Compared with the CTT, IRT (Lord, 1980; Rasch, 1960; Rasch, 1980; Lazarsfeld & Henry, 1968) is considered as a modern psychometric theory. The idea of IRT is to model the

relationship between person proficiency level (the continuum) and the probability of correct response to an item. The assumptions of IRT are stronger than CTT, which assume that the probability of observed responses is determined by examinee's ability and the parameters that characterize the items. Though there could be infinite number of possible IRT models used to estimate item parameters and person proficiencies. Only a few of them are the most applied IRT models.

3.2.1 Unidimensional IRT

Unidimensional IRT (UIRT) assumes a single underlying trait or a common composite of traits explains persons' performance on the test items. The simplest commonly used UIRT model is the one-parameter logistic model. It is used for dichotomous scored items. It has one parameter for describing the ability of the person and one parameter for describing the difficulty of the item. The equation for the model is given by

$$P(X_{ni} = 1) = \frac{e^{(\theta_n - \delta_i)}}{1 + e^{(\theta_n - \delta_i)}} \tag{3.3}$$

where X_{ni} is the response of n^{th} student to the i^{th} item.

θ_n is the latent trait (the ability parameter) of the n^{th} student and

δ_i is the item difficulty of the i^{th} item.

Note that the only observable quantity is the X_{ni} and student ability parameter as well as the item difficulty parameter will be estimated by maximizing the total likelihood (see section

3.3.3 for more details). That is, as long as we determine the model and collect the data (X_{ni}), the rest will be a pure mathematical process to get the person ability parameter and the item difficulty parameter estimated.

The one-parameter logistic model is for items that are dichotomously scored. For the items that are polytomously scored, there are other UIRT models. One of the most commonly used models for polytomous item is the partial credit model (PCM) (Masters 1982). PCM was designed for test items with two or more ordered categories and is appropriate for test items in which the scores on the item represent levels of performance, with each higher score meaning that the examinee accomplishes more of the desired task. The boundaries between adjacent scores are labeled "thresholds". Threshold parameters are the locations on the ability scale that students with those abilities will have the same probability of getting the two adjacent score points. For example, the assessment items in this study are all polytomous item and usually have 4 score points: 1, 2, 3, and 4. Then, there are three thresholds, the threshold between level 1 and 2 (d_1), between level 2 and 3 (d_2), and between level 3 and 4 (d_3). If a student's ability is d_1, then he/she will have 50% of the chance to get a Level 1 or a Level 2 score. If his/her ability increases, the probability of getting a level 2 will be higher than the probability of getting a level 1.

The OMC, MTF questions and the CR items have two or more score categories, and higher score requires accomplishing more of the desired task. Therefore, it is appropriate to use PCM to model our data. The probability of student *n* being graded into level *x* for item *i* is given by

$$P(X_{ni} = x) = \frac{\exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\theta_n - \delta_{ij})}$$

(3.4)

where X_{ni} is the observed score of person n on item i,

m_i is the maximum score on item i,

θ_n is the proficiency of person n,

and δ_{ij} is the threshold parameter for the j^{th} score category for item i.

The unidimensional PCM can be applied to estimate person proficiency and item parameters.

3.2.2 Multidimensional IRT (MIRT)

The actual interactions between person and test items are often more complicated than what the UIRT implied. MIRT is an extension of the UIRT model used to describe situations where multiple skills and abilities are needed to respond to the test items. MIRT describes students' abilities in a multidimensional space with each construct as a line in the space. MIRT identify a mathematical model that can represent the connection between the probability of a response to an item and the location of a person in a multidimensional space (Reckase, 2009).

The potential use of MIRT for educational assessment has been recognized for more than twenty years (e.g. Embretson, 1984; Reckase, 1990). It is a useful methodology for assessing competencies in educational assessment and provides a more accurate representation of the complexity of tests. MIRT models provide tools for gaining more detailed information than the information gained from more traditional, classical measurement models. Applying MIRT models can help understand the latent traits that an item measures, for example, apply MIRT models to fit the data can tell how many latent traits influence performance on an item. The examinee's proficiency levels on each latent trait are estimated from the MIRT models and the measurement error estimated tells how precisely different composites of ability are being measured (Ackerman, 1994a, 1994b, 1996; Muraki & Carlson, 1995; Reckase, 1985, 1997; Reckase & McKinley, 1991; Yao & Schwarz, 2006). MIRT models are useful for understanding both the items and the students' abilities in complex domains. So MIRT analysis can provide more detailed information about the items and the test, which can inform the instrument construction.

It is appropriate to apply MIRT model to analyze the assessment data since the assessment has items in multiple formats and the items assess students' understanding of different processes. Researchers found that several assessments that contain a mixture of MC and CR items measure more than one trait (Yao & Boughton, 2009). In addition, the assessment measures students' understanding of different carbon transforming processes and different scientific practices, which may be intrinsically multidimensional.

In addition, science assessments often require numerous knowledge and skills, for instance, knowledge of different subject matter areas and a variety of skills, such as conceptual understanding and scientific investigation. So science assessments are likely to be

multidimensional. When responding to a particular item, students often rely on more than a single ability. The multidimensionality that underlies science assessments has been recognized by researchers (Reckase & Martineau, 2004; Wei, 2008). Therefore, more complex multidimensional models can be applied to describe the data.

The multidimensional model, called multidimensional random coefficients multinomial logit model (MRCMLM), is used in this study. It is specified in Adams, Wilson, & Wang's paper (1997). It assumes that a set of traits underlie the persons' responses. It is a general model that includes both dichotomously and polytomously scored test items. The expression for the full model is given by

$$P(X_{ik} = 1 \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\xi}, \boldsymbol{\theta}) = \frac{e^{\mathbf{b_{ik}}\boldsymbol{\theta} + \mathbf{a_{ik}}'\boldsymbol{\xi}}}{\sum_{k=0}^{K_i} e^{\mathbf{b_{ik}}\boldsymbol{\theta} + \mathbf{a_{ik}}'\boldsymbol{\xi}}} \tag{3.5}$$

where $\mathbf{A}$ is a design matrix with vector elements $\mathbf{a_{ik}}$ that select the appropriate item parameter for scoring the item;

$\mathbf{B}$ is a scoring matrix with vector elements $\mathbf{b_{ik}}$ that indicate the dimension or dimensions that are required to obtain the score of *k* on the item;

$\boldsymbol{\xi}$ is a vector of item difficulty parameters;

$\boldsymbol{\theta}$ is a vector of coordinates for locating a person in the construct space.

K_i is the highest score category of the item; k represents the score category.

X_{ki} is an indicator variable that indicates whether or not the observed response is equal

to k on Item i. If the score is k, the indicator variable is assigned a 1; otherwise, it is 0.

Suppose an item has four response categories (0, 1, 2, 3) and three latent abilities are required to solve the item. The MRCMLM model is specified using the design matrix and the scoring matrix as the following:

$$\mathbf{A}=\begin{bmatrix}0 & 0 & 0\\ 1 & 0 & 0\\ 1 & 1 & 0\\ 1 & 1 & 1\end{bmatrix}$$

$$\mathbf{B}=\begin{bmatrix}0 & 0 & 0\\ 1 & 0 & 0\\ 1 & 1 & 0\\ 1 & 1 & 1\end{bmatrix}$$

A is the design matrix. The elements of **A** matrix select the appropriate item parameter to score the item. The rows of the matrix correspond to the scoring categories and the columns associate with the item parameters. The elements of the **A** matrix are specified by the test developer rather than obtained through the statistical estimation procedures. **B** is the scoring matrix. The rows of the **B** matrix represent the scoring categories and the columns of the matrix correspond to latent dimensions. The elements of **B** matrix indicate the dimension or dimensions that are required to obtain the score of *k* on the item. For example, the abilities in all three dimensions are required to obtain score 3, so the fourth row of the **B** matrix is [1,1,1]. In this case, the MRCMLM is specified as a multidimensional PCM and different achievement levels require different latent abilities. Multidimensional PCM is used in this study.

The response categories are modeled as

$$P(\mathbf{X}_{10}=1;\ \mathbf{A},\ \mathbf{B},\ \xi \mid \theta) = 1/D$$

$$P(\mathbf{X}_{11}=1;\ \mathbf{A, B,}\ \xi \mid \theta) = \exp(\theta_1 + \xi_1)/D$$

$$P(\mathbf{X}_{12}=1;\ \mathbf{A, B,}\ \xi \mid \theta) = \exp(\theta_1 + \theta_2 + \xi_1 + \xi_2)/D$$

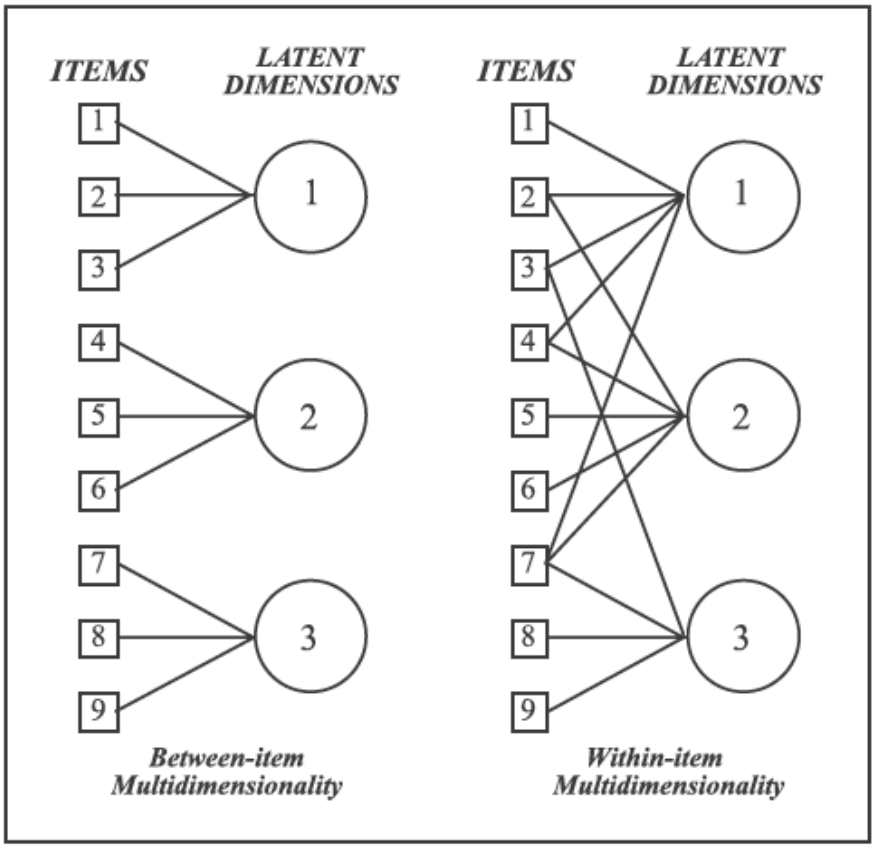
$P(\mathbf{X}_{13}=1; \mathbf{A}, \mathbf{B}, \xi \mid \theta) = \exp(\theta_1 + \theta_2 + \theta_3 + \xi_1 + \xi_2 + \xi_3) / D$

where $D = 1 + \exp(\theta_1 + \xi_1) + \exp(\theta_1 + \theta_2 + \xi_1 + \xi_2) \exp(\theta_1 + \theta_2 + \theta_3 + \xi_1 + \xi_2 + \xi_3)$

Adams et al. (1997) specified two subclasses of the MRCMLM model. One subclass is for a between-item multidimensionality test, a test that consists of several unidimensional subscales. Each item of the test relates to only one latent dimension. The other subclass of the model is used for within-item multidimensional tests, the case in which each item of the test relates to more than one latent dimension. In this study, both within and between-item multidimensional PCM are applied. These two types of multidimensionality can be modeled by having appropriate design and scoring matrices in the MRCMLM model. Figure 3.1 below shows the difference between the within-item and between item multidimensionality.

For between-item model, the B matrix has nonzero elements that specify the coordinate dimension that is measurement target for the item, and the other elements are all zeros. For within item dimensionality, the B matrix has more than one nonzero element and these nonzero elements specify the coordinate dimensions that influence the performance of the test item.

Figure 3.1 Structures of between-item and within-item multidimensionality



From Adams, Wilson & Wang (1997) p. 9

The MRCMLM can be estimated by marginal maximum likelihood method (MML, Bock & Aitkin, 1981, see more details about MML in 3.2.3.). Bock & Aitkin's formulation of the EM algorithm[2] (Dempster, Laird, & Rubin, 1977) can be used to estimate structural item parameters.

[2] Expectation-maximization (EM) algorithm is a technique using iterative procedures to find the maximum likelihood estimators when there are unobserved latent parameters. The iterative procedures involve two major steps: first, use initial guess of the parameters to calculate the expectation of the log likelihood conditioned on the latent parameters; second, estimate the parameters by maximizing the expectation of the log likelihood obtained from the previous step. Then, use the new parameters as input to repeat the two steps until the likelihood does not increase much. More details can be found from: http://en.wikipedia.org/wiki/Expectation-maximization_algorithm

Estimation of the parameters in the MRCMLM is implemented in the ConQuest program (Wu, Adams, & Wilson, 1998). The use of MRCMLM model follows a confirmatory procedure that used to check hypotheses of a dimensional structure when a test has been designed to measure specific constructs. When estimating the item parameters of the model, the vector-valued person parameter θ is assumed to follow a multivariate normal distribution. ConQuest provides estimates of the person parameters, item parameters, means, variances, covariances and correlations of the latent dimensions, and deviance of the model.

There are also exploratory approaches to examine the dimensionality of the assessment data. In the exploratory approaches, the number of dimensions needed to accurately model the relationships in the item response matrix need to be determined first. This is determined from the interaction between a particular sample of examinees and the particular sample of items (Reckase, 2009). Parallel analysis, scree plots and residual correlation matrix are often used to determine the number of coordinate axes[3]. Software such as DIMTEST (Stout, Douglas, Junker & Roussos,1999; Stout, Froelich, & Gao. 2001), Ploy-DIMTEST and DETECT (Zhang & Stout, 1999) are commonly used to implement the procedures for determining the number of dimensions to model the item response matrix.

There are advantages and disadvantages of both confirmatory and exploratory approach. The advantage of the confirmatory approach is that the design and scoring matrices of the MRCMLM model the data explicitly according to the test developer's intended structure, so the results are easier to interpret and the fit statistics can be used as a diagnostic tool to confirm whether the theorized model is an acceptable description of the latent traits. But the confirmatory

[3] More details about the methods used to determine the number of coordinate axes such as parallel analysis, scree plots and residual correlation matrix can be found in Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York: Springer.

approach may ignore the relationships that are not specified in advance and may result in less model fit compare to the exploratory approach. The exploratory approach can achieve more data-model fit by selecting a model that fits the data best. But since the model is not specified in advance for confirmation, the result is often difficult to interpret and it sacrifices the use of fit statistics as a diagnostic tool to confirm the theorized model. In this study, since there are hypotheses and theories about what the assessment items assess, the confirmatory approach is applied.

Model fit indexes including the Akaike's (1973) information criterion[4] (AIC, Bozdogan, 1987) and the Bayesian information criterion[5] (BIC; Schwarz, 1978), can be used to compare the posited models. Chi-square[6] test can also be used to determine the model fit. In this study, the model fit indexes are estimated using ConQuest (Wu, Adams, & Wilson, 1998) and chi-square

---

[4] The Akaike's (1973) information criterion (AIC) is a statistics that tell the relative goodness of fit of nested models. It is defined as $AIC = -2\ \log(L_{max})+2k$, where L_{max} is the maximized likelihood and k is the number of model parameters. *AIC* is used to compare the relative goodness of fit of nested models and penalize the number of parameters in the model. The model corresponding to smaller *AIC* is a better model.

[5] The Bayesian information criterion is a statistics that tells the relative goodness of fit of nested models. It is defined as $BIC = -2\ \log(L_{max})+k\ log\ (N)$, where k is the number of model parameters and N is the number of data points. L_{max} is the maximized likelihood. The model with smaller *BIC* is a better model.

[6] In the context of this dissertation, the chi-square goodness of fit is a way to estimate how well the model fit the data. In general, the residual between data and model follows normal distribution. Then, the sum of the residual square will follow a chi-square distribution. A good fit require the reduced chi-square (the sum of the residual square divided by the number of the degree of freedom) ~ 1. If the reduced chi square $>> 1$, the model under fits the data. If the reduced chi square $<< 1$, then the model is over fit the data. The difference in the model deviance between two models approximately follows a chi-square distribution with degrees of freedom equal to the number of additional parameters estimated in the more complex model (Haberman, 1977). A significant test result will indicate that the full model fits the item response data significantly better than the reduced model. For more details, one can refer to: http://en.wikipedia.org/wiki/Goodness_of_fit

test is used to determine the model fit.

3.2.3 Maximum likelihood estimation

As I discussed in the previous section, the person ability and item parameters can be estimated by maximum likelihood estimation. Maximum likelihood estimation is a procedure of finding the value of one or more parameters that makes the observed data distribution the most probable. The maximum likelihood estimate for a parameter μ is denoted $\hat{\mu}$. Here I illustrate the maximum likelihood estimation method using the normal distribution as an example. Suppose there is a set of data, denoted as $\{X_i \mid i=1, 2, \ldots n\}$. X_i is the value of the i^{th} data in the set. If the data follow a normal distribution with mean μ and variance σ^2, then the likelihood function for each data point is given by:

$$L(X_i|\,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\,[-\frac{(x_i-\mu)^2}{2\sigma^2}] \tag{3.6}$$

The joint likelihood for all data points is given by:

$$f(X_1,\ldots,X_n|\mu,\sigma) = \prod_{i=1}^{N}[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\,[-\frac{(x_i-\mu)^2}{2\sigma^2}]] \tag{3.7}$$

$$= \frac{(2\pi)^{-\frac{n}{2}}}{\sigma^n} \exp\,[-\frac{\Sigma(x_i-\mu)^2}{2\sigma^2}] \tag{3.8}$$

The log likelihood function is

$$\ln(f) = -\frac{1}{2}\, n \ln(2\pi) - nln\sigma - \frac{\sum(x_i - \mu)}{\sigma^2}$$

(3.9)

Then the estimators of μ and σ can be obtained by maximizing the likelihood function. Taking the partial derivatives of ln(f) with respect to each of the parameters and setting it equals to zero yields:

$$\frac{\partial(lnf)}{\partial\mu} = \frac{\Sigma(x_i - \mu)}{\sigma^2} = 0$$

(3.10)

$$\frac{\partial(lnf)}{\partial\mu} = -\frac{n}{\sigma} + \frac{\Sigma(x_i - \mu)^2}{\sigma^3} = 0$$

(3.11)

Solving equation 3.10 and 3.11 yields:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

(3.12)

These are the well-known estimators for the mean and variance, but we obtained them through the maximum likelihood approach.

The above example shows the essence of the maximum likelihood estimation. As long as we write done the likelihood function, the model parameters can be easily estimated by maximizing the likelihood function. In the one-parameter logistic (1PL) model case, the corresponding likelihood is

$$L = \prod_{n=1}^{N} \prod_{i=1}^{k} \frac{e^{x_{ni}(\theta_n - \delta_i)}}{[1 + e^{x_{ni}(\theta_n - \delta_i)}]} \tag{3.13}$$

where X_{ni} is the response of nth student to the ith item.

θ_n is the latent trait (the ability parameter) of the nth student and

δ_i is the item difficulty of the ith item.

and the person ability as well as item difficulty can be estimated by maximizing this likelihood.

Marginal maximum likelihood estimation (MML) is a special case of the maximum likelihood estimation. Suppose we have a likelihood function with 4 parameters, for example, L (a,b,c,e). But we are only interested in two parameters, say, a and b. Then, we can marginalize the likelihood by integrating out the parameter c, and e, i.e., L(a,b) = $\int dcde\ L(a,b,c,e)$. Now, inistead of maximizing the likelihood L(a,b,c,e), we can get the marginal maximum likelihood estimators for a and b by marginalizing the marginal likelihood L(a,b). Here we can maximize a simpler function, but we need to integrate out the other parameters first.

3.2.4 Information function

In the previous sections, I introduced the general features of the IRT models, and outlined how to estimate the ability parameter as well as the item difficulty/threshold parameters by the maximum likelihood estimation. Since the central issue is to estimate the ability parameter for each student, a very important question is how precise the ability parameter is estimated. To address this question, another important concept is needed, the information.

In general, the term information tells us how much we know about something. The more information we have for a given quantity, the less uncertain we will be about it. Therefore, the information should be proportional to the inverse of the uncertainty. As I have shown that the person ability parameter is estimated by maximum likelihood estimation, then what we want to know is the information that is proportional to the inverse of the uncertainty of this maximum likelihood estimation.

In statistics, a mathematically consistent definition for information of such nature is called Fisher information, which is the inverse of the variance of the maximum likelihood estimator. It is defined as:

$$I(\theta) = E[\left(\frac{\partial}{\partial\theta} logL(\theta|X)\right)^2 |\theta]$$

(3.14)

Where E denotes taking expectation. In practice, taking the expectation for a complex function can be difficult. Therefore, an observed Fisher information is defined as:

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2} logL\ (\theta_{max}|X)$$

(3.15)

where $\theta_{\max}$ is the maximum likelihood estimator of θ. This equation shows that as long as we write done the likelihood function and obtain the maximum likelihood estimators of the parameters, we will be able to calculate the information of that parameter. It can be shown from statistical theory that the above defined Fisher information equals to the inverse of the variance of the parameter, i.e.

$$\mathrm{I}(\theta) = \frac{1}{\mathrm{var}(\hat{\theta}|\theta)} \tag{3.16}$$

where θ is the true ability, and $\hat{\theta}$ is the maximum likelihood estimator of the ability θ. Therefore, as long as we know the information, we will know the variance of the parameter we estimated via the maximum likelihood estimation and therefore know the precision of our estimation.

In our application, we want to know not only the ability parameter, but also its uncertainty. The information function can tell how precise one can estimate the ability parameter. The more precise we can measure the ability parameter, the better we can tell the difference among persons. Therefore, we want the information of the test to be adequately large.

3.3 Information and hypothesis testing

3.3.1 Some basic notation

A test with high information is desirable since it can differentiate smaller ability differences. There are two cases in terms of the ability differences: 1) the difference between two individual students; or 2) the difference between two groups of students. For an individual student, his/her true ability is denoted by θ_T, which is not directly measurable. What we have is a measured ability θ and associated measurement error on θ, denoted as ϵ. The measured θ and

the true θ_T, given the measurement error ϵ is sampled from a normal distribution, whose density is shown in the following equation:

$$p(\theta|\theta_T) = \frac{1}{\sqrt{2\pi\epsilon^2}} \exp\left[-\frac{(\theta-\theta_T)^2}{2\epsilon^2}\right] \tag{3.17}$$

The measurement error on θ relates the test information for the same θ as $\epsilon = 1/\sqrt{I(\theta)}$. Clearly, the higher the information is, the more precise we can measure the θ.

Now, let's consider a group of students, each with a true ability θ_{Ti}. In the group, all the true abilities of students follow a normal distribution as:

$$p(\theta_{T_i}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta_{T_i}-\mu)^2}{2\sigma^2}\right] \tag{3.18}$$

In terms of the corresponding observed θ_i, we have

$$p(\theta_i) = \int d\theta_{Ti} p(\theta_i|\theta_{Ti}) p(\theta_{Ti}) = \frac{1}{\sqrt{2\pi(\sigma^2+\epsilon_i{}^2)}} \exp\left[-\frac{(\theta_i-\mu)^2}{2(\sigma^2+\epsilon_i{}^2)}\right] \tag{3.19}$$

That is, the observed θ_i for the ith student follows a distribution with mean μ and variance $\sigma^2 + \epsilon_i^2 = \sigma^2 + 1/I(\theta_i)$. σ is the true group standard deviation. The maximum likelihood estimator of the group mean μ is given by:

$$\mu = \frac{\sum_{i=1}^{N}\left(\frac{\theta_i}{\sigma^2 + \epsilon_i^2}\right)}{\sum_{i=1}^{N}\left(\frac{1}{\sigma^2 + \epsilon_i^2}\right)}$$

(3.20)

Here, we need to note that the σ is not directly measurable. But it can be estimated by maximizing the likelihood

$$L(\theta) = \prod_{i=1}^{N} p(\theta_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi(\sigma^2 + \epsilon_i{}^2)}} \exp\left[-\frac{(\theta_i - \mu)^2}{2(\sigma^2 + \epsilon_i{}^2)}\right]$$

(3.21)

In the case $\in_i$ << σ, the well known estimators for μ and σ are recovered:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \theta_i$$

$$\sigma^2 = \frac{1}{(N-1)}\sum_{i=1}^{N} (\theta_i - \mu)^2$$

(3.22)

3.3.2 Hypothesis testing

A. Significance level and test power

In hypothesis testing, two types of errors are of interest. Type I error refers to the situation that we reject the null hypothesis when it is correct while the Type II error refers to the situation that we accept the null hypothesis when it is incorrect. The probability of making type I error is called significance level, usually denoted by α. The probability of NOT making type II error is called the power, often denoted by 1- β. In practice, as a rule of thumb, we usually choose α = 0.05 and 1- β = 0.8. In our current application, two types of hypothesis testing will be considered: 1) for any two students, we want to know if their true abilities are the same; 2) for two groups of students, we want to know if the group means are the same. I will show how the test information will affect these two types of hypothesis testing in this section.

B. Individual student case

For individual student, what we want to test is whether two students are different in their true θ_T s. We can run a *t*-test with the *t*-static defined as

$$t = \frac{(\theta_1 - \theta_2)}{\sqrt{\epsilon_1^2 + \epsilon_2^2}} = (\theta_1 - \theta_2)\sqrt{\frac{I(\theta_1)I(\theta_2)}{I(\theta_1) + I(\theta_2)}} \tag{3.23}$$

where θ_1 and θ_2 are the true abilities of the two students. ϵ_1 and ϵ_2 are their corresponding measurement errors. For a given significance level and test power, the greater the information, the smaller the difference between two students' abilities can be detected (θ_1 - θ_2). If the test

information is 10, it can detect the difference between two abilities with a difference of 0.92 on the logit scale at the significant level of .05 and power of .8.

C. Group student case

For the group-wise hypothesis testing, we want to know whether two groups of students are different in their mean true abilities θ_{T}. The *t*-statistics for two groups of students is given as:

$$\mathrm{t} = \frac{\overline{\theta_1} - \overline{\theta_2}}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \tag{3.24}$$

where

$\overline{\theta_1}$ denotes the mean of the true abilities of group 1.

$\overline{\theta_2}$ denotes the mean of the true abilities of group 2.

σ_1 denotes the standard deviation of the ability estimates of group 1.

σ_2 denotes the standard deviation of the ability estimates of group 2.

n_1 and n_2 are the sample size of group 1 and group 2.

However, the true ability is not directly measurable and we can only estimate them based on the measured θ. Though the mean of θ_T can be well estimated by the mean of θ, the variance of θ_T is related to the variance of the measured θ by

$$Var(\theta_T) = Var(\theta) - \epsilon^2$$

(3.25)

where we have assumed all the $\in_i$ are equal to ϵ for simplicity. If we do not have any information about the measurement errors on θ, we can only estimate the variance of the true ability as Var(θ). This will over-estimate the variance of the true ability, decreasing the sensitivity of the testing. However, if we know the measurement errors, we will be able to estimate the variance of the true ability via the Equation (3.25) above.

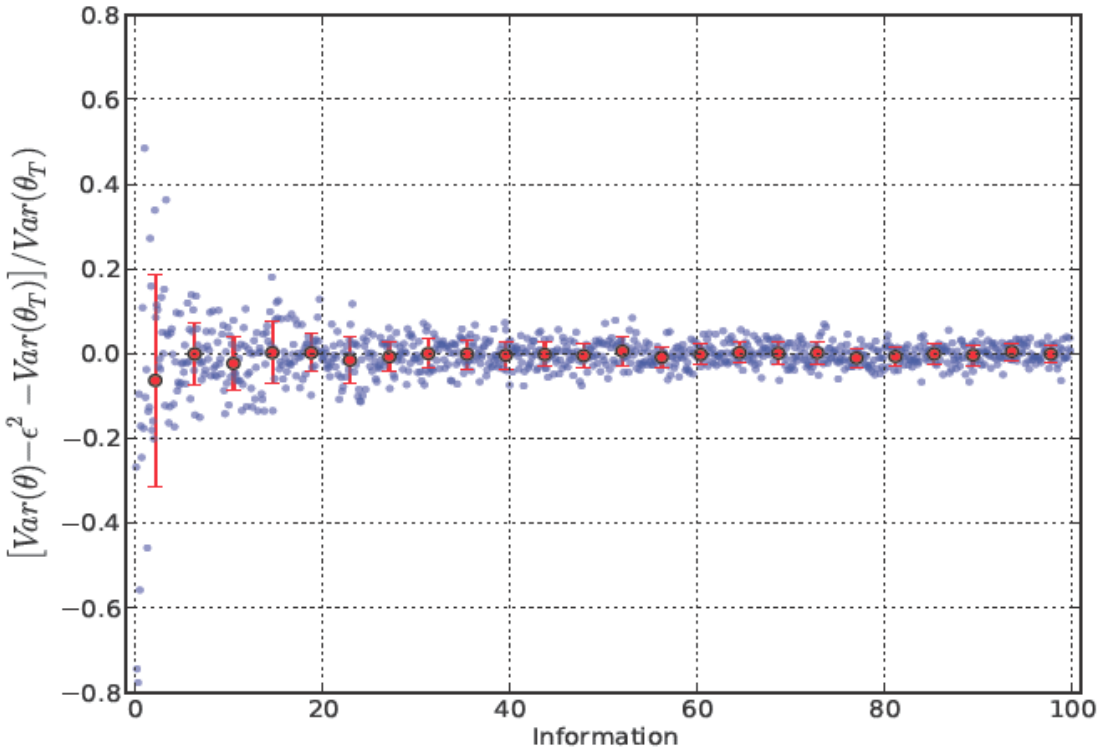
Measurement error is related to the *t*-statistics because it contributes to a small portion of the group variance (σ_1 and σ_2). Smaller measurement error can increase the differentiability a little bit by decreasing the group variance but it does not necessarily lead to a smaller variance of the group. If the test information is 10, then the test can detect a certain difference between two groups of students at a certain significance and power level. For example, if each group has 30 students and the true variance of the ability estimates of each group is assumed to be 0.6, the test can detect a difference of 0.87 between the mean ability estimates at the significance level of .05 and the power level of 0.8. If the test information is 5 and the other variables remain the same, then the test can detect a difference of 0.93 between the mean ability estimates at the same significance and power levels.

According to Equation (3.25), it seems that we can always recover the variance of the true ability, no matter how large the measurement errors are. Then, why the measurement errors need to be small? The key part here lies in that if the measurement error is large, then we cannot recover the variance of the true ability at the same precision as when the measurement error is small. To demonstrate this, I used a Monte Carlo simulation. First, I generated 100 θ_T from N(0;

1). Then, I assumed the measurement error ϵ varied from 0.1 to 100 with an increment of 0.1 each time. At each measurement error ϵ, the measured ability $\theta = \theta_T$+N (0,1) * ϵ Then, the recovered variance of the true ability θ_T was calculated by $Var(\theta) - \epsilon^2$. After that, I calculated the fractional error of the recovered variance with respect to the true variance $[Var(\theta) - \epsilon^2 - Var(\theta_T)] / Var(\theta_T)$ as a function of the test information (inverse of the measurement error square). The results are shown in Figure 3.2.

Figure 3.2 Fractional error of the recovered variance as a function of test information.

For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.



Note: The blue dots are results and the red dots are the mean in each bin of size 5. The red error bar is the standard deviation in each bin.

Clearly, as the information increases, the dispersion of the fraction error decreases, while when the information is small, the scatter of the fractional error can reach to 50%. If the information is sufficiently large, we will be able to recover the variance of the true ability with significantly less uncertainty. However, in practice, increasing the test information requires increasing the number of test questions, which is constrained by the amount of available test time. Therefore, a rule of thumb choice for test information is 10, which corresponds to a fractional error of ~ 10%. From the graph, the test information of 5 also corresponds to a fractional error that closes to 10%. So if the test is designed to test the difference between two groups of students rather than two individual students, information of 5 can be considered as sufficient. As long as we can estimate the variance of the true ability reliably, it is straightforward to calculate the *t*-statistic and run the *t*-test.

3.4 Learning progression and IRT

In the Environmental Literacy project, the researchers are trying to make claims such as: 1) the learning progression framework and the carbon cycle assessment are unidimensional. The learning progression levels, which represent increasing understanding and complexity, can be ordered along a continuum; 2) the assessment can accurately locate individual student's understanding within the framework; 3) the same student should be at similar levels across items.

Applying IRT methods can investigate the validity of these claims. According to the first claim, the learning progression framework is unidimensional, there should be a single dimension defined by the achievement levels that accounts for a significant portion of the variance in student performance. The dimension analysis can verify whether the unidimensional claim is true. It can tell whether students use the same ability or different abilities to answer the items. So

it has implication for the learning progression framework and the assessment development. Students' abilities are indicated by the achievement levels defined in the learning progression framework. Students' abilities are also indicated by the IRT ability estimates. If these two different definitions of ability reconcile with each other, it will provide evidence for the validity of the learning progression framework and the validity of the single latent construct defined by the four achievement levels. So IRT analysis can be conducted to examine both the validity of the framework and the validity of the assessment.

Second, applying IRT methods can reduce the measurement error so that students can be more accurately classified into learning progression levels. This ensures the reliability of the assessment and supports the second claim above.

Third, IRT methods can be used to check how consistent do students respond across items. If the item difficulty and item thresholds are similar across items, it means the same student will be at similar levels across items. So the third claim is valid.

There are other advantages of applying IRT methods. Generally speaking, person ability estimated from IRT models is a better measure of person's proficiency than the raw levels assigned by raters. It considers item characteristics such as item difficulty so it indicates students' ability more accurately. Because of all these advantages, IRT methods are applied in this study to test the validity of the learning progression claims.

Chapter 4 Methodology

In this chapter, I introduce the assessment developed, the assessment data collected, and the methods used to analyze the data. Section 4.1 describes the sample for this study. Section 4.2 introduces the test design. Section 4.3 describes the data scoring process and Section 4.4 briefly introduces the data analyses.

4.1 Data

During the 2009-2010 school year, the Environmental Literacy project collected written assessment data from elementary to high school students. The data were collected from twelve science teachers' classrooms, with four teachers at each level. These teachers used the teaching materials designed by the Environmental Literacy project research team. They administered the tests before and after they taught the teaching materials. So about half of the test data were collected in the pretest and the rest were collected in the posttest. In total, there are 1500 test papers from 10 rural and suburban schools. The more specific numbers of tests from each grade level and from pre or posttest are described in Table 4.1.

Table 4.1 Number of test papers collected during 2009-2010

	Pre	Post	Grade Sum
Elementary	167	149	316
Middle	288	439	727
High	262	195	457
Pre/Post Sum	717	783	Total: 1500

4.2 The carbon cycle test design

The carbon cycle assessment was designed for each of the three grade levels: elementary, middle and high school. At each grade level, there were three alternative forms. The items administered at each grade level were selected to be appropriate for students at that grade level.

Table 4.2 below describes the three test forms and the items that have been administered at the high school level.

Table 4.2 Three alternative high school test forms[7]

High Form A	High Form B	High Form C
[Photosynthesis items]	[Digestion items]	[Cellular respiration items]
CARBPATH	EATAPPLE	GLUGRAPE
ENERPLNT	*INFANT*	**WTLOSS**
THINGTREE	CARBBODY	**BODYTEMP**
PLANTGAS	*ENERPEOP*	AIRNBODY
[Decomposition items]	[Combustion items]	[Human energy system items]
TREEDECAY	GASOLINE CAR	*GLOBWARM*
POTATO	**BRNMATCH**	LAMPELEC
BREADMOLD	WAXBURN	KLGSEASON
[Cross process items]	[Cross process items]	[Cross process items]
EATBRTHE	GRANJOHN	**DEERWOLF**
ECOSPHERE	DIFEVENTS	**TROPRAIN**
[Linking items]	[Linking items]	[Linking items]
ENERPEOP(AB)	PLANTGAS(AB)	**BRNMATCH(BC)**
TROPRAIN(AC)	**WTLOSS(BC)**	**BREADMOLD(AC)**

All items assess students' understanding of matter or/and energy transformation in six carbon transforming processes— plant growth, animal growth, animal function, combustion, decomposition and cross-process events. The items in bold font are the OMC + CR items. The items in italic font are the MTF+CR items. The rest are the CR items. There are linking items across the forms, which are in the last two rows of Table 4.2. About 20% of the items are the linking items. Most of the items administered at high school level are different from those

[7] In Table 4.2, each item is named using some key words in the item. For example, EATAPPLE item asks students the question: An **apple** is **eat**en by a boy and digested in his body. What happens to the apple when it is digested?

administered at the elementary level. Items administered at the middle school level are a combination of the elementary and the high school items. Some items are used across all three-grade levels as vertical linking anchor items.

There are 43 items in total, including 25 CR items, 7 OMC + CR items, 10 MTF + CR items and 1 MTF item. All the items are listed in Table 4.3 below. This item pool was developed based on the item pool used in the previous year (2008-2009). Some of the items were used in the previous year but were modified according to the assessment results, while some were newly developed. Except four elementary items, all the other items have more than 100 responses. The numbers of responses for each item are shown in Table 4.3. These numbers are different from item to item, depending on whether the item is an anchor item that is used across forms or grade levels.

Table 4.3 Number of responses per item

Item ID	Item format	Item label	Number of OMC/MTF responses	Number of CR responses
1	OMC+CR	ACRON	612	601
2	OMC+CR	BREADMOLD	418	397
3	OMC+CR	BRNMATCHA (M/H)	516	525
4	OMC+CR	DEERWOLF	232	233
5	OMC+CR	TROPRAIN	637	641
6	OMC+CR	BODYTEMP	133	137
7	OMC+CR	WTLOSS	886	885
8	MTF+CR	AIREVENT	189	198
9	MTF+CR	ANIMWINTER	79	80
10	MTF+CR	ENERPEOP	469	900
11	MTF+CR	ENERPLNT	508	522
12	MTF+CR	GLOBWARM(M)	172	185
13	MTF+CR	GLOBWARM(H)	132	136
14	MTF+CR	INFANT	457	455
15	MTF+CR	OCTAMOLE	148	148
16	MTF+CR	THINGTREE	585	598
17	MTF+CR	STONEWIN	191	198
18	MTF	POTATO	162	N/A

Table4.3 (Cont'd)

19	CR	AIRNBODY	N/A	339
20	CR	APPLEROT	N/A	253
21	CR	BRNMATCH (E)	N/A	182
22	CR	BRNMATCHB (M/H)	N/A	492
23	CR	CARBODY	N/A	254
24	CR	CARBPATH	N/A	361
25	CR	CARGAS	N/A	355
26	CR	CONNLIFE	N/A	220
27	CR	CUTTREE	N/A	75
28	CR	DIFEVENT	N/A	412
29	CR	EATAPPLE	N/A	449
30	CR	EATBRTHE	N/A	348
31	CR	ECOSPHERE	N/A	283
32	CR	GIRLRUNNAB	N/A	69
33	CR	GIRLRUNNC	N/A	68
34	CR	GLUGRAPE	N/A	265
35	CR	GRANJOHN (D)	N/A	121
36	CR	GRANJOHN (P)	N/A	102
37	CR	GROWTH	N/A	352
38	CR	KLGSEASON	N/A	124
39	CR	LAMPELEC	N/A	318
40	CR	PLANTGAS	N/A	359
41	CR	TREEDECAYAB	N/A	741
42	CR	TREEDECAYC	N/A	653
43	CR	WAXBURN	N/A	260

An example of MTF + CR item is given below:

Example MTF item: The baby gained more and more weight as she grew. Where did her weight come from? Please circle Yes or No for each of the following and explain your choices.

a. Sunlight Yes / No

b. Water Yes / No

c. Air Yes / No

d. Nutrients Yes / No

e. Foods Yes / No

f. Exercises Yes / No

Paired CR item: Please explain your answer. Try to explain what happens inside the girl's body to each of the materials that you circled "Yes."

4.3 Scoring rubrics and coding process

Students' responses were coded by nine raters in the Environmental Literacy research team. These raters majored in either science education or educational measurement. All raters were familiar with the learning progression framework and the scoring rubrics. They all had some experience in coding similar written assessment data. Seven of the raters had worked in the project for over two years and had coded hundreds of responses collected every year. The other two raters who joined in the project in the last year went through the coding training and coding practice to ensure high accuracy for their coding.

The responses to the CR items and the CR part of the two tier items were coded using the generic rubrics and the item specific rubrics. The generic rubric has general level descriptions that describe the general characteristics across items (See Table A.1 in the Appendix). The item-

specific rubrics have specific level descriptions of each level for each item and representative example responses for each level (See Table A.2 in the Appendix for an example). These rubrics were developed in previous studies and refined during the coding process to distinguish responses more clearly. Ten percent of the responses were double coded by a second rater. The inter-rater reliability[8] between the first and second raters was higher than 80% for all items. Discrepancies in coding were discussed and final agreements were reached for each response.

Students' responses to the OMC questions were recoded into levels according to the level of understanding the option represents. Appendix A lists all the items and the level of each OMC option. Students' responses to the MTF item, which included a string of T or F responses, were also recoded into levels based on the number of correct choices made by the students. For example, level 1 means the student made one correct choice and level 4 means the student made four correct choices. The correct answers of the MTF items are in bold in Appendix A.

4.4 Data analysis

Both classical test statistics and IRT based statistics were used to evaluate the quality of the items. Item discrimination indices were analyzed. It is the correlation between the students' scores (the final score agreed by both raters) on the item and their total scores. IRT based statistics such as item fit indices, item difficulty, step difficulty and measurement error were also used to evaluate the quality of item. Since there were common anchor items across grade levels and test forms, the entire data matrix followed the common anchor item design. The combined set of items used at all grade levels was calibrated through a concurrent calibration using IRT models. ConQuest was used to estimate both the item and person ability parameters. Open source

[8] This is the exact plus adjacent agreement that includes differences within 0.5 level between two raters.

software R and Microsoft Excel were used to calculate the summary statistics.

The ConQuest program can provide various IRT based statistics. For example, it can provide person ability estimates such as the Expected A Posterior (EAP)[9] estimates and maximum likelihood estimates (MLE, see section 3.2.3 for detailed explanation of MLE). It can also provide fit statistics for individual items. These are the residual-based indices such as the weighted and unweighted fit statistics developed by Wright and Masters (1982). Weighted fit statistics are usually preferred because they are less sensitive to unexpected responses made by persons for whom the item of interest is far too easy or far too difficult. Wu (1997) has shown that these statistics have approximate scaled chi-square distributions and can be transformed to approximate normal deviates (*t*-values). An item is considered as a misfit item if the absolute value of its associated *t*-statistic is greater than 2.0. A *t*-value greater than 4.0 or less than - 4.0 indicates serious misfit. Items that do not converge or show poor IRT fit may have low quality and may cause potential problems when included into the test.

[9] The Expected A Posterior (EAP) estimates of person combine the item calibrations, prior rough idea of person ability, and the observed responses to obtain improved, a posteriori person ability measure. It is based on the posterior probability distribution of the ability parameter. Suppose we want to estimate the ability parameter and we know the posterior distribution is

$$\mathrm{p}(\theta|\mathrm{data}) = \frac{\mathrm{p}(\mathrm{data}|\,\theta)\mathrm{p}(\theta)}{\mathrm{p}(\mathrm{data})}$$

Then, we have

$$\mathrm{EAP}(\theta) = \frac{\int_{-\infty}^{\infty} \theta \mathrm{p}(\theta|\mathrm{data})\mathrm{d}\theta}{\int_{-\infty}^{\infty} \mathrm{p}(\theta|\mathrm{data})\mathrm{d}\theta}$$

Chapter 5 Design of a Test Consisting of Items in Multiple Formats

5.1 Research purpose and procedure

The first research question is how to design a test using items in multiple formats to precisely classify students into levels. This is one central question of this study. To measure students' learning progression as accurately as possible, the most appropriate items need to be selected. Three steps of data analyses were conducted to address this question:

First, the dimensionality analysis was conducted to see whether items in different formats assess the same ability or not. A confirmatory approach was applied to analyze the dimensionality using the subjective classification of items according to their formats. Both unidimensional PCM model and the multidimensional PCM model were used to fit students' OMC, MTF and CR scores. The appropriate model was selected according to the chi-square goodness of fit test and how well it is supported by the theoretical underpinning.

Second, after the item parameters were calibrated using the selected model, items were selected to form a test based on certain test design criteria. More details about the design criteria and how items were selected to meet these criteria were discussed in Section 5.4.

Third, to design OMC, MTF items to accurately classify students, discriminative OMC/MTF options are needed. Correlation and cross tabulation analyses were conducted to see how well the OMC/MTF choices relate to students' abilities or their CR levels. To design better CR items, whether the CR items accurately classified students into levels was analyzed.

5.2 Classical test statistics—Item discrimination index

The item discrimination was analyzed as a first check of the item quality. It was computed as the correlation between the students' scores (the final scores agreed by both raters)

on the item and their total scores. The item discrimination indices of most items were higher than 0.3. However, the item discrimination indices of nine OMC and MTF items were lower than 0.3 and four of them were even lower than 0.2. These OMC items were BODYTEMP (H) (0.20), TROPRAIN (0.22) and WTLOSS (0.24). The MTF items were AIREVENT (0.24), ANIMWINT (0.17), BODYTEMP (0.25), INFANT (0.16), POTATO (0.18) and THINGTREE (0.19).

A discrimination index value below 0.30 indicated that an item might not be measuring what it was intended to measure, and should be reviewed. The discriminations of all CR items were higher than .30. Since the test mainly consisted of CR items, the low discrimination of OMC and MTF items indicated that students' OMC and MTF scores did not strongly correlate to their CR scores. There might be multidimensionality in terms of item format. This was examined in the dimension analysis.

5.3 Dimensionality in terms of item format

The dimensionality analysis was conducted to test whether items in different format assessed the same ability. Two models were used to fit students' response codes: a unidimensional Partial Credit Model (1PCM) and a three-dimensional Partial Credit Model (3PCM) that classified items into three dimensions according to their formats. To compare the goodness of fit between these two models, a chi-square test was performed on the difference between the deviances of these two models. The difference in the model deviance approximately follows a chi-square distribution with degrees of freedom equal to the number of additional parameters estimated in the more complex model (Haberman, 1977). The difference between model deviances was 420. The additional number of parameters of the 3PCM was 5. A chi-square statistics of 420 with degree of freedom of 5 was statistically significant at 0.001 level. So the 3PCM fit the data significantly better than the 1PCM.

There was increase in model fit by applying 3PCM, which suggested the additional dimensions could explain some of the variance in student performance. However, there were moderate correlations among these three latent dimensions and strong correlations between the abilities estimated using the 1PCM and the abilities in each dimension estimated using the 3PCM. The correlation between the OMC and the CR dimension was 0.69 and the correlation between the MTF and the CR dimension was slightly higher, which was 0.76. The disattenuated correlations[10] between the unidimensional ability estimates and the ability estimates in three dimensions were all over 0.9. Table 5.1 provided all these correlations. These high correlations indicated that one dimension was sufficient to approximate student's ability.

Though the dimensionality analysis suggested that the OMC, MTF and CR items might measure different components of the construct, the goal of the assessment is to design OMC, MTF and CR items that assess the same construct. The OMC and MTF items need to be revised to predict students' CR levels more accurately (Section 5.5.1 and 5.5.2 discuss about this in detail). A unidimensional model is supported by the cognitive theories underlie the assessment design. Multiple dimensions are only necessary when we want to consider the nuisance dimensions for a particular measurement purpose.

Table 5.1 Correlations among the EAP estimates in each dimension

	Dimension 1 (OMC)	Dimension 2 (CR)	Dimension 3 (MTF)
Dimension 1 (OMC)	1		
Dimension 2 (CR)	0.69	1	
Dimension 3 (MTF)	0.44	0.76	1
Unidimensional ability	0.94	0.99	0.96

[10] Disattenuated correlation is the correlation between two sets of parameters that accounts for measurement error contained within the estimates of those parameters. The measurement error of the EAP ability estimates is accounted in the disattenuated correlations.

Note: the correlations in the last row between unidimensional ability estimates and multidimensional ability estimates are disattenuated correlation.

So the 1PCM was applied to analyze the data. When the 1PCM was applied to the data, a student's ability measured could be considered as a composite of the abilities in the multidimensions. The results showed most items fitted well with the 1PCM model. The MNSQs were within [0.67,1.33] and the *t*-statistics are within [-2,2]. This indicated in general, the learning progression framework and the unidimensional assumption were supported. Table A.3 in the Appendix has a list of the item difficulty parameter estimates and the fit statistics from the 1PCM results. Nine items did not fit well with the 1PCM model (see Table A.3 in the Appendix). These items need to be reviewed.

5.4 Select items to meet the design criteria

5.4.1 Design criteria of the learning progression-based carbon cycle assessment

There are some general considerations for the test design such as reliability and validity. However, each research project has its own unique goals and the design criteria must be made in a way so that the results can support the inferences that one wants to make. In the Environmental Literacy project, the following criteria are what we are specifically interested in:

1) High information at the boundaries between levels on the IRT ability scale (we are less concerned with information within the boundaries). So the boundaries need to be defined first.
2) Similar item step thresholds across items. For leaning progression-based items, ideally, students at the same ability level will get the same level across all items. This means that the item step thresholds need to be similar across items.
3) Detect the differences between classes of about 30 students (e.g. pre posttest difference at

the class level)

4) Use more OMC and MTF items in the test to reduce scoring effort.

5.4.2 Design a test to meet each criterion

1) Define boundaries on the IRT scale

First, if the ability defined by the IRT scale reconciles with the ability defined by the achievement level codes, it means the construct—the learning progression framework is generally supported. To test whether these two defined abilities reconcile with each other, boundaries should be set on the IRT scale first to classify students into levels.

Classifying students based on the IRT scale has some advantages. A student may be at different levels on different items. How to decide his/her achievement level in general? The ability estimated from IRT analysis is based on students' responses to all items and takes the item characteristics into account. So it's a better measure of students' ability. Therefore, we want to define boundaries on the IRT scale to classify students into levels. This is usually conducted in standard setting process which is a complex process that beyond the scope of this study.

A simpler way was applied in this study to set the boundaries. The mean of the item thresholds across a set of good items was taken as where the estimated boundaries should be. The set of good items were the items that fitted well with the 1PCM model and their thresholds were in the correct order. The boundaries were set based on 38 items that considered as good items (about 2/3 of all items). Table A.4 in the Appendix listed the items included in the good item set and Table A.5 in the Appendix listed the 22 items that were excluded. The threshold parameters of these excluded items do not represent the boundaries well. Half of these excluded items were OMC and MTF items that did not fit well with the unidimensional model. The other half are CR items. Some of these CR items have problems in their scoring rubrics and some of

these CR items cannot differentiate students at some achievement levels. The problems of these CR items will be discussed in more detail in section 5.5.3. Therefore, it's reasonable to exclude these 22 items when setting the boundaries.

The thresholds of the good items were relatively close to each other. The mean of the thresholds between level 1 and 2 (d_1), level 2 and 3 (d_2), and level 3 and 4 (d_3) were -1.7, 0.5 and 1.9 respectively. So these were considered as where the estimated thresholds should be. When the boundaries are set, students can be classified into levels according to their ability estimates.

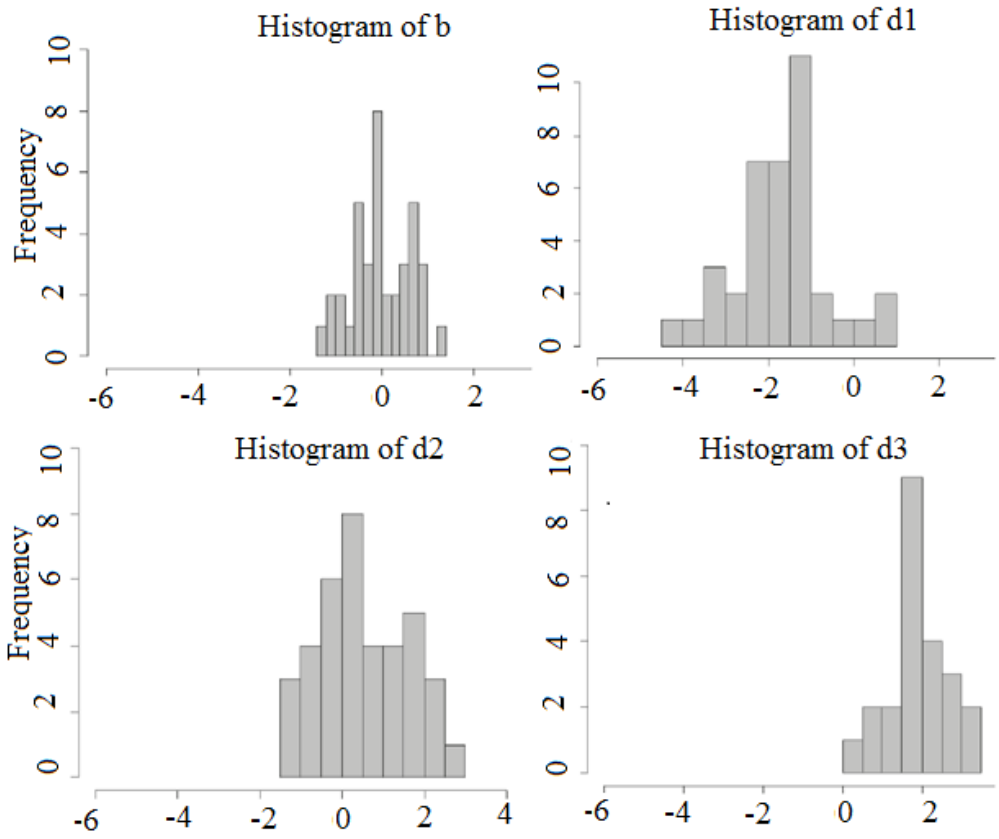
2) Similar item step thresholds across items

Some items had thresholds that were closer to the estimated boundaries than others. Table 5.2 included the descriptive statistics of the item parameters and Figure 5.1 showed the distribution of item parameters, including the item difficulty parameter (b) and item step threshold parameters (d_1, d_2, d_3). The second design criterion is to have similar locations of thresholds across items. Ideally, all the items should have thresholds at those boundaries.

Table 5.2 Descriptive statistics of the item parameters of 38 selected good items

	b	d_1	d_2	d_3
Mean	0.000	-1.72	0.52	1.91
Median	-0.03	-1.67	0.30	1.82
S.D.	0.65	1.12	1.10	0.80
Skewness	-0.04	0.16	0.32	0.10
Maximum	1.32	0.96	2.90	3.50
Minimum	-1.21	-4.43	-1.28	0.42

Figure 5.1 Item difficulty (b) and threshold (d) parameter distribution



It can be seen that the threshold parameters across items vary a bit. If the items and rubrics are well designed and the scorings are reliable, then the variance of the threshold parameters, d_1, d_2 and d_3 should be small. The items with d_1, d_2 and d_3 deviated far away from the mean values is a sign that either the item or the scoring is not appropriate. For example, on the "histogram of d_1 graph", the d_1 of some items (e.g. TROPRAIN, GLUGRAPE, GLOBWARM) are much smaller than the others. These items are not discriminative at the lowest level—Level 1. However, either because the scoring rubric did not clarify that these items were not discriminative at Level 1, or because of coding mistakes, a small proportion of the

responses were coded as Level 1. So the threshold parameter between level 1 and level 2 (d_1) of these items are much smaller than the d_1 of the other items.

3) Detect the differences between classes

To accurately classify students into levels, the measurement errors at the boundaries should be small. This means the information at the defined boundaries needs to be high. Since the third design criterion is to detect the difference between classes, information of 5 or above on these boundaries can be considered as sufficient (see Section 3.3.2 about why selecting 5 as the criterion). Items were selected to form a test that could get information above 5 at the defined boundaries.

Items were randomly selected from the good item set to see how many items were needed. The result suggested around 16 items were needed. The information curve formed by these 16 items is shown in Figure 5.2 below. According to the second design criterion, all item thresholds should be close to the defined boundaries. So some ideal items that have thresholds at those boundaries are simulated. Then these ideal items are selected to see how many items are needed to achieve information above 5 at the boundaries. The result shows around 14 items are needed. The information curve formed by these 14 ideal items is shown in Figure 5.2 below.

Figure 5.2 Information curve of 16 real items

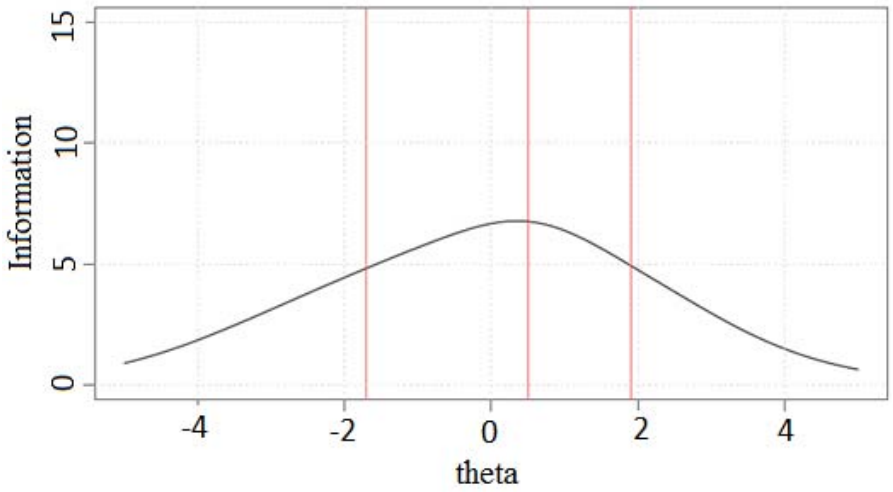
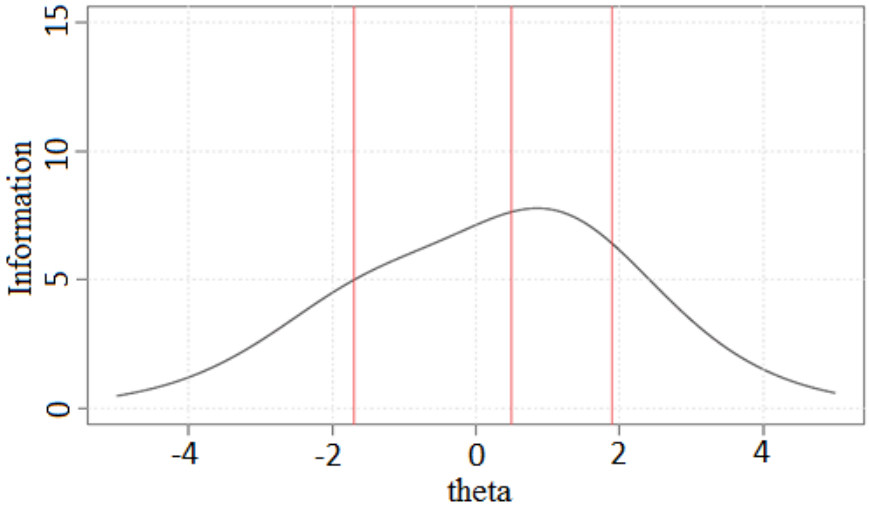


Figure 5.3 Information curve of 14 simulated ideal items



Sixteen real items are required to get information above 5 at the boundaries, but only 14 items are required to get information above 5 at the boundaries when item thresholds are the same across items. So adjusting rubrics to get similar thresholds across items will slightly reduce the number of items needed to reach the information criterion at the boundaries. Most importantly, having similar thresholds across items is an important criterion to design learning

progression-based items. So it is worth to modify the item or adjust the rubrics to get similar item thresholds across items.

4) Using more OMC and MTF items in the test

The OMC and MTF item formats can be used as alternative formats to reduce the scoring effort and administration time as long as they can accurately classify students into levels and can elicit responses consistent with those elicited by the CR items. Including OMC/MTF items in the test might also give information about students' other abilities such as the ability to identify the best/correct answer. However, in the prior analysis, some OMC items and MTF items do not fit well with the unidimensional model. This indicates too much randomness in the data so these OMC/MTF items do not perform well to classify students into levels.

Three problems may cause the misfit. The first problem is the OMC and MTF items assess different aspects of the construct (e.g. the ability to recognize a correct answer) as discussed in the dimensionality analysis previously. So the unidimensional model does not fit well. The second possible problem is the quality of the OMC and MTF items. Some of the OMC and MTF options cannot discriminate students appropriately. Section 5.5.1 and 5.5.2 will discuss this problem and how to design better OMC and MTF options in detail. The third problem is the PCM model may not be the best model for OMC and MTF items. The randomness in the OMC and MTF data might due to the restricted range of responses and the guessing effect. In addition, some OMC items have two options representing understanding at the same level, so the probabilities at each level by chance are different. Some new models are under development for the OMC items but they are limited to certain types of OMC items which may not be appropriate for the OMC items in this study.

The misfit of the OMC and MTF items can be resolved in two ways. First, since the PCM

might not be the best model for OMC and MTF item, the one-parameter logistic model (1PL) is applied to the OMC and MTF data by recoding the OMC and MTF items as dichotomous items (1- if student choose the best answer or choose all correct answers, 0-otherwise). The result shows that one OMC item shows slight misfit, all the other recoded OMC and MTF data fit well with the 1PL model. The difficulties of the 7 OMC and 10 MTF items are listed in Table 5.3 below. Most of the MTF items are difficult which indicates that students need to have high ability to choose all correct answers. If the item difficulty is close to a boundary, the item can still work well to classify students between levels.

Table 5.3 OMC and MTF item difficulty (recoded as dichotomous items)

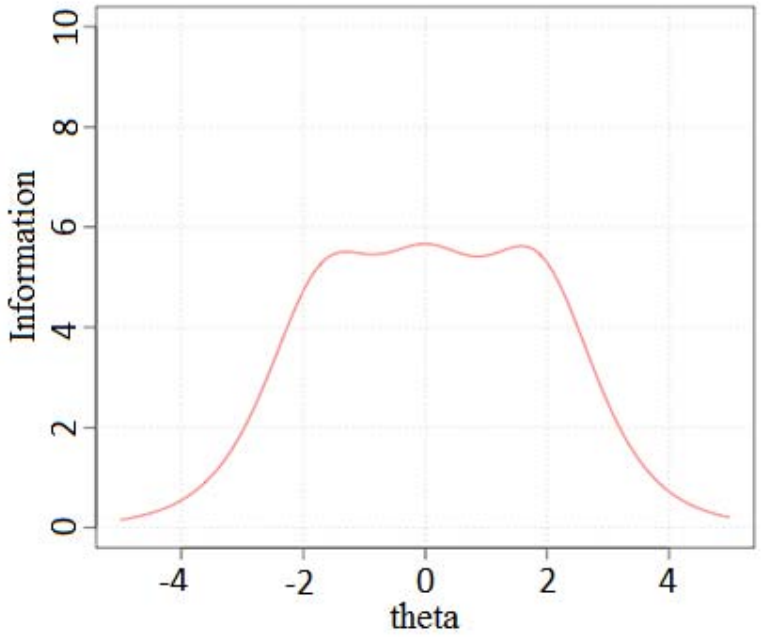
OMC items	Difficulty
ACRON_OMC	-0.411
BODYTEMP_OMC	-2.037
BREAD_OMC	-1.271
BRNMATCH(M)_OMC	0.405
DEERWOLF_OMC	0.863
TROPRAIN_OMC	0.172
WTLOSS_OMC	-0.236
MTF items	Difficulty
AIREVENT_MTF	-0.552
ANIMWINT_MTF	2.652
STONEWIN_MTF	2.355
ENERPEOP_MTF	1.138
ENERPLNT_MTF	1.622
GLOBWARM(M)_MTF	1.428
GLOBWARM(H)_MTF	3.914
INFANT_MTF	1.904
OCTAMOLE_MTF	1.798
POTATO_MTF	2.498

The second way to solve misfit is that, since some MTF options are not as discriminative as others (Section 5.5.2 will discuss this problem in detail), responses from the less discriminative sub-questions can be excluded from data analysis. Students' MTF scores are calculated based on their responses to the discriminative options only. I discuss about this further

in Section 5.5.2. Because some MTF items don't have any discriminative option, this approach is not sufficient to improve the fit statistics of MTF items significantly. Section 5.5.1 and 5.5.2 address other possibilities to improve the quality of OMC and MTF items.

Besides using OMC and MTF items, an alternative way to reduce the scoring effort and administration time is to develop some dichotomously scored items such as multiple-choice or True-or-False items in the test. If the difficulties of the dichotomous items are at the defined boundary, these items can help to accurately classify students. For example, if there are three well-designed CR items that have thresholds at the defined boundaries and there are 14 dichotomous items with difficulty at the boundaries (five at -1.7, five at 1.9 and four at 0), the information curve formed by these items is shown in the graph below:

Figure 5.4 Information curve formed by 3 CR with thresholds at the boundaries and 14 ideal dichotomous items with item difficulties at the boundaries



The information curve is above 5 across all boundaries. So these 3 CR and 14 dichotomous items can form a reliable test to classify students into four achievement levels. The next question that arises is how to develop dichotomous items that have difficulties close to the boundaries. The difficulties of the recoded OMC items can provide some information about what kind of dichotomous items that may need to be developed.

5.5 Design discriminative OMC, MTF and CR items

Using OMC/MTF items can reduce the test administration time and scoring effort. In order to use items in these formats in the test, OMC and MTF items need to either accurately classify students into several achievement levels or distinguish students between two adjacent levels (OMC and MTF items act as dichotomous items). The previous analysis shows when

OMC and MTF items are scored as dichotomous items, they fit well with the 1PL model in general. When the rescored OMC and MTF items have difficulties that are close to one of the estimated boundaries (-1.7, 0.5, 1.9), they can be used to classify responses between two levels. For example, OMC items are generally easy and can be used to classify level 1 and level 2 students. MTF items are generally more difficult and can be used to classify level 3 and level 4 students.

On the other hand, OMC and MTF items can be designed to perform better as polytomous items. The quantitative analyses suggest many OMC or MTF items have poor item statistics when coded as polytomous items. The item discrimination indices of some items are lower than 0.20. Some items do not fit the unidimensional model or the step thresholds of the item are not in the correct order. Since the test mainly consisted of CR items, the low discrimination indices of OMC and MTF items indicate that students' OMC and MTF scores do not strongly correlate to their CR scores. The misfit and the incorrect order of the step thresholds indicate problems with the OMC and MTF items. So sections 5.5.1 and 5.5.2 discuss how to design OMC and MTF options to align with CR items.

5.5.1. OMC options

To evaluate how well students' OMC levels predict their CR levels, students' levels on the OMC questions are cross-tabulated with their levels on the paired CR questions. The result shows the OMC level can predict the CR level to some extent, but there are cases that the OMC level over-predicts or under-predicts CR levels. More often, the OMC level over-predicts the CR level. Take the ACORN item as an example, this item asks students to identify where the weight of the tree comes from.

ACORN

OMC part: A small acorn grows into a large oak tree. Where does most of the weight of the oak tree come from? (Circle the best explanation from the list below).

A). From the natural growth of the tree (level 1)

B). From carbon dioxide in the air and water in the soil (level 3)

C). From nutrients that the tree absorbs through its roots (level 2)

D). From sunlight that the tree uses for food (level 1)

Paired CR part: Please explain why you think that the answer you chose is better than the others. (If you think some of the other answers are also partially right, please explain that, too.)

Table 5.4 Cross-tabulation between OMC levels and CR levels for ACORN item

	OMC response levels			
CR levels	Level 1 (A)	Level 3 (B)	Level 2 (C)	Level 1 (D)
0	6	7	8	2
1	105	14	69	36
2	21	88	170	24
3	3	26	9	4
4	0	2	0	0

Table 5.4Table 5.4 shows how students' choices for the OMC part cross-tabulate with their levels of the paired CR part. In the table, the numbers in the grey cells represent the number of cases that the OMC level is consistent with the CR level. There are relatively large counts in the grey cells, which means that students' CR responses were coded the same as their OMC responses. So the OMC options can predict students' levels for the CR part to some extent.

The counts in the other cells served as evidence that the different item formats are not eliciting consistently coded responses from students. The numbers in the cells above the grey

cells represent the cases that the OMC part over-estimate students' CR levels and the numbers in the cells below the grey cells represent the cases that the OMC part under-estimate students' CR levels. There are more over-estimations than under-estimations, which indicate that the OMC question is easier than the CR question. Many students could identify that the weight came from CO_2 and water but could not explain how CO_2 and water contributed to weight gain. For instance, one student selected the choice B but his/her response to the CR question was "the carbon in the air and the water in the soil makes the weight more at the bottom then on top". Clearly, the students could not explain the photosynthesis process. This is also true with the other OMC+CR items. Students' OMC levels are usually higher than their paired CR levels. The hypothesis for the discrepancy between students' OMC levels and their CR levels is that students perform better when identifying the input and output of carbon transforming processes than explaining what happens during the processes. Since most OMC questions assess the former ability and most CR questions assess the latter ability, students' OMC levels are usually higher than their CR levels.

The cross-tabulation analysis also shows that in most cases, the OMC options associated with range of levels rather than a single level. For example, the BODYTEMP item (below) asks students where human body heat mainly comes from. Students who selected the correct option (option C) of the OMC question tended to be at multiple levels in the corresponding CR part. For instance, one student who selected option C gave a level 4 explanation: "because the food is then broken down and the chemical energy in that food is then changed into thermal energy to keep you warm". However, another student who selected option C provided a level 1 explanation: "I think all of these answers were legitimate, but I choose C because you have to eat food, everyone does, so it's a natural function, it would make sense to make heat from something you do daily."

So the OMC option cannot predict the level of the student's CR response very well.

BODYTEMP

OMC part: Your body needs heat to keep its normal temperature. Where does the heat mainly come from? Please choose ONE answer that you think is best.

A) The heat mainly comes from sunlight. (Level 1)

B) The heat mainly comes from the clothes you are wearing. (Level 1)

C) The heat mainly comes from the foods you eat. (Level 3)

D) When people exercise, their bodies create energy. (Level 2)

Paired CR part: Please explain why you think that the answer you chose is better than the others. (If you think some of the other answers are also partially right, please explain that, too.)

In total, 91 students selected option C. Among these students, 59 of them are at level 3, 20 are at level 4 and the others are at level 1 or 2 on the paired CR part.

Table 5.5 shows how the OMC levels cross-tabulate the CR levels. Most of the OMC options of this item represent understanding at lower levels. Hence, these OMC options under-predict students' CR levels.

Table 5.5 Cross-tabulation between OMC levels and CR levels for BODYTE item

	OMC response levels			
CR response levels	Level 1 (A)	Level 1 (B)	Level 2 (D)	Level 3 (C)
1	8	3	11	3
2	1	1	11	9
3	1	0	6	59
4	0	0	0	20

In summary, OMC options can predict students' CR level to some extent. But in most cases, the OMC questions are not true OMC questions. There are cases that OMC level over-predict or under-predict CR levels. Often times, OMC level over-predict CR levels. When all the OMC options represent understanding at low levels, the OMC levels may under-predict students' real achievement level. So in order to design OMC options that can better predict students' CR levels, the OMC options need to represent understanding at multiple levels.

5.5.2 MTF options

The MTF options of the MTF+CR items were analyzed in similar ways to see how students' T or F responses were related to their CR responses. If students' T or F responses can predict their levels on the paired CR question, then MTF can be used instead of the CR format to detect students' achievement levels.

First, how students' response strings to the set of T or F questions were related to their levels on the paired CR item was analyzed. Then, the relation between students' responses to each sub T or F question and their CR levels was analyzed. The main findings were summarized below:

1) Students who selected all correct responses were usually at very high CR levels and had high ability estimates. This suggests that the set of T or F questions is useful to identify the students at the high ability range.
2) Students' number of correct choices can detect the achievement level of students who are at the middle and lower ability range.
3) The patterns of T or F response string do not associate with students' paired CR levels clearly. This is mainly because students often select both low level and high-level options since they do not have enough sophisticated understanding to rule out the lower

level distracters.

4) Some of the “T” or “F” questions work better to differentiate students than others.

In the following paragraphs, examples are given to illustrate each of these four main findings about the MTF items.

First, take the ENERPLNT item as an example. Table 5.6 describes the percentages of the most common T or F response strings. Among 508 students, only 4% of them (20 students) correctly identify "sunlight" as the only energy source for plant growth. The average of their paired CR level is 3.9 and the average of their ability estimates is .906, which is significantly higher than the other students. This is a common pattern across most of the MTF items. Students who gave all correct answers are those at very high ability levels. So MTF items are very useful to identify these students.

ENERPLNT

MTF part: Which of the following are sources of energy for plants? Circle yes or no for each of the following:

a). Water	Yes / No
b). **Sunlight**	Yes / No
c). Air	Yes / No
d). Nutrients in soil	Yes / No
e). They make their own energy	Yes / No

Paired CR part: Explain what you think is energy for plants.

Table 5.6 Percentages of each response string of the ENERPLNT item and the average ability estimates for each response string

ENERPLNT (n=508)							
Percent-ages	Water	Sunlight	Air	Nutrient	Own energy	CR level average	Average ability estimates
4%	**No**	**Yes**	**No**	**No**	**No**	**3.9**	**.906**
28%	Yes	Yes	Yes	Yes	No	1.9	-0.29
25%	Yes	Yes	Yes	Yes	Yes	1.9	0.13
14%	Yes	Yes	No	Yes	No	1.8	-0.09
9%	Yes	Yes	No	Yes	Yes	2.1	-0.38
Others (20%)	…	…	…	…	…	…	…

Note: The correct response string is in bold.

Second, the patterns of T or F response strings are not clearly associated with students' paired CR levels. Take the ENERPLNT item as an example, the average CR level of the students who selected "Y" to the first four options and "N" for the last option was 1.9. The average CR level of the students who selected "Y" to all five options was also 1.9. There was no clear pattern in terms of how the responses strings associate with the CR levels. Students who gave different response strings were at similar CR levels. The main problem is that students often select both low level and high-level options because they do not have enough sophisticated understanding to rule out the lower level distracters. In this case, most students selected sunlight as the energy source but they selected the others as the energy source as well.

Data from the ANIMWINTER item give another good example for this problem. This item asks what happens to the fat that the animal lost during hibernation. The correct answer is T for the second option, "the fat was turned into water and gases that the animal breathed out" and F for the other options. Some options such as "turned into waste in the digestive system and left the body as poop" are designed as a lower level distracter. Students who selected the correct

answer also select T to those lower level distracters. However, in the CR part, when the question is asked in an open-ended way and there are no low level distracters, students are more likely to be at higher levels. For example, many students selected T to the high and low level options, but their explanations are mainly on fat and heat/energy conversion, which are at level 3.

ANIMWINTER:

MTF part: During winter, many animals have problems finding food and may hibernate (sleep through the winter). These animals lose weight by spring. What do you think happens to the fat that the animal lost during hibernation? Circle True OR False for each possibility.

True False The fat was turned into heat to keep their bodies warm during the winter

True False The fat was turned into water and gases that the animal breathed out

True False The fat was turned into waste in the digestive system and left the body as poop.

True False The fat was turned into other materials in the body that don't weigh as much.

True False The fat was used up in the animal's body and disappeared.

CR part: Think about your responses above. Please explain as much as you can about what happens to the fat in the animal's body during hibernation.

Third, since students often choose both low and high-level options, and it's hard to judge their understanding level based on their T or F response strings, therefore, another approach was applied to analyze the correlation between students' T or F response strings and their paired CR levels. Students' MTF responses were recoded into scores according to the number of correct choices students made for the T or F questions. The IRT analysis suggested students' MTF scores based on their number of correct choices generally fit the 1PCM model and the step thresholds of the MTF items were in the correct order. This suggests MTF items can also measure students who are in the middle or low ability range using students' number of correct

choices as their MTF scores.

One MTF item (THINGTREE) showed serious misfit and three MTF items showed slight misfit (AIREVENT, INFANT, POTATO). The THINGTREE item below asks students to identify things that a tree needs in order to grow from a list: sunlight, soil, water and air.

THINGTREE

A small oak tree was planted in a meadow. After 20 years, it has grown into a big tree, weighing 500 kg more than when it was planted. Do you think the tree will need any of the following things to grow and gain weight? Please circle Yes or No and explain your choice. If you circled yes, explain how the tree uses it. What happens to it inside the tree?

Sunlight	**YES**	NO
Soil	**YES**	NO
Water	**YES**	NO
Air	**YES**	NO

Most of the students selected that all four things were needed regardless of their ability level. For example, a student who selected Yes to all four options was clearly at high ability level. He/she provided a level 4 response as the one follows:

During photosynthesis, chloroplasts absorb and use light energy. CO_2 and H_2O combined into organic materials and release O_2. Light energy is converted into chemical energy. The sugar produced by photosynthesis is converted to starch, which involves in the synthesis of amino acid, protein, and lipid. So the weight of tree increases.

Meanwhile, another student who selected T to all four options was at low ability level. He/she gave a level 1 response:

It needs sunlight so it can grow. Without soil the tree won't grow healthy and strong. It always needs water so it can grow just like it needs sunlight. Without the air it won't even be able to grow.

These two students were at significantly different ability levels. However, they got the same number of correct answers. The number of correct choice is not a good measure of students' ability. So this MTF question is not well designed. The other three MTF items that show misfit have similar problems and need to be reviewed to include discriminative options.

Fourth, some of the "T" or "F" questions work better to differentiate students than others. Take the ENERPLNT item as an example, the "water", "air", "nutrients" options are most effective to detect students' differences. Table 5.7 below shows the percentages of students at each CR level for two groups of students: the group who selected Y and the group who selected "No" to the question. For three options, water, air and nutrient, students who circled "No" were more likely to be at higher CR levels than those who selected Y. The *t*-test indicated for these three options, there was significant group difference in terms of students' CR levels between the students who selected Y and who selected N. But for the other options (b. sunlight; e. plant make their own energy), the differences between the groups who selected Y and who selected N were not significant. This means that these two options are less effective to differentiate students. The *t*-test results are in Table 5.8.

Table 5.7 Percentages of students at each CR level for students who selected Y and those who selected N to each T or F question

Level	WATER (%)		LIGHT (%)		AIR (%)		NUTRIENT (%)		OWN ENERGY (%)	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Level 1	34	9	31	36	34	26	32	25	32	30
Level 2	44	25	42	27	45	35	46	22	33	48
Level 3	22	38	23	36	20	29	22	32	34	15
Level 4	0	28	4	0	0	9	0	21	0	6

Table 5.8 Compare the average CR level of two groups of students: students who selected Y and those who selected N

	WATER		LIGHT		AIR		NUTRIENT		OWN ENERGY	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Mean CR	1.87	2.85	1.99	2.00	1.85	2.21	1.89	2.48	2.02	1.97
n	455	65	499	22	319	198	431	87	233	279
Sig	.012		.394		.000		.000		.120	

Excluding less discriminative options and including more discriminative options will improve the quality of the MTF items to some extent. Table A.6 in the Appendix is a list of what are the effective sub-questions for each MTF item according to the *t*-statistics. After recalculating students' MTF scores using only their responses to the effective sub T or F questions, the item discrimination of most MTF items improves. However, since some items don't have any discriminative option or only have some slightly effective options, by excluding less effective sub T or F question is not sufficient to improve the quality of MTF items.

In summary, the set of T or F questions of the MTF questions are very useful to identify students at very high achievement levels. Those students often make correct choices to all T or F

questions. The number of correct T or F choices students made can also measure students who are in the middle or low ability range to some extent. For the MTF items that do not fit well with the PCM model or have low discrimination, they need to be redesigned to have a better combination of T or F questions since some T or F options are more effective than others in terms of differentiating students. The design of MTF options needs to be informed by more research to include the most efficient indicators and combinations of indicators to differentiate students.

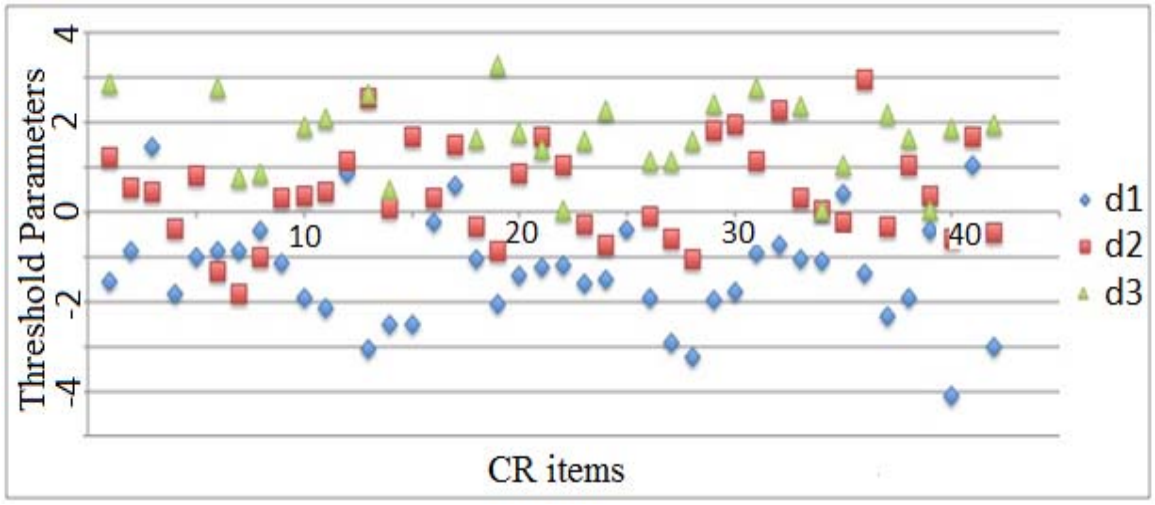
5.5.3 CR items

The CR items should be able to elicit responses at multiple levels of a learning progression. If the item can only elicit responses at particular levels, we need to know what levels that the item is discriminative at and then use the item appropriately. For example, if an item can only distinguish low-level responses, then the item is most appropriate to be used for low ability students. To find out the levels that each CR item is discriminative at, the item threshold parameters are analyzed. If the threshold parameters are not in the correct order, it may indicate that the item is not discriminative at certain levels or the item classify students inaccurately. Knowing this will allow us to either make modifications to make the CR item be discriminative for a wider range of levels or to use it more appropriately.

The result indicates that most of the CR items are effective for differentiating students among levels. Figure 5.5 shows the item threshold parameters of all the CR items. Among all 42 CR items/questions (25 CR items and the 17 CR questions from the OMC+CR items and the MTF+CR items), the item threshold parameters of 33 items are in the correct order. The threshold parameters (d_1, d_2, d_3) of 9 items are very close to each other or not in the correct order, which suggests that there are too many or too few responses at a particular level

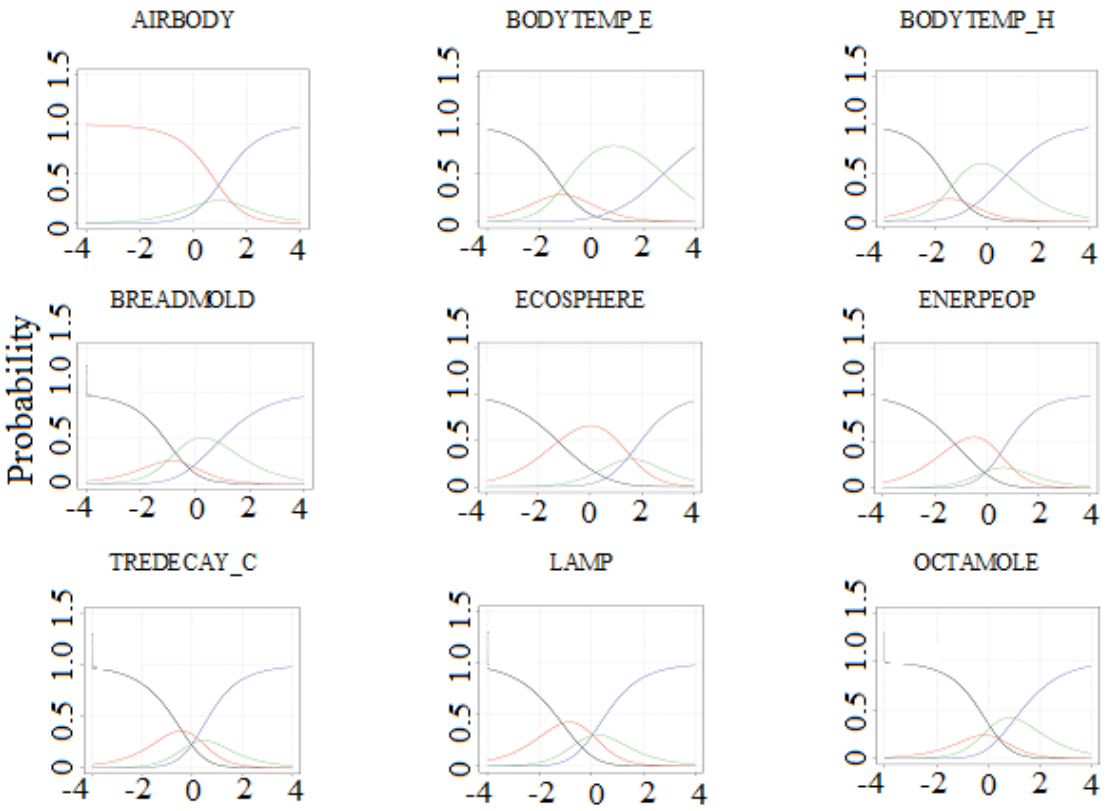
comparing to the proportion of responses at that level from other items. So the item does not accurately classify students at these particular levels.

Figure 5.5 Item threshold parameters of all the CR items



The threshold parameters of the following nine items are not in the correct order: AIRBODY, BODYTEMP_E, BODYTEMP_H, BREADMOLD, ECOSPHERE and ENERPEOP, TREEDECAY_C, LAMP and OCTAMOLE. Figure 5.6 below shows the item characteristic curves by categories for these nine items.

Figure 5.6 The characteristic curves by category of nine CR items



Note: the black line represents the probability of getting score 1; red line for score 2, green line for score 3 and blue line for score 4.

The AIRBODY item has 87% of the students at level 2, 9% of the students at level 3 and 4% of the students at level 4. The majority of the students are at level 2. Though this item was administered at middle and high school level, it did not elicit many high level responses. Students gave very brief responses such as "lung" or "blood" to sub-question A. They explained the process as breathing in oxygen and breathing out carbon dioxide, which seemed less than they might actually know. Though some sub-questions of this item ask students to explain "how", these questions still did not work well to elicit detailed responses at cellular or atomic-molecular scale. These questions may need to be revised to elicit higher-level responses. For

example, subquestion C can be revised as "where does the carbon in carbon dioxide that people breathe out comes from?"

AIRNBODY

Humans get oxygen from the air they breathe in, and they breathe out carbon dioxide.

a. Where in the body does the oxygen get used?

b. How does the oxygen get used?

c. How is the carbon dioxide produced in the body?

d. Does breathing help your body use energy? If so, how?

Four items, BODYTEMP_E, BODYTEMP_H, BREADMOLD, and OCTAMOLE have much fewer level 2 responses but more level 3 responses compared to the other items, so the step difficulties are not in the correct order. These items do not classify level 2 and 3 clearly. Both the BODYTEMP_E and the BODYTEMP_H items ask students to identify the energy source of human body heat. Many students could identify "food" as the energy source of human body heat but did not provide further explanations in terms of how food was used in human body to provide energy, and these responses were coded as Level 3. So some students who might be actually at level 2 got level 3 for these two items. The BREADMOLD and the OCTMOLE items are mostly discriminative at the higher levels such as level 3 and 4, so there are relatively fewer level 2 responses.

The other four items, ECOSPHERE, ENERPEOP, TREDECAY and LAMP did not classify students between level 2, 3 and 4 precisely. There were relatively fewer level 3 responses and most responses were scored either as level 2 or level 4. All these four items assess students' understanding of energy transformations. The problem with these items might due to the scoring rubrics do not clearly distinguish the adjacent levels or due to coding mistakes. The

general level descriptions of students' understanding of energy transformation at level 2, 3 and 4 are as follows:

Level 4: Students clearly distinguish matter from energy and knows energy degradation.

Level 3: Students do not consistently distinguish matter from energy, mixing forms of matter with forms of energy

Level 2: Students do not clearly distinguish energy from other enablers, so they may identify sunlight or other enablers (or nothing) as inputs and heat or other products (or nothing) as outputs.

Based on the level description, both level 2 and level 3 students confuse matter and energy. They cannot trace energy consistently. The distinction between level 2 and level 3 is not clear in the level description. The level description does not emphasize what a level 3 student can do but a level 2 student cannot. So the responses were coded as either level 2 or level 4.

The results of CR items indicate most of the CR items are effective to differentiate students among levels. However, the threshold parameters (d_1, d_2, d_3) of some items are very close to each other or not in the right order, which suggests that the items do not accurately classify students at some levels. The items and the scoring for these items need to be reviewed to classify students more accurately.

When the Environmental Literacy project was developing the scoring rubric, the levels that the item was not discriminative at on conceptual grounds were specified in the scoring rubric. Then the responses of this item were classified into the discriminative levels only. As discussed previously in this section, the statistical analyses suggested additional items that had too few or too many responses at a particular level. The statistical analyses can help to identify additional items that might not be discriminative at a particular level. Then all the responses

coded at that level should be recoded into the adjacent levels to reduce the measurement error. Sometimes, there are too few or too many responses at a particular level due to the ambiguity of the scoring rubric or due to coding mistakes. Then the rubric needs to be revised and the coding mistakes need to be corrected.

In summary, this chapter discusses how to design a learning progression-based test using items in the OMC, MTF and CR format. Though the dimensionality analysis suggests that 3PCM fits better than the 1PCM, there are moderate correlations among these three latent dimensions and strong correlations between the abilities estimated using the 1PCM and the abilities in each dimension estimated using the 3PCM. These high correlations indicate that one dimension is sufficient to describe student's ability. The additional dimensions may only capture subtle differences among item format.

A unidimensional model is supported by the cognitive theories underlie the assessment design. The results show most items fit well with the 1PCM model. This indicates in general, the learning progression framework and the unidimensional assumption are supported. The levels defined by the learning progression framework reconcile with the ability defined by the IRT scale. This is evidence that the learning progression framework is valid.

This chapter also discusses how to design a learning progression-based test to accurately locate students' understanding within the learning progression achievement levels. First, the boundaries between levels are defined on the IRT scale and then items in multiple formats are selected to achieve the information criterion at all the defined boundaries. This ensures the accuracy of the classification. One important design criterion for the learning progression-based item is that ideally, students should be at the same level across items. So the items threshold parameters should be similar. This chapter provides the calibrated item parameters to inform the

future development of carbon cycle items. Finally, how to design OMC and MTF items to predict students' CR levels more accurately and what are the valid ranges of the CR items are discussed.

Chapter 6 Design of a Test to Assess a Particular Process or Practice

6.1 Research purpose and procedure

The carbon cycle assessment is designed to assess students' understanding of six carbon-transforming processes: animal functioning, animal growth, plant growth, decomposition, combustion, and cross-process events. The assessment also assesses five practices: macroscopic, mass/gases/amount, energy/causes, microscopic, and large-scale practices (See section 2.2.3 for a detailed description of these processes and practices). These processes are the key carbon-transforming processes in socio-ecological systems and the practices help to reason about the carbon-transforming processes in complex systems. Appendix A includes the information about the practice and the process that each item measures.

In some cases, we want to know students' understanding of a particular process or practice. This can provide teachers information about students' performances on the particular process or practice so that they can adjust their teaching. In order to design such a test, the dimensionality of the assessment data is investigated first to see whether items of different processes or practices assess the same ability or different abilities and what are the correlations among the constructs. If items of different processes or practices measure students' ability in different dimensions, then to assess students' understanding of a particular process or practice, only items of that process or practice should be used.

The following steps were followed to investigate how to design a test to assess a particular process or practice. First, dimensionality analysis was conducted to see whether items of different processes/practices assessed the same latent construct. In the current item pool, there are 42 CR items. The hypothesis made by the Environmental Literacy project is that students use the same ability to respond to these items that assess different processes and practices. In other

words, students' understandings of different processes are highly correlated, which can be considered as the same type of ability, so as their abilities for different practices. Dimensionality analysis can test this hypothesis.

To evaluate the dimensionality of the item response data, the unidimensional PCM and the multidimensional PCMs were applied to the data and the model fits were compared. Both multidimensional within-item and between-item models were used to account for the structure of the assessment. The dimensions were defined in terms of carbon transforming processes and/or the scientific practices. In total, four models were used to fit the data to explore the correlations among the constructs. The model parameters were estimated using the ConQuest software.

Figure 6.1 is a graphical representation of these four models. Model 1 is the unidimensional PCM. The assumption is that there is one latent general construct that determine students' performances on all items. Model 2 is a between item multidimensional PCM, and the dimensions are defined in terms of processes. Model 3 is a between-item multidimensional PCM and the dimensions are defined in terms of process. The last model, Model 4, is a within-item multidimensional PCM and the dimensions are defined by both process and practice. Each item is associated with one process and one practice.

Figure 6.1 Graphical representation of the unidimensional and multidimensional models

Model 1. Unidimensional PCM

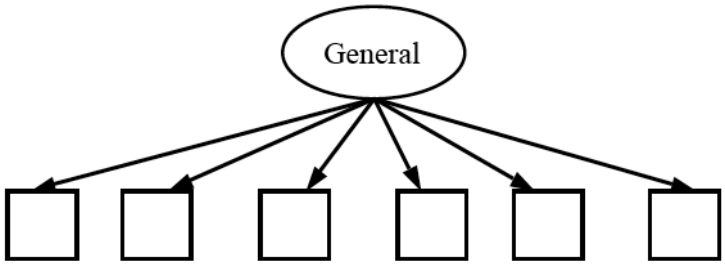
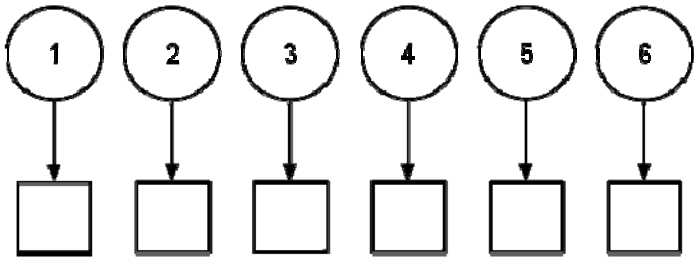
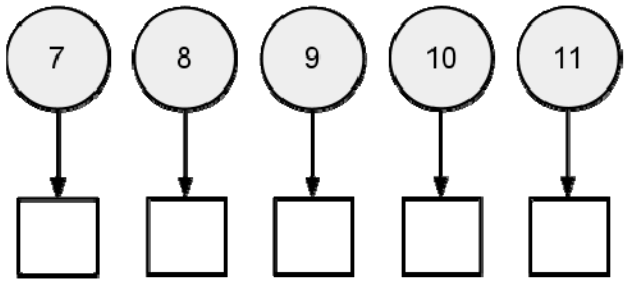


Figure 6.1 (Cont'd)

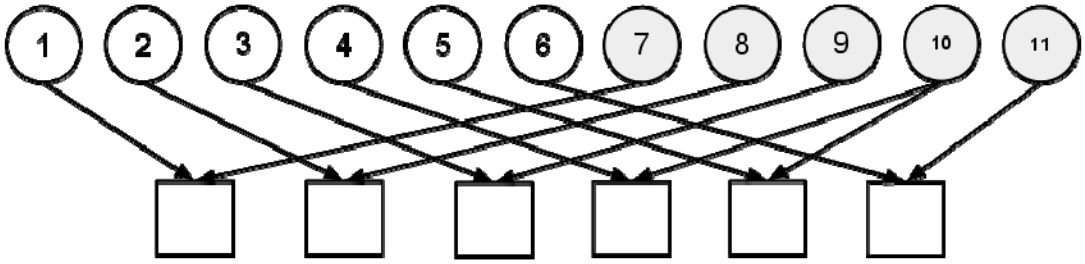
Model 2. Multidimensional between item model (processes as dimensions)



Model 3. Multidimensional between-item model (practices as dimensions)



Model 4. Multidimensional within-item model with processes and practices as dimensions



The names of the dimensions are listed below:

Process	Practice
Dimension 1—Plant growth	Dimension 7—Macroscopic
Dimension 2—Animal growth	Dimension 8—Mass/gases/amount
Dimension 3—Animal function	Dimension 9—Energy/causes
Dimension 4—Combustion	Dimension 10—Microscopic
Dimension 5—Decomposition	Dimension 11—Large-scale practices
Dimension 6—Cross-processes	

Second, the fit of these four models were compared. Chi-square goodness of fit tests were performed on the difference of the deviance of each model. Meanwhile, the model fit was compared in terms of how well the model was supported by the cognitive theories. Then item parameters were calibrated by applying the most appropriate model to fit the data.

Third, items were selected to assess a particular process/practice based on the dimensionality result. If the dimensionality is greater than one in terms of process, then only items of a particular process or practice should be selected to assess students' understanding of that process or practice. The selected items need to yield the amount of desired test information at the boundaries for the process or the practice being measured. This item selection and test design procedures were similar to what were conducted in Chapter 5.

6.2 Dimensionality in terms of process or practice

The results from the chi-square goodness of fit tests are given in Table 6.1. From the results, one can see that Model 2 is the most appropriate model. We can have the following observations based on Table 6.1.

- Model 3 and Model 1: There is no significant difference between Model 1 and Model 3. The unidimensional model, Model 1, fits the data as well as Model 3.
- Model 2 and Model 1: Model 2 fits the data significantly better than Model 1.
- Model 4 and Model 1: Model 4 fits the data significantly better than Model 1.
- Model 2 and Model 4: A further comparison between Model 2 and Model 4 shows that Model 2 explains the data as well as Model 4. There is no significant difference between Model 2 and Model 4. But since Model 2 has fewer parameters, it is more parsimonious than Model 4. Therefore, Model 2 is the best model among these four models.

Table 6.1 Goodness of fit test among four models

	Difference between deviance	Difference between # of parameters	p-value
chi-square test (Model 2 vs. 1)	94.217	20	<.001
chi-square test (Model 3 vs. 1)	5.637	14	0.975
chi-square test (Model 4 vs. 1)	114.179	65	<.001
chi-square test (Model 2 vs. 4)	19.962	45	0.999

The dimensionality analyses suggest that the multidimensional between item model using processes as dimensions fit the data best. This suggests that the data are multidimensional in terms of the processes but not in terms of the practices. In general, the increase in model fit by applying multidimensional model is not big. There are moderate to strong correlations among the process dimensions and strong correlations among the practice dimensions, which suggest that different practices can be considered as the same construct. Table 6.2 and Table 6.3 are the correlations among the process dimensions and the correlations among the practice dimensions based on Model 2 and Model 3 results respectively.

Table 6.2 Correlations among process dimensions

	D1 Plant growth	D2 Animal growth	D3 Animal function	D4 Combust ion	D5 Decom-position	D6 Cross-processes
D1 Plant growth						
D2 Animal growth	0.675					
D3 Animal function	0.870	0.732				
D4 Combustion	0.769	0.857	0.799			
D5 Decomposition	0.859	0.633	0.808	0.717		
D6 Cross-process	0.854	0.680	0.857	0.754	0.845	
Variance	1.985	1.846	0.910	2.184	1.763	1.067

Table 6.3 Correlations among practice dimensions

	D7 Macroscopic	D8 Mass/gases / amount	D9 Energy/causes	D10 Microscopic	D11 Large-scale
D7 Macroscopic					
D8 Mass/gases/amount	0.934				
D9 Energy/causes	0.879	0.886			
D10 Microscopic	0.894	0.907	0.869		
D11 Large-scale	0.789	0.756	0.755	0.798	
Variance	2.179	1.666	0.665	1.046	1.594

In general, there are moderate to high correlations among process and practice dimensions. This suggests that a single latent construct explains most of the variance in students' responses. One possible reason for the high correlations among the practice and process dimensions is that the student sample does not show much variation in these dimensions. Even if the items are sensitive to the difference between dimensions, the item response data do not show strong multidimensionality.

The high correlations among the practice dimensions may also suggest that the understanding of the different practices is in fact strongly psychologically linked. For instance, the ability to explain at microscopic scale is associated with the ability to explain changes at the macroscopic scale. The process dimensions are highly correlated but the correlations are not as strong as those among the practice dimensions. The same student's responses to items of different processes are different to some extent.

In this study, it is hard to tell whether the high correlations among the process or practice dimensions should be attributed to the psychological links among the dimensions or to the lack of variation of the sample in these dimensions. This can be tested in the future with a different sample that does have variance in these dimensions (e.g. a group of students who have learned photosynthesis and another group of students who do not). Since students' understanding of

different processes might be different, the design of a test to assess a particular process is discussed in the following section.

6.3 Design a test to assess a particular process

There are 42 CR items in total including: six plant growth items, three animal growth items, ten animal function items, six combustion items, five decomposition items and twelve cross-process items. To assess a particular process, only items of that process should be selected. Figure 6.2 Figure 6.2 shows for each process, the information collected by using all the items of that process in the current item pool. There are more cross-processes items than items of any other processes, so it is not surprising that the information of the cross-process items is higher than the information of the items of other processes. But from this graph, we can see that except for the cross-processes, the test information of the other processes cannot reach to 5 across the boundaries even if all the items are selected.

Figure 6.2 Information of all items of each process

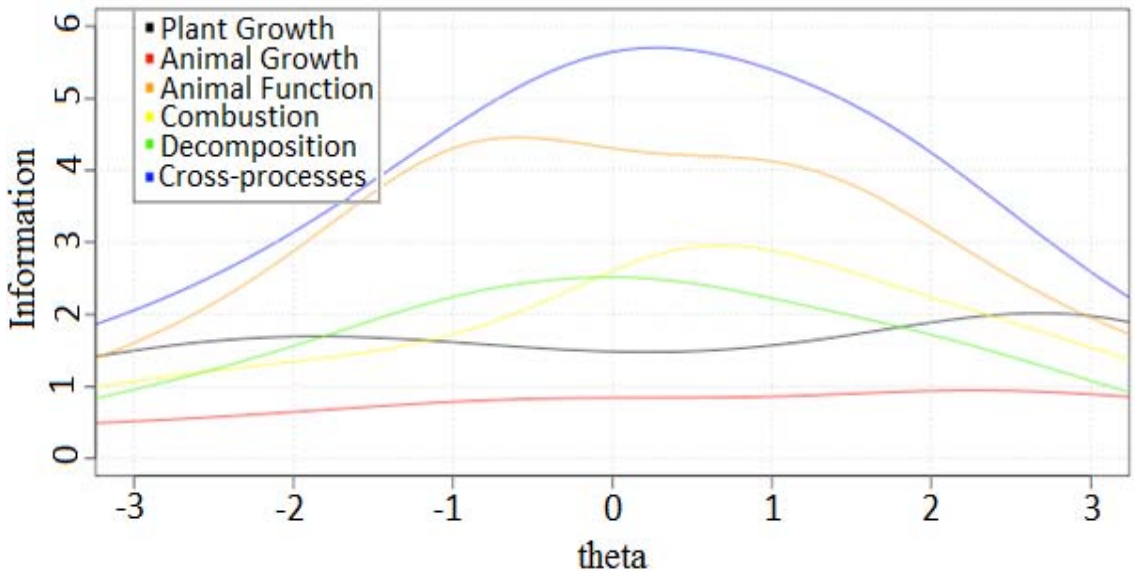


Table 6.4 below shows the item difficulty parameters and threshold parameters derived from the Model 2 results. The difficulty parameter (b) indicates the relative difficulty of the item among all items of the same process. The threshold parameters are the boundaries between levels. Since the design criterion of learning progression-based item is to have similar thresholds across items, the items with boundaries that are away from other items need to be reviewed. Additional items need to be developed to form a test that can reach the information criterion at the defined boundaries. The procedures are similar to those discussed in Chapter 5.

Table 6.4 The item parameters of each process

Process	Item label	b	d_1	d_2	d_3
Process 1: Plant Growth	ACRON	0.534	-2.560	0.958	3.203
	CARPATH	0.435	-4.154	2.479	2.979
	ENERPLNT	-0.564	-2.727	-0.728	1.764
	GRANJOHN_P	-0.305	-2.496	1.886	
	PLANGAS	0.568	-1.924	3.060	
	THINTREE	-0.668	-3.576	-0.801	2.373
Process 2: Animal Growth	CARBODY	0.582	0.177	0.988	
	EATAPPLE	-0.718	-3.473	-1.683	3.003
	INFANT	0.135	-1.679	1.950	
Process 3: Animal Function	AIRBODY	1.329	1.710	0.949	
	ANIMWINT	-1.125	-2.001	-0.249	
	BODYTEMP_E	0.464	-0.909	-1.178	3.481
	BODYTEMP_H	-0.434	-0.860	-1.622	1.179
	EATBRTHE	0.680	-1.345	1.141	2.244
	ENERPEOP	0.171	-1.160	1.278	0.393
	GIRLRUN_AB	0.315	-1.664	-0.648	3.255
	GIRLRUN_C	-0.283	-0.283		
	GLUGRAPE	-0.738	-3.299	-0.905	1.989
	WTLOSS	-0.379	-3.140	-0.305	2.308
Process 4: Combustion	BRNMATCH_E	-1.079	-2.232	0.074	
	BRNMATCH_MA	-0.025	-2.850	0.192	2.584
	BRNMATCH_MB	0.018	-3.052	0.312	2.795
	CARGAS	-1.116	-3.762	-0.359	0.774
	OCTAMOLE	0.516	0.516	-0.507	1.540
	WAXBURN	1.685	0.921	2.449	
Process 5: Decomposition	APPLEROT	-0.485	-1.719	0.748	
	BREADMOLD	-0.243	-0.814	-1.037	1.122
	GRANJOHN_D	0.623	-2.440	1.920	2.388
	TREEDECAY_AB	0.148	-2.581	1.021	2.003
	TREEDECAY_C	-0.042	-0.850	0.390	0.335
Process 6: Cross-process	AIREVENT	-0.459	-1.358	0.440	
	CONNLIFE	-0.685	-2.996	1.626	
	CUTTREE	-0.108	-0.570	0.355	
	DEERWOLF	1.276	0.548	2.003	
	DIFEVENT	0.036	-1.383	-0.373	1.865
	ECOSPHERE	0.698	-1.472	1.765	1.799
	GLOBWARM_M	-0.401	-2.387	-0.253	1.436
	GLOBWARM_H	-0.814	-3.288	-0.633	1.480
	GROWTH	1.152	-1.221	1.201	3.477
	KLGSEASON	0.762	-1.166	0.491	2.963
	LAMP	-0.354	-1.405	0.063	0.280
	TROPRAIN	-1.103	-4.628	-0.749	2.069

The column headers are:

b is the item difficulty parameter

d_1 is the first item threshold parameter. It is the cutting point on the ability scale between score 1 and 2

d_2 is the second item threshold parameter. It is the cutting point on the ability scale between score 2 and 3

d_3 is the third item threshold parameter. It's the cutting point on the ability scale between score 3 and 4

In short, science assessments often require numerous types of knowledge and skills, which are likely to be multidimensional. To design a learning progression-based science assessment, we need to understand whether the assessment measures a single construct or several constructs and how items are associated with the constructs being measured. Only items that assess the construct of interest should be used in the test. To scrutinize the dimensionalities among different processes/practices, we need data from student samples that show variances among the process/practice dimensions to investigate the links among the process/practice dimensions.

The current assessment does not have enough items to assess a particular process accurately. Two things need to be done to design a test for a particular process. First, according to the design criterion of learning progression-based items, items that have thresholds away from the other items of the same process need to be examined. Second, items that have thresholds close to the estimated boundaries need to be developed to form a test that can reach the test

information criterion at the boundaries. The current items can be used as a reference to develop new items.

Chapter 7 Item Characteristics

7.1 Research purpose and procedures

Items were evaluated quantitatively in the Chapter 5 and 6 in terms of item fit indices, discrimination indices, item difficulty and threshold parameters. According to the quantitative evaluation, some items are better than others. These items are more discriminative, fit well with the model, and have thresholds that close to the estimated boundaries. In this chapter, item characteristics are analyzed to find the characteristics that are related to good item statistics. These characteristics can be used as guidelines to design learning progression-based items in future.

The characteristics of the items are coded in terms of the following aspects:

A. Whether the item includes picture(s) or not

B. The number of sub-questions included in the item

C. The familiarity of the example(s) to students

D. The process that the item assesses

E. The practice that the item assesses

F. The scale of the item (e.g. microscopic item, macroscopic item, large-scale item)

In addition, each item received a qualitative rating from two science education researchers. The qualitative evaluation results are analyzed to provide additional suggestions to write good learning progression-based items.

7.2 Write good learning progression-based items: How item statistics are related to the item characteristics

I investigated how the item characteristics above were related to the item statistics. The results from *t*-test and ANOVA analyses indicated there were no significant group differences in item statistics (e.g. item difficulty and item step thresholds) in terms of the following item characteristics:

A. Whether the item includes picture(s) or not;

B. The number of sub-questions included in the item;

C. The familiarity of the example(s) to students;

D. The process that the item assesses and

E. The practice that the item assesses.

However, the item difficulty and the location of the step thresholds were significantly different among items of different scales. The difficulties of microscopic items were significantly higher than those of the macroscopic items or large-scale items. The mean difficulty of microscopic items was around 0.7 while as the mean difficulty of macroscopic items and that of the large-scale items were both around -0.2. About half of the microscopic items were not discriminative at level 1. For example, the microscopic items such as CARBPATH, CARBBODY, GRANJOHN could not elicit level 1 responses. Students whose understandings were constrained at macroscopic scale were not able to explain the movement of atoms and molecules at all. One fourth of the macroscopic items were not discriminative at the highest level—level 4.

Since there are no general connections between the superficial item characteristics listed above and the item statistics except the scale of the item, in the following section, more detailed item characteristics analysis is conducted to find out some rules to write items that will result in good item statistics.

7.3 Write good learning progression-based items: How item statistics are related to suggestions from qualitative evaluation

The following suggestions might help the items to perform better according to both feedback provided by a group of science education researchers and the quantitative results.

7.3.1 CR items

First, some items are not discriminative at some achievement levels. It's important to notice the levels that the item is discriminative at and classify responses into those levels only. This will improve the fit statistics and make the thresholds in the correct order. In Section 5.5.3, nine CR items are identified that do not have item threshold parameters in the correct order. This may due to the item is not discriminative at a particular level. If so, collapsing score categories will help to make the thresholds in the correct order.

Second, as mentioned in section 7.2, many macroscopic items are not valid for level 4 and microscopic items are often not valid for level 1. Depending on the students who will take the item, the same question can be asked at different scales to measure students more precisely. For example, the OCTAMOLE item and the CARGAS item both assess the concept of the combustion of gasoline. These two items basically assess the same concept. However, the OCTAMOLE item is proposed at microscopic scale and the CARGAS item is proposed at macroscopic scale. The difficulty of the OCTMOLE item is 0.5 and the difficulty of the CARGAS item is only -0.8. The item information curves are in the graph below. The OCTAMOLE item has information peaked in the high ability range and the CARGAS has relatively flat information curve over a wider range of abilities. Therefore, depending on the students who will take the item, the same question can be asked in different scales to measure students more precisely.

OCTAMOLE

Gasoline is mostly a mixture of hydrocarbons such as octane: C_8H_{18}. Decide and circle whether each of the following statements is true (T) or false (F) about what happens to the atoms in a molecule of octane when it burns inside a car.

T F Some of the atoms in the octane are incorporated into carbon dioxide in the air.

T F Some of the atoms in the octane are incorporated into air pollutants such as ozone or nitric oxide.

T F Some of the atoms in the octane are converted into energy that moves the car.

T F Some of the atoms in the octane are burned up and disappear.

T F Some of the atoms in the octane are converted into heat.

T F Some of the atoms in the octane are incorporated into water vapor in the atmosphere.

a. When the gas tank is empty and the car stops, where is the energy that was in the gasoline?

b. What was the original source of energy of gasoline?

c. Is air needed for the car to use the gasoline? If so, how does the air change as the car runs?

CARGAS

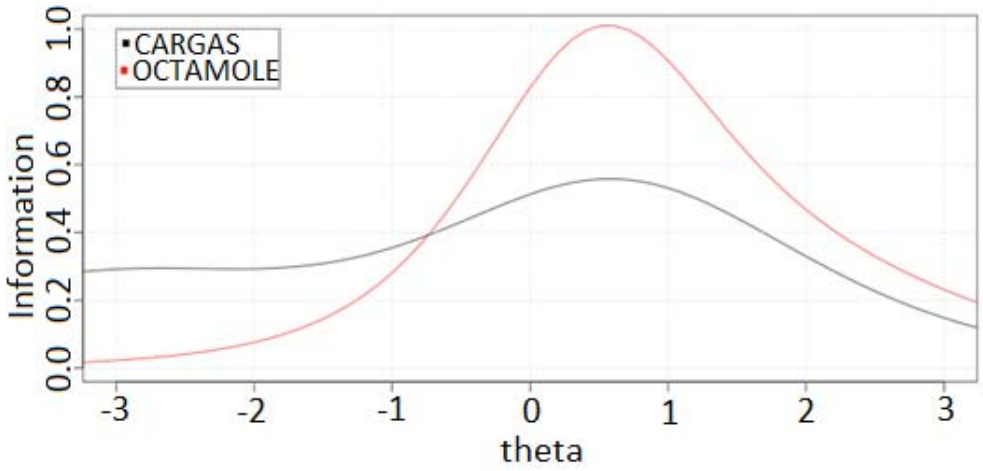
When you are riding in a car, the car uses gasoline to make it run. Eventually the gasoline tank is empty.

a. What happens to the materials the gasoline is made of when the car uses the gasoline?

b. Is air needed for the car to use the gasoline? If so, how does the air change as the car runs?

c. Where does energy come from to make the car run?

Figure 7.1 Item information curve of OCTAMOLE and CARGAS



Third, it is pointed out in Section 7.2 that some of the macroscopic items are discriminative at all four levels while others are not discriminative at level 4 as indicated by their thresholds. Then, the question is what are the item characteristics that make some items work better than others? One feature of the items that are discriminative at all four levels is that these items explicitly require students to trace matter or trace energy. The EATAPPLE item and the GRUGRAPE item below are good examples of this.

EATAPPLE

An apple is eaten by a boy and digested in his body.

a. What happens to the apple when it is digested?

b. Do you think the apple the boy ate can help all parts of his body (like his fingers) to grow? Please circle one: YES NO

If you answered YES, please explain how can an apple that goes to the boy's stomach help his fingers to grow. If you answered NO, please explain how the boy's body makes his fingers grow.

GLUGRAPE

The grape you eat can help you move your body parts such as your legs.

a. Please describe how the substances from the grape provide energy to move your legs. Describe as many intermediate stages and processes as you can.

b. Can the substances of the grape also be involved in helping to keep your body warm? Please explain your answer.

These two questions successfully elicited both low and high level responses. The EATAPPLE item asks students to trace apple through the body and the GLUGRAPE item asks students to trace energy from the grape people eat to energy that help people to move body parts. These questions have a clear focus on matter/energy transformations and encourage students to explain the transformations.

Some other items that measured the same concept were not able to elicit high-level responses. For example, the INFANT item also measured the concept of animal growth. Same as the EATAPPLE item, it was administered to students at all three grade levels. However, the INFANT item mainly elicited level 1 and 2 responses (97%). The question is not as focused as

the EATAPPLE item and it does not explicitly require students to trace matter.

INFANT

Do you think the baby girl will need any of the following things to grow and gain weight? Please circle Yes or No and explain your choice. If you circled yes, explain how the girl's body uses it. What happens to it inside the inside the girl's body?

Sunlight Yes **No**

Water Yes **No**

Air Yes **No**

Food **Yes** No

The response that the same student gave for the INFANT item tended to be at lower levels than the one he/she gave for the EATAPPLE item, though these two items measure the same concept generally. For example, below are the responses that a high school student gave to the INFANT item and the EATAPPLE item respectively:

> INFANT: The girl does not need sunlight. She needs water to stay hydrated and live. The girl needs air to breath and respire. She needs food to grow.
>
> EATAPPLE: The body removes all of the helpful vitamins, nutrients, and/or helpful substances out of the apple and the rest becomes waste. The apple can help all parts of his body to grow. The energy from the apple goes all over the body because the body cells pick of the energy from the apple in the villi and take it to cells all around the body.

The student's response to the INFANT item was at level 2, but his/her response to the EATAPPLE item was at level 3. The EATAPPLE item explicitly requires students to trace matter from the apple to human body parts. This helped to elicit higher level responses.

Fourth, it is better to propose a specific problem than a general problem to elicit

responses focusing on the measured concept. For example, both the KLGSEASON and the CUTTREE item assess students understanding of global warming and how photosynthesis is related to global warming. The KLGSEASON specifically focus on the changes in concentration of CO_2 in the atmosphere, it asks students why the atmospheric carbon dioxide levels decreasing in the summer and fall every year and increasing in the winter and spring. But the CUTTREE item asks students why cutting down trees increases global warming. It requires students to make two connections, one between the CO_2 level and global warming, and the second between trees and CO_2. Some students fail to make one of the connections and cannot provide any related answer to this question. The question would be better if it asks students why cutting down tree will increase the CO_2 concentration.

The item discrimination (the correlation between students' scores of the item and their total scores) of the CUTTREE item is .38 and the discrimination of the KLGSEASON is .54, which is much higher. Both items fit the PCM model. The CUTTREE item is not valid for level four and the thresholds are -0.24 (d_1) and 0.30 (d_2). The KLGSEASON item is valid for all four levels and the thresholds are -1.05 (d_1), 0.30 (d_2), and 2.34 (d_3). The KLGSEASON item did a better job to differentiate students over a wider ability range.

Fifth, some types of CR items cannot elicit detailed explanations so these items do not measure students as precisely as others. For example, when the item gives students several examples and ask students to explain each example, students often do not provide enough details for each example. The CARBPATH item asks students where carbon can be found inside a tree. The item gives students three locations (leaves, wood, and roots) and asks them to judge whether

they can find carbon at each location and how carbon gets there. Students often do not give detailed explanation about how carbon gets to each location. As a result, 96% of the responses are at level 1 or 2, and only around 4% of the responses are at level 3 or 4. This makes the item invalid to measure high-level students.

Similarly, students often do not provide detailed explanations to items that require them to make comparisons, connections or analogies among examples. Their explanations usually focus on the observable similarities or differences between the events and ignore deeper connections between events at atomic-molecular scale. For example, the GROWTH item asks students to think of ways that the plants and animals are different in ways they use water, air and nutrients to grow. For example, a typical student's response is that "plants use water to suck it up with their roots, animals drink it. Plants make air, animals breathe in air. Plants use nutrients to grow, animals eat nutrients." There are not many deep descriptions of what is happening and how food might be different for plants and animals.

In addition, there are some other general rules to follow when writing CR items. For example, the item needs to be scientifically rigorous, irrelevant information should not be provided, words or phrases that may involve construct irrelevant difficulty should not be included in the item stem. Researchers who evaluated the items pointed out some of our items had these problems. For example, the stem of the TROPRAIN item has two words that may produce construct irrelevant difficulty, one is "recycled" and the other is "ecosystem". These words may increase the difficulty for students to understand the item. As a result, this item has a lot of missing responses, which suggests the item is not well received by examinees.

7.3.2 OMC and MTF items

This study only includes 7 OMC and 11 MTF items. It's difficult to draw conclusions based on this small sample of items in terms of how item characteristics are related to the item statistics. Section 5.4.2 discussed how to solve the misfit of OMC and MTF items and how to design better OMC and MTF options were discussed in Section 5.5.1 and 5.5.2.

7.4 Recommendations for writing items in future

Based on the above analyses, the following recommendations can be made for writing good learning progression-based items:

- The scale of the item has an impact to the item difficulty and the discriminative range. When the question is asked at the microscopic scale, it often cannot discriminate lower level students. When the question is asked at the macroscopic scale, it often cannot discriminate high level students. So depending on the target group, the same question can be asked at different scales.
- Since some items are not discriminative at some particular achievement levels. It is important to notice the levels that the item is discriminative at and classify responses into those levels only.
- Items that explicitly require students to trace matter or trace energy are more likely to be discriminative at all achievement levels.
- The items that ask concrete and specific problems usually work better to elicit detailed and focused responses.
- To gather more detailed responses, avoid asking students to explain too many examples in one item.

- The items that ask students to make comparisons, connections, or analogies need to be carefully designed since students will make comparisons, connections or analogies in all possible aspects, which may not address the construct being assessed.
- According to the analyses of the OMC and MTF options in Chapter 5, OMC and MTF options need to be designed carefully.
    - When the OMC options are associated with a restricted range of learning progression levels, the OMC item often under or over predict students' real learning progression levels. So in order to design OMC options that can better predict students' real levels, the OMC options need to represent understanding at multiple levels.
    - The design of MTF options needs to be informed by more research to include the most efficient indicators and combinations of indicators to differentiate students.
    - The OMC and MTF options should be designed based on the ability of the examinees. Options need to be discriminative for the examinees who take the item.
    - OMC and MTF items can be treated as dichotomous items (1- if student choose the best answer or choose all correct answers, 0-otherwise) to provide information about students' achievement level. In general, the recoded MTF items are difficult and the recoded OMC items are easy. So MTF items can be used to classify high level students and OMC items can be used to classify low level students.

Chapter 8 Discussion and Conclusions

8.1 Summary of main findings and implications

The focus of this dissertation is to investigate how to design learning progression-based science assessments to accurately classify students into achievement levels. It investigates 1) how to design OMC, MTF and CR items to classify students among levels, 2) how to design a test for a particular process or practice and 3) what are the item characteristics that support the use of items to classify students among levels. The followings are the main findings from the investigation of these three research foci.

8.1.1. Items in different formats are associated with one main construct but also measure slightly different aspects of the construct

The analysis suggests that OMC, MTF and CR items are associated with one main construct but they may also measure slightly different aspects of the construct. The data from most of the items fit well with the unidimensional model. This suggests a single construct explains most of the variances in students' performances on the carbon cycle assessment. The abilities defined on the IRT scale reconcile with the abilities defined by the learning progression level, which provides evidence of the validity of the learning progression framework and the single construct defined by the four achievement levels. So the unidimensional hypothesis of the carbon cycle learning progression framework is generally supported.

There are additional dimensions in terms of item format that can explain some amount of the variance in student performance. There are moderate correlations among students' abilities in the CR, OMC and MTF dimensions and high correlations among students' unidimensional and multidimensional ability estimates. So the unidimensional data might be sufficient to describe

the data. But multiple dimensions are necessary when we want to consider the nuisance dimensions for a particular purpose.

This finding can inform the assessment design. Depending on the purpose of the assessment, items in these formats can be reasonably used to measure what is intended to measure. For example, to measure students' general understanding of carbon cycle, items in all these formats can be used. If the assessment focuses on students' abilities to organize, integrate and synthesize their knowledge and their abilities to solve novel problems, then CR format is preferred. In addition, items in each format need to be improved in certain ways.

8.1.2. Improve the quality of the OMC, MTF and CR items

The OMC and MTF questions have some problems indicated by the discrimination indices, item fit indices, and item thresholds. In terms of the OMC items, students' OMC levels can predict the CR level to some extent. Students' OMC levels are significantly correlated with their levels on the paired CR questions. But in some cases, the OMC level over-predict or under-predict the CR level. Often times, OMC levels over-predict CR levels. When all the OMC options represent understanding at low levels, the OMC levels may under-predict students' real achievement levels.

In terms of the MTF items, the set of T or F questions are useful to identify students at very high achievement levels. Those students often make all correct choices to the T or F questions. The number of correct T or F choices students made can indicate the ability level of students who are in the middle or low ability range to some extent. The MTF items that have low discrimination indices or show misfits with the PCM model need to be revised. Some T or F options are more effective than others in terms of differentiating students. Therefore, the design

of the MTF options needs to be informed by more research to include the most efficient indicators and combinations of indicators to differentiate students.

Most of the CR items fit well with the unidimensional model and are effective to differentiate students among levels. The threshold parameters (d_1, d_2, d_3) of a few CR items are very close to each other or not in the correct order. This suggests that these items do not accurately classify students at some levels. This can be improved by adjusting the scoring rubrics, correcting the coding mistakes or recoding the data into levels that the items are discriminative at.

8.1.3. Use items in multiple formats to meet the test information criterion

To detect the difference between two groups of students, I set "5" as information criterion at the boundaries between levels on the ability scale. This amount of test information is sufficient to detect the difference between two groups of students (30 students in each group, with a difference of 0.93 between the group mean ability estimates) at the significance of 0.05 and the power level of 0.8.

To accurately classify students into levels, first, the boundaries between levels on the IRT scale were defined by using the means of the threshold parameters across a set of good items. Then items are selected to achieve high information at these defined boundaries. To reach test information above 5 at the boundaries, we need 16 items from the current item pool. When the thresholds of the items are close to the defined boundaries, only 14 items are needed to reach the same amount of information (above 5) at the defined boundaries. So adjusting rubrics to get similar thresholds will reduce the number of items we needed to reach the same amount of test information at the defined boundaries. Most importantly, one design criterion of learning

progression-based items is that, ideally, students at the same ability level will get the same level across all items. This means that the item thresholds (d_1, d_2 and d_3) should be similar across items. The item threshold parameters of our current items vary a bit. The items and the scoring need to be reviewed to adjust the thresholds to be similar across items.

Our current assessment has about 10-12 items on each form, which is not too long to be administered in one science class. If the rubrics can be adjusted to make the items classify students more consistently and if the items can be revised to be more discriminative, then around 14 items on one test form will be sufficient to detect the difference between student groups. The analyses also suggest other ways to reduce the test administration time and scoring effort, some dichotomously scored items that have difficulties at the boundaries can be used. For example, using 14 dichotomous items (around 4~5 items with difficulty at each boundary) and 3 polytomous items that have thresholds at the defined boundaries will also reach the information criterion.

8.1.4. Design a test to assess a particular process or practice

To design a learning progression-based science assessment, we need to understand whether the assessment measures a single construct or several constructs and how items are associated with the constructs being measured. This study examines whether different carbon transforming processes and different scientific practices are associated with a single latent construct or different constructs. The results show there is multidimensionality in terms of process but not practice. In general, the correlations among process/practice dimensions are moderate to strong. It is not clear whether the high correlations are due to the inherent links among processes/practices or due to the student sample does not show much variation in these

process/practice dimensions. Future data are needed to examine the dimensionality in terms of process/practice in more detail.

8.1.5. Implications from the item characteristics analysis

Based on item characteristics analysis, recommendations are made at the end of Chapter 7 in terms of how to write items to achieve better item statistics. These can provide guidelines for item writers to write learning progression-based items besides the general item writing guidelines and tips that can be found in literature.

Finally, the results from this study can inspire another iteration of assessment design, in which we refine the learning progression framework, modify existing items, develop new assessment tasks, collect more data and analyze the data with an appropriate statistical model to understand students' learning progression of carbon cycle.

8.2 Discussion of the results

8.2.1. Items in different formats

1) OMC format

Students' OMC levels can predict their CR level to some extent. However, OMC level often over-predict the CR levels. The hypothesis for the discrepancy between students' OMC levels and their CR levels is that students perform better to name the input and output of carbon transforming processes, which is assessed by most OMC questions, than to explain what happens during the processes which is assessed by most CR questions. When the OMC options mostly represent understanding at low levels, the OMC levels will under-predict students' real achievement level. These over and under-predictions need to be noticed when using OMC items. Options at multiple levels need to be developed to reduce cases of over or under predictions.

Since OMC options associated with a more restricted range of learning progression levels compare to CR items, the OMC items do not measure students as precisely as the CR items at the extremes of the ability distribution. This result is consistent with some previous studies mentioned in the literature review. For example, Lee, Liu & Linn (2011) found that compared to MC items, CR items discriminate between high and low knowledge integration ability students much more effectively and measure a wider range of knowledge integration levels. Ercikan, Schwarz, Julian, Burket, Weber, and Link (1998) and Wilson and Wang (1995) have similar conclusions from their studies.

Less discrimination of the OMC items at the low end of the ability distribution may result from guessing involved when answering OMC item. And less discrimination of the OMC items at the high end of the ability distribution might due to fewer options designed at the high levels, especially at level 4. It is difficult to write OMC options at higher achievement levels without using "science-y" terminologies. So in order to reduce measurement errors due to the low discrimination of OMC format at the high end, we need to either develop high-level options without using science-y terminologies or use OMC items mainly to measure median and low-level understanding.

2) MTF format

The MTF items work especially well to assess students' commitments to fundamental principles. Since MTF item allow students to select multiple answers, to answer the item correctly, students need not only to identify all the correct answer(s) but also to exclude all the incorrect answer(s). This requires students to have deep understanding of the principles being assessed and apply those principles consistently. The result shows only a small proportion of students with high abilities can correctly answer all the T or F questions.

The number of correct choices students made can indicate their ability level to some extent. Some of the T or F questions distinguish students better than others. More research is needed to find out the most effective options and most efficient combinations of options. When designing the MTF items, the item writers need to design the options to avoid the situation that the students will make the same choices regardless of their ability levels. If students at different ability levels will give different T or F response strings, then the MTF item is effective.

The fixed options in OMC or MTF question restrict students' thinking. Some of the lower level distractors lowered their responses. Especially for the MTF items, since students can chose T for more than one option, they may choose T for both low level options and high level options. But when the same question is asked in an open-ended way, their responses are at higher levels. Thus, some of the MTF questions can be asked in a more open-ended so there will be less influence from the item itself on students' responses. Students might be able to provide more focused and detailed responses then.

3) CR format

Most of the CR items are effective for differentiating students. There are a small number of CR items that do not have item thresholds parameters in the right order. This is a sign that either the item or the scoring rubric is not appropriately designed. So the item or the scoring needs to be reviewed.

As Anderson et al (2007) pointed out; one challenge with developing learning progression grounded items is that it is difficult to write items that provide opportunities for students to respond at multiple levels of a learning progression. The results of this study also suggest some CR items can only elicit responses at particular levels rather than all levels. For

example, most items proposed at microscopic scale are only discriminative for level 2 and above. These items need to be used appropriately so they are discriminative for the examinees.

8.2.2. Assessing a particular process

Our previous study (Mohan, Chen, Baek, Choi, Lee, & Anderson, 2009) showed that students had a similar level of reasoning on different carbon transforming processes. So the assessment and the learning progression framework are essentially unidimensional. This study indicates that there is multidimensionality in terms of carbon transforming processes but not in terms of scientific practices. But the correlations among process/practice dimensions are moderate to strong. It is not clear whether the strong correlations are due to the inherent links among processes/practices or due to the student sample does not show much variation in these process/practice dimensions. Future data are needed to examine the dimensionalities in terms of process/practice in detail.

There were some differences between the assessments used in the previous study and those used in this study. In the previous study, students answered only one or two items of each of those six processes. In this study, each student answered around three to five items of each process but the test included items of fewer processes. So students' ability in each dimension was measured more precisely. Hence, the conclusion based on this study is more reliable.

After knowing more about what are the latent constructs that the assessment measures, and how items are associated with the latent constructs, items can be selected more purposefully to measure a particular construct such as knowledge of a carbon transforming process. This can provide teachers information about students' performances on the particular process so that they can adjust their teaching of a particular unit.

8.3 The broader implications to learning progression-based assessments

Findings from this study can generalize to other research on designing learning progression-based science assessment. Dimensional analysis is a way to provide evidence for the construct validity of the assessment. It is an approach that can be applied to the design of other learning progression-based science assessments. Science assessments often assess various knowledge and skills, for instance, knowledge of different subjects and a variety of skills, such as conceptual understanding and scientific investigation. So science assessments are likely to be sensitive to differences on multiple dimensions. The dimensionality analysis is one way to understand the construct being measured by the assessment and how items are associated with the construct being measured.

Furthermore, this study establishes some typical procedures that can be followed to design other learning progression-based tests. For example, first, the test developers should set a test information criterion depending on the purpose of the test (e.g. detect the difference between individual students, or detect the difference between groups of students). Then, the test developers can use either statistical approaches (e.g. take the mean of the item threshold parameters across a set of items) or standard setting approaches to set the boundaries between levels and select items to reach the information criterion at the boundaries. Third, since for learning progression-based items, ideally, the thresholds should be similar across items, the items with thresholds different from those of the other items should be examined.

In addition, findings about the item formats can inform the future use of these items in other learning progression-based assessments. We know that items in different formats might assess slightly different aspects of the construct. So depending on the goal of the assessment, an appropriate item format(s) should be selected to measure what is intended to measure. This study

also provided suggestions to improve the quality of the OMC, MTF and CR items respectively. These suggestions are based on both statistical analyses and qualitative item characteristic analyses. These suggestions are generalizable to the design of other learning progression-based items in these formats.

8.4 Limitations of this study and future work

Four problems limit the validity or generalizability of the findings from this study and suggest directions for future work.

First, the OMC, MTF and CR items are not completely independent items. Each OMC question is paired with a CR question and each MTF question is paired with a CR question as well. The OMC/MTF question and the CR question share the same item stem. So there is inherent correlation between a student's OMC response and his/her response to the paired CR question. This may inflate the correlations between students' ability measured by the OMC/MTF items and their ability measured by the CR items a little bit. But since the correlations between the abilities in the OMC/MTF dimension and the CR dimension are calculated based on students' responses to all CR items (42 in total) instead of just the paired CR items, the correlations between the OMC/MTF and the CR dimension are not inflated much by the inherent relations between the OMC/MTF questions and their paired CR questions.

Second, the partial credit model is used to fit OMC and MTF responses in this study. However, PCM might not the best model for the OMC and MTF items. Guessing was not accounted for in the model. In addition, for some OMC items, there are two options at the same level. So the PCM model may not be the best choice. Some new models are under development to fit OMC responses. For instance, Briggs & Alonzo (in press) introduce the Attribute Hierarchy Method (AHM; Leighton, Gierl & Hunka, 2004) as a relatively novel approach for modeling

OMC items. Since the new models are still under development and they are mainly suitable for certain types of OMC items, this study did not apply the novel models for the OMC items.

Third, this study mainly focused on the assessment items rather than the students who took the items. The information about the students such as their general science achievement or the science courses they had taken were not considered in this study. Since not much information was known about the student sample, it was not clear whether the unidimensionality in terms of practice was because the practices were psychologically linked or because the student sample did not show much variation in these practice dimensions. In this study, it is difficult to tell whether the unidimensionality is attributed to the former or the latter. In the future, more information about students can be collected. And this can be tested with a different sample that does show variation in the practice dimensions (e.g. a group of students who have learned energy and another group of students who do not).

Finally, this study only involves 7 OMC and 11 MTF items. So the findings about the OMC and MTF formats are based on the data from relatively small numbers of items. These findings need to be verified in the future with data from more OMC and MTF items.

APPENDICES

Appendix A Item list

ACORN [Plant growth, Mass/gases/amount]
A small acorn grows into a large oak tree. Where does most of the weight of the oak tree come from? (Circle the best explanation from the list below).
a. From the natural growth of the tree. (Level 1)
b. From carbon dioxide in the air and water in the soil. (Level 3)
c. From nutrients that the tree absorbs through its roots. (Level 2)
d. From sunlight that the tree uses for food. (Level 1)
Explain why you think that the answer you chose is the best answer.

AIREVENT [Cross-processes, Macroscopic]
The 4 pictures below show 4 events happening. Do you think air is needed for each of the events? Please circle Yes or No and explain your choice.
A. Plant growth B. Girl running C. Burning wood D. Food decay

Events	Does the event need air? (Circle)	If you circled yes, explain how is air used in the event?
A. Plant growth	**Yes** No	
B. Girl running	**Yes** No	
C. Burning wood	**Yes** No	
D. Food decay	**Yes** No	

Do the events that you circled "Yes" use air in similar ways or in different ways? Please explain your answer.

AIRNBODY [Animal function, Macroscopic]
Humans get oxygen from the air they breathe in, and they breathe out carbon dioxide.
a. Where in the body does the oxygen get used?
b. How does the oxygen get used?
c. How is the carbon dioxide produced in the body?
d. Does breathing help your body use energy? If so, how?

ANIMWINTER (E) [Animal function, Macroscopic]
During winter, many animals have problems finding food and may hibernate (sleep through the winter). These animals lose weight by spring. What do you think happens to the fat that the animal lost during hibernation?
Circle True OR False for each possibility.
True **False** The fat was turned into heat to keep their bodies warm during the winter
True False The fat was turned into water and gases that the animal breathed out
True **False** The fat was turned into waste in the digestive system and left the body as poop.
True **False** The fat was turned into other materials in the body that don't weigh as much.
True **False** The fat was used up in the animal's body and disappeared.

Think about your responses above. Please explain as much as you can about what happens to the fat in the animal's body during hibernation.

APPLEROT [Decomposition, Mass/gases/amount]
When an apple is left outside for a long time, it rots.
a. What causes the apple to rot?
b. The weight of the apple decreases as it rots. What do you think happens to the matter or stuff that was once in the apple?
c. Is there energy involved when the apple rots?
Circle one: Yes / No
Please explain your answer.

BODYTEMP (E) [Animal function, Energy/causes]
You are playing outside in a cold winter. You find a stone on the ground. When you pick up the stone, you find that the stone is very cold. Why can people keep warm on a cold day, but stones cannot? Which of the thing(s) from the list below can help to keep people's bodies warm? Please circle YES or NO for each thing in the list below.

a. Water	YES	**NO**
b. Food	**YES**	NO
c. Air	YES	**NO**
d. Exercise	YES	**NO**

Try to write an explanation of how people's bodies stay warm that includes ALL of the things you circles "YES" for in the list above.

BODYTEMP (H) [Animal function, Energy/causes]
Your body produces heat to maintain its normal temperature. Where does the heat mainly come from? Please choose the ONE answer that you think is best.
a. The heat mainly comes from sunlight. (Level 1)
b. The heat mainly comes from the clothes you are wearing. (Level 1)
c. The heat mainly comes from the foods you eat. (Level 2)
d. The heat mainly comes from your body when you are exercising. (Level 1)
Please explain why you think that the answer you chose is better than the others. (If you think some of the other answers are also partially right, please explain that, too.)

BREADMOLD (M, H) [Decomposition, Mass/gases/amount]
A loaf of bread was left inside its plastic bag for two weeks on a balance measuring its mass. Three different kinds of mold grew on it. Assuming that the bread did not dry out, which of the following is a reasonable prediction of the weight of the bread and mold together?
A) The mass has increased, because the mold has grown. (Level 1)
B) The mass remains the same as the mold converts bread into biomass. (Level 2)
C) The mass decreases as the growing mold converts bread into energy. (Level 3)
D) The mass decreases as the mold converts bread into biomass and gases. (Level 4)
Please explain your answer and indicate any important transformations.

BRNMATCH(E) [Combustion, Mass/gases/amount;]
When a match burns, it loses weight and becomes smaller.
a. What does the flame need to keep burning the match?
b. What happens to the materials the match is made of as the match burns?
c. Is air needed for the match to burn? Please circle one: YES NO If you answered YES, please explain how does the air change as the match burns.
d. Where does energy needed for the match to burn come from?

BRNMATCH (M, H) [Combustion, Microscopic, Energy/causes]
When a match burns, the released energy
a. comes mainly from the match. (Level 3)
b. comes mainly from the air. (Level 2)
c. is created by the fire. (Level 1)
d. comes from the energy that you used to strike the match. (Level 2)
e. none of the above. (Level 1)
Please explain your answer.

CARBBODY [Animal growth, Microscopic]
Use the table below to explain where you think that carbon is found inside a person's body and how it gets there.

Location		If you circled yes, explain how the carbon gets to that location. Include molecules in your explanation if you can.
Do people have carbon in their muscles?	**Yes** No	
Do people have carbon in their fat?	**Yes** No	
Do people have carbon in their blood?	**Yes** No	

CARBPATH [Plant growth, Microscopic]
Use the table below to explain where you think that carbon is found inside a tree and how it gets there.

Location	Circle Yes or No	If you circled yes, explain how the carbon gets to that location. Include molecules in your explanation if you can.
Does a tree have carbon in its leaves?	**Yes** No	
Does a tree have carbon in its wood?	**Yes** No	
Does a tree have carbon in its roots?	**Yes** No	

CARGAS [Combustion, Macroscopic]
When you are riding in a car, the car uses gasoline to make it run. Eventually the gasoline tank is empty.
a. What happens to the materials the gasoline is made of when the car uses the gasoline?
b. Is air needed for the car to use the gasoline? If so, how does the air change as the car runs?
c. Where does energy come from to make the car run?

CONNLIFE [Cross-processes, Large-scale practices]
Explain how the following living things are connected with one another:
Grass
Cows
Human beings
Decomposing bacteria

CUTTREE [Cross-processes, Large-scale practices]
Some people say that cutting down trees in a forest will increase global warming. Do you agree?
Circle one: YES NO
Please explain your answer.

DEERWOLF [Cross-processes, Large-scale practices]
A remote island in Lake Superior is uninhabited by humans. The primary mammal populations are white-tailed deer and wolves. The island is left undisturbed for many years. Select the best answer(s) below for what will happen to the average populations of the animals over time.

a. The deer will all die or be killed. (Level 1)
b. The wolves will all die or be killed. (Level 1)
c. On average, there will be a few more deer than wolves. (Level 2)
d. On average, there will be a few more wolves than deer. (Level 1)
e. On average, there will be many more deer than wolves. (Level 3)
f. On average, there will be many more wolves than deer. (Level 1)
g. On average, the populations of each would be about equal. (Level 1)
h. None of the above. My answer would be:
Please explain your answer to what happens to the populations of deer and wolves.

DIFEVENTS [Cross-processes, Energy/causes]
A. Eating a hamburger B. Filling up a car with gasoline C. Watering plants
The pictures above show three things happening.
A science teacher says that pictures “A” and “B” are similar events, but picture “C” is different from “A” and “B”. What reason do you think the science teacher might have for saying that? Explain as much as you can.

EATAPPLE [Animal growth, Microscopic]
An apple is eaten by a boy and digested in his body.
a. What happens to the apple when it is digested?
b. Do you think the apple the boy ate can help all parts of his body (like his fingers) to grow?
Please circle one: YES NO

If you answered YES, please explain how can an apple that goes to the boy's stomach help his fingers to grow. If you answered NO, please explain how the boy's body makes his fingers grow.

EATBRTHE [Animal function, Macroscopic]
Humans must eat and breathe in order to live and grow. Are eating and breathing related to each other? (Circle one)
YES NO
If you circled "Yes" explain how eating and breathing are related. If you circled "No" then explain why they are not related. Give as many details as you can.

ECOSPHERE [Cross-processes, Energy/causes]
NASA scientists invented the EcoSphere – inside a sealed glass container, there are air, water, gravel, and three types of living things – algae, shrimp, and bacteria. Usually, these three living things can stay alive in the container for two or three years until the shrimp become too old to live. The picture above shows an EcoSphere and its contents.
Do you think that the living things need to get energy from outside of the EcoSphere to keep living?
Circle one: YES / NO
If your answer is NO, how can the living things stay alive without getting energy from the outside world? If your answer is YES, what form of energy do they get from outside of the EcoSphere?
Do you think the living things will release energy out of the EcoSphere?
Circle one: YES NO
Please explain your answer.
If you circled YES above, what's the form of energy that is released out of the EcoSphere?

ENERPEOP [Animal function, Energy/causes]
People need energy to live and grow. Which of the following is/are energy source(s) for people? Circle yes or no for each of the following and explain your answers.
a. Water YES **NO**
b. Food **YES** NO
c. Nutrients YES **NO**
d. Exercise YES **NO**
e. Sunlight YES **NO**
g. Carbon dioxide YES **NO** (E, M don't ask for CO_2)
h. Oxygen YES **NO**
Please explain ALL your answers, including why the things you circled "No" for are NOT sources of energy for humans.

ENERPLNT(E,M,H) [Plant growth, Energy/causes]
Which of the following is (are) energy source(s) for plants? Circle yes or no for each of the following.
a. Water YES **NO**
b. Light **YES** NO
c. Air YES **NO**
d. Nutrients in soil YES **NO**

e. Plants make their own energy. YES **NO**
Please explain ALL your answers, including why the things you circled "No" for are NOT sources of energy for plants.

GIRLRUNN [Animal function, AB- Energy/causes, C-Macroscopic]
The following picture shows a girl running.
a. When a girl runs, how does her body make her legs move? Try to list everything that the girl's body needs to make her legs move and explain how her body uses those things.
b. Is food one of the things that the girl's body needs to move her legs?
Circle one: YES NO
If you answered "YES," try to explain how food that goes to the girl's stomach can help her legs to move.
c. Is air needed for her to run? Circle one: YES NO
If you answered "YES", explain how the air that goes into her lung helps her run? Is the air she breathes out different from the air she breathes in?

GLOBWARM (M, H) [Cross-processes, Large-scale practices]
a. How would you define or describe global warming?
b. What events from the list below do you think could cause global warming? Events Will the event contribute to global warming? (Note: Please circle "Yes" even if you think the contribution is small) If you circled Yes, please explain why the event will contribute to global warming.
[High]

Driving trucks long distances on the highway	**Yes**	No
Cutting down forests to have land for farming	**Yes**	No
Running a refrigerator with electricity	**Yes**	No
Using aerosol (spray can) hairspray	Yes	**No**
Eating lots of beef for dinner	**Yes**	No

[Middle]

Driving trucks long distances on the highway	**Yes**	No
Cutting down forests to have land for farming	**Yes**	No
Burning 95 candles on your great-great-aunt's birthday cake)	**Yes**	No
Using aerosol (spray can) hairspray	Yes	**No**

GLUGRAPE [Animal function, Energy/causes]
The grape you eat can help you move your body parts such as your legs.
a. Please describe how the substances from the grape provide energy to move your legs. Describe as many intermediate stages and processes as you can.
b. Can the substances of the grape also be involved in helping to keep your body warm? Please explain your answer.

GRANJOHN [Decomposition, Plant growth, Macroscopic, Microscopic]
Grandma Johnson had very sentimental feelings toward Johnson Canyon, Utah, where she and her late husband had honeymooned long ago. Because of these feelings, when she died she requested to be buried under a creosote bush in the canyon. Describe below the path of a carbon

atom from Grandma Johnson's remains, to inside the leg muscle of a coyote. NOTE: The coyote does not dig up and consume any part of Grandma Johnson's remains.

GROWTH [Cross-processes, Macroscopic]
Both plants and animals need air, water, and nutrients to grow. Can you think of ways that the plants and animals are different in the ways they use water, air, and nutrients to grow?

INFANT(E,M,H) [Animal growth, Mass/gases/amount]
Do you think the baby girl will need any of the following things to grow and gain weight? Please circle Yes or No and explain your choice. If you circled yes, explain how the girl's body uses it. What happens to it inside the inside the girl's body?
Sunlight Yes **No**
Water Yes **No**
Air Yes **No**
Food **Yes** No

KLGSEASON [Cross-processes, Large-scale practices]
The graph given below shows changes in concentration of carbon dioxide in the atmosphere over a 47-year span at Mauna Loa observatory at Hawaii, and the annual variation of this concentration.

a. Why do you think this graph shows atmospheric carbon dioxide levels decreasing in the summer and fall every year and increasing in the winter and spring?
b. Why do you think this graph shows atmospheric carbon dioxide levels increasing from 1960 to 2000?

LAMPELEC [Cross-processes, Energy/causes]
When you turn on a lamp, you can see the light. Where does the light energy come from? Trace the energy as far as you can. You may or may not fill up all of the spaces in the table.

	What form of energy was it? Where was it?
	Light energy of the light
Before that…	
Before that…	
Before that…	
Before that…	
Before that…	
Before that…	

OCTAMOLE [Combustion, Microscopic]

Gasoline is mostly a mixture of hydrocarbons such as octane: C8H18. Decide and circle whether each of the following statements is true (T) or false (F) about what happens to the atoms in a molecule of octane when it burns inside a car.
T F Some of the atoms in the octane are incorporated into carbon dioxide in the air.

T **F** Some of the atoms in the octane are incorporated into air pollutants such as ozone or nitric oxide.
T **F** Some of the atoms in the octane are converted into energy that moves the car.
T **F** Some of the atoms in the octane are burned up and disappear.
T **F** Some of the atoms in the octane are converted into heat.
T F Some of the atoms in the octane are incorporated into water vapor in the atmosphere.
a. When the gas tank is empty and the car stops, where is the energy that was in the gasoline?
b. What was the original source of energy of gasoline?
c. Is air needed for the car to use the gasoline? If so, how does the air change as the car runs?

PLANTGAS [Plant growth, Macroscopic]
Plants take in gas(es) from their environments. Please circle the gas(es) that plants take from their environments (You may circle more than one). You may also write down other gas(es).

Oxygen Carbon dioxide Other:_____
Explain what happens to the gas(es) once it is (they are) inside the plant.

POTATO [Decomposition, Microscopic]
A potato is left outside and gradually decays. One of the main substances in the potato is the starch amylose, which is made of many glucose molecules bonded together. What happens to the atoms in amylose molecules as the potato decays? Circle True (T) or False (F) for each option.

T **F** Some of the atoms are converted into nitrogen and phosphorous: soil nutrients.
T **F** Some of the atoms are used up by decomposers and disappear.
T F Some of the atoms are incorporated into carbon dioxide.
T **F** Some of the atoms are turned into energy by decomposers.
T F Some of the atoms are incorporated into water.

STONEWIN (E, M) [Animal function, Energy/causes]

You are playing outside in a cold winter. You find a stone on the ground. When you pick up the stone, you find that the stone is very cold. Why can people keep warm on a cold day, but stones cannot? Which of the thing(s) from the list below can help to keep people's bodies warm? Please circle YES or NO for each thing in the list below.
a. Water YES **NO**
b. Food **YES** NO
c. Air YES **NO**
d. Exercise YES **NO**
Try to write an explanation of how people's bodies stay warm that includes ALL of the things you circles "YES" for in the list above.

THINGTREE (E, M, H) [Plant growth, Mass/gases/amount]
A small oak tree was planted in a meadow. After 20 years, it has grown into a big tree, weighing 500 kg more than when it was planted. Do you think the tree will need any of the following things to grow and gain weight? Please circle Yes or No and explain your choice.If you circled yes, explain how the tree uses it. What happens to it inside the tree?

Sunlight **YES** NO
Soil **YES** NO
Water **YES** NO
Air **YES** NO

TREEDECAY [Decomposition, AB-Macroscopic, C-Energy/causes]
A tree falls in the forest. After many years, the tree will appear as a long, soft lump on the forest floor.

a. The lump on the forest floor weighs less than the original tree. What happened to it? Where would you find the matter that used to be in the tree?
b. What caused those changes in the wood? Explain as much as you can how these changes happened.
c. Is energy involved when the tree decays?
Circle one: Yes / No
If your answer is yes, please explain how energy is involved.

TROPRAIN [Cross-processes, Large-scale practices]
A tropical rainforest is an example of an ecosystem. Which of the following statements about matter and energy in a tropical rainforest is the most accurate? Please choose ONE answer that you think is best.
a. Energy is recycled, but matter is not recycled. (Level 2)
b. Matter is recycled, but energy is not recycled. (Level 4)
c. Both matter and energy are recycled. (Level 3)
d. Both matter and energy are not recycled. (Level 2)
Please explain why you think that the answer you chose is better than the others. (If you think that some of the other answers are partially right, please explain that, too.)

WAXBURN [Combustion, Mass/gases/amount]
A burning candle is put into an air-tight container. After some time, the candle stops burning.
a. Predict whether the air inside the candle will have more, the same, or less of the gases below. Explain where the gases come from or go to.

Gas	Prediction (circle)	Explanation: How did burning the candle produce or use the gas?
Oxygen	More Same **Less**	
Carbon dioxide	**More** Same Less	
Water vapor	**More** Same Less	

b. Where does the energy for burning come from? Please explain your answer.

WTLOSS(M,H) [Animal function, Microscopic]
When a person loses weight, what happens to some of the fat in the person's body? Choose ONE answer that you think is best.
a. The fat is broken down and leaves the person's body as water and gas. (Level 3)
b. The fat is converted into energy. (Level 2)
c. The fat is used up providing energy for the person's body functions. (Level 2)

d. The fat is broken down and leaves the person's body as feces and urine. (Level 1)
Please explain why you think that the answer you chose is better than the others. (If you think some of the other answers are also partially right, please explain that, too).

Table A.1 Descriptions of the four achievement levels of carbon cycle learning progression

Explaining	Specific Level Description
Level 4. Linking processes with matter and energy as constraints	**Macro:** Describe systems as conserving matter and energy in hierarchy of scales; Link macroscopic processes to chemical reactions with matter and energy as constraints; Link macroscopic processes to large-scale carbon cycle and energy flow. **Gases:** Correct explanation of gases (CO2 or O2) change in chemical reactions or in global-scale changes. **Micro:** Atomic-molecular accounts conserving atoms in chemical changes and/or conserving energy with degradation. **Large:** Describe matter cycle involving carbon transforming between organic and inorganic forms; Describe energy flow with degradation or connected to chemical reactions.
Level 3. Changes of Molecules and Energy Forms with Unsuccessful Constraints	**Macro:** Describe actors as systems containing matter and energy in hierarchy of scales, but do not conserve matter/energy successfully; Link macroscopic changes to chemical changes and describe chemical changes as changes involving atoms, organic molecules, and energy forms, but do not successfully conserve matter and energy. (e.g., organic molecule and energy conversion; energy conservation without degradation) **Gases:** Describe air as mixture of gases including CO2 or O2; Describe gas cycle as changes between CO2 and O2 and CO2 and O2 as different substances or molecules; Connect gas cycles with chemical changes. Identify CO2 as the product of combustion or cellular respiration. **Micro:** Trace materials to and from cells; Provide incomplete atomic-molecular accounts about changes of molecules and energy forms. **Large:** Link large-scale processes to macro or atomic-molecular processes, but without full conservation of matter and energy; Describe large-scale processes as materials passing on without organic carbon generation and oxidation; Describe energy passing on without degradation or connecting to chemical reactions.

Table A.1 (Cont'd)

Level 2. Force-dynamic accounts with hidden mechanisms	**Macro:** Still focus on actors, enablers, and results, but link the macroscopic changes to hidden mechanisms that involving changes of materials and energy in general. **Gases:** May use CO2 or O2 to describe the quality of the air. Describe gas changes in life-related events. **Micro:** Link macro-processes with unobservable mechanisms or hidden actors (e.g., decomposer) **Large:** Describe networks of actors & enablers (e.g., food chains with emphasis on eating rather than matter/energy flow.)
Level 1. Macroscopic force-dynamic accounts	**Macro:** Describe macro-processes in terms of the action-result chain: actors use enablers to accomplish their goals; interactions between actors and enablers are like macroscopic physical push-and-pull that does not involve any change of matter/energy. **Gases:** Air (fresh air, bad air) as enablers or waste products of the actor. No explicit gas. exchange. **Large:** No connections to larger systems. Simple food chains as series of events. **Micro:** Connections to subsystems limited to parts student can see or feel.

Table A.2 The specific rubric of the CARGAS item

	CARGAS(E,M): When you are riding in a car, the car uses gasoline to make it run. Eventually the gasoline tank is empty. a. What happens to the materials the gasoline is made of when the car uses the gasoline? b. Is air needed for the car to use the gasoline? If so, how does the air change as the car runs? c. Where does energy come from to make the car run?	
Explaining	Specific level description for this item	Typical example
Level 4. Linking processes with matter and energy as constraints	Macro: - Describe systems as conserving matter and energy in hierarchy of scales; - Link car running to the combustion (or burning) of gasoline in which they trace matter OR energy successfully. Tracing matter successfully means that they explain the consumption of gasoline by stating that materials of gasoline and air (NOTE: mentioning O2 is not necessary because not asked for) change into CO2 which goes into the air. Tracing energy successfully means that they explain that the energy that makes the car run ultimately comes from high-energy bonds (C-C, C-H) or chemical energy in the materials of gasoline. To be level 4, they also must not confuse matter and energy in tracing them (e.g., No matter/energy conversion). - Link macroscopic processes to large-scale carbon cycle and energy flow. Gases: - Correctly explain that air is needed and CO2 is produced in the combustion of gasoline. Micro: - Correctly describe atomic-molecular accounts tracing carbon through combustion (materials of gasoline -> CO2). - Identifies the materials of gasoline and air as reactants and carbon dioxide as a key product. Large: - Describe matter cycle involving carbon transforming between organic and inorganic forms; - Describe energy flow with degradation or connected to chemical reactions. - Explain combustion at atomic-molecular level, consistently trace matter and energy through the process.	a. The gasoline is turned into CO2 and H2O through combustion. b. The air is connected to carbon and hydrogen molecules in the fuel and is turned into CO2 and H2O. c. The fuel - which is chemical energy. a. when the gas runs out, that means all of the high energy bonds were broken down b. Yes / the air helps break down the bonds c. the ultimate source of energy in the gasoline is c_c, and C-H bonds

Table A.2 (Cont'd)

Level 3. Changes of Molecules and Energy Forms with Unsuccessful Constraints	Macro: - Describe actors as systems containing matter and energy in hierarchy of scales, but do not conserve matter/energy successfully; - Link car running to the combustion (or burning) of gasoline in which they trace matter or energy yet unsuccessfully. This may entail one of several things: (1) they trace both matter and energy, but they confuse matter and energy in tracing them (e.g., the materials of gasoline change into energy that makes the car run); (2) they trace matter only and unsuccessfully (e.g. they explain that the materials of gasoline change into some gases without identifying them); (3) they trace energy only and unsuccessfuly (e.g. they mention some forms of energy other than chemical energy or "bonds" without identifying them.) Gases: - Describe air as mixture of gases including CO2 or O2; - Describe gas cycle as changes between CO2 and O2 and CO2 and O2 as different substances or molecules; - Connect gas cycles with chemical changes. - Identify CO2 as a product of the combustion of gasoline. Micro: - Trace materials to and from cells - Provide incomplete atomic-molecular accounts about changes of molecules and energy forms. Large: - Link large-scale processes to macro or atomic-molecular processes, but without full conservation of matter and energy; - Describe large-scale processes as materials passing on without organic carbon generation and oxidation; - Describe energy passing on without degradation or connecting to chemical reactions.	a. They get made into different types of energy. b. No c. The gas being changed into motion energy.

Table A.2 (Cont'd)

Level 2. Force-dynamic accounts with hidden mechanisms	Micro: Link macro-processes with unobservable mechanisms or hidden actors (e.g., decomposer) Large: Describe networks of actors & enablers (e.g., food chains with emphasis on eating rather than matter/energy flow.) Macro: Still focus on actors, enablers, and results, but link the car running to hidden mechanisms that involve changes of materials and energy in general (e.g., identify gasoline as the source of the energy that makes the car run). Gases: May use CO2 or O2 to describe the quality of the air without any explanation of mechanism. Mention air, smoke or/and ash as byproducts of burning. Describe gas changes in life-related events. Micro: Link macro-processes with unobservable mechanisms or hidden actors. May describe explosion or any process to provide energy in engine, cylinder and piston. Large: Describe networks of actors & enablers (e.g., car running with emphasis on burning gasoline)	A: The materials get burned and they come out of your gas pipe where smoke comes out. B: No, cause when the materials burn it makes smoke and air. C: The energy comes from the gas in order for the car to run.
Level 1. Macroscopic force-dynamic accounts	Macro: Describe car running in terms of the action-result chain: car uses enablers (e.g. gasoline) to run; interactions between actors and enablers are like macroscopic physical push-and-pull that does not involve any change of matter/energy (e.g. without gasoline your car cannot work, Energy comes from running of the car). Gases: Describe air in term of tire, engine, people breath in car running OR state that air is not involved in this process. Large: No connections to larger systems Micro: Connections to subsystems limited to parts student can see or feel Describe how the materials the gasoline is made of get used using single-process/step words (e.g., "evaporate," "gets burned," "used up") without providing additional steps (e.g., "go out in the air"), hierarchy in structure, or smaller entities.	A. It is used up. B. No C. The engine.

Table A.3 Unidimensional PCM results

				WEIGHTED FIT		
ID	Item	ESTIMATE	ERROR^	MNSQ	CI	T
1	ACRON_MC	0.09	0.06	1.18	(0.91, 1.09)	3.9
2	ACRON_CR	0.842	0.065	1.01	(0.89, 1.11)	0.2
3	AIREV_CR	-0.169	0.091	0.94	(0.81, 1.19)	-0.6
4	AIRBO_CR	0.95	0.091	0.96	(0.70, 1.30)	-0.2
5	ANIMW_CR	-1.097	0.106	1	(0.75, 1.25)	0
6	APPLR_CR	-0.107	0.085	0.9	(0.84, 1.16)	-1.2
7	BODYE_CR	0.19	0.075	1.02	(0.85, 1.15)	0.3
8	BODYH_MC	-1.842	0.097	1.18	(0.74, 1.26)	1.3
9	BODYH_CR	-0.64	0.081	1.04	(0.79, 1.21)	0.4
10	BREAD_MC	-1.024	0.057	1.09	(0.90, 1.10)	1.7
11	BREAD_CR	-0.194	0.054	0.96	(0.88, 1.12)	-0.6
12	MATCHEL_CR	-0.407	0.09	0.91	(0.82, 1.18)	-1
13	MATCHM_MC	-0.372	0.061	1.16	(0.90, 1.10)	3.0
14	MATCHMA_CR	0.104	0.061	0.98	(0.88, 1.12)	-0.3
15	MATCHMB_CR	0.127	0.065	0.93	(0.87, 1.13)	-1
16	CARBO_CR	0.991	0.094	1.03	(0.74, 1.26)	0.3
17	CARPA_CR	0.696	0.096	0.94	(0.76, 1.24)	-0.5
18	CARGA_CR	-0.65	0.066	0.91	(0.85, 1.15)	-1.2
19	CONNL_CR	-0.432	0.099	0.99	(0.80, 1.20)	-0.1
20	CUTTR_CR	0.032	0.112	1.07	(0.65, 1.35)	0.4
21	DEERW_MC	0.292	0.085	1.13	(0.77, 1.23)	1.1
22	DEERW_CR	1.037	0.098	0.97	(0.74, 1.26)	-0.2
23	DIFEV_CR	0.086	0.061	0.98	(0.88, 1.12)	-0.4
24	EATAP_CR	0.101	0.062	0.94	(0.89, 1.11)	-1.1
25	EATBR_CR	0.415	0.071	1.01	(0.85, 1.15)	0.2
26	ECOSP_CR	0.62	0.08	1.05	(0.81, 1.19)	0.5
27	ENERP_CR	-0.05	0.044	0.92	(0.90, 1.10)	-1.5
28	ENPLN_CR	-0.094	0.056	0.96	(0.89, 1.11)	-0.7
29	GIRLAB_CR	0.009	0.105	0.95	(0.72, 1.28)	-0.3
30	GIRLC_CR	-0.389	0.121	0.83	(0.79, 1.21)	-1.6
31	GLOBM_CR	-0.297	0.084	0.92	(0.80, 1.20)	-0.8
32	GLOBH_CR	-0.809	0.092	0.94	(0.78, 1.22)	-0.5
33	GLUEG_CR	-0.906	0.079	0.92	(0.84, 1.16)	-1
34	GRAND_CR	0.749	0.107	1.01	(0.69, 1.31)	0.1
35	GRANP_CR	0.086	0.111	0.92	(0.72, 1.28)	-0.5
36	GROWT_CR	0.992	0.078	0.92	(0.85, 1.15)	-1.1
37	INFAN_CR	0.767	0.076	0.92	(0.89, 1.11)	-1.4
38	KLGSE_CR	0.532	0.094	0.98	(0.78, 1.22)	-0.2
39	LAMPE_CR	-0.345	0.061	1.11	(0.85, 1.15)	1.4
40	OCTAM_CR	0.417	0.096	0.85	(0.79, 1.21)	-1.4

Table A.3 (Cont'd)

41	PLANG_CR	0.796	0.084	0.9	(0.88, 1.12)	-1.7
42	THINT_CR	-0.151	0.057	0.82	(0.90, 1.10)	-3.5
43	TREDEAB_CR	0.252	0.056	0.92	(0.89, 1.11)	-1.4
44	TREDEC_CR	-0.015	0.048	0.94	(0.88, 1.12)	-0.9
45	TROPRA_MC	-1.041	0.063	1.17	(0.89, 1.11)	3.0
46	TROPRA_CR	-0.94	0.061	0.97	(0.89, 1.11)	-0.6
47	WAXBUR_CR	1.367	0.1	0.95	(0.70, 1.30)	-0.3
48	WTLOSS_MC	-1.488	0.057	1.14	(0.91, 1.09)	3.0
49	WTLOSS_CR	-0.512	0.051	1.02	(0.91, 1.09)	0.4
50	AIREVE_MTF	-1.2	0.089	1.22	(0.83, 1.17)	2.3
51	ANIM_MTF	0.003	0.11	1.25	(0.67, 1.33)	1.5
52	BODY_MTF	0.562	0.105	1.09	(0.82, 1.18)	1
53	ENERPE_MTF	-0.431	0.055	1.12	(0.88, 1.12)	1.9
54	ENERPL_MTF	0.141	0.065	1.11	(0.83, 1.17)	1.2
55	GLOBM_MTF	0.107	0.104	1.11	(0.79, 1.21)	1
56	GLOBH_MTF	0.693	0.093	1	(0.80, 1.20)	0.1
57	INFA_MTF	0.794	0.088	1.26	(0.79, 1.21)	2.2
58	OCTAM_MTF	0.281	0.086	1.08	(0.76, 1.24)	0.7
59	POTATO_MTF	0.572	0.624	1.36	(0.72, 1.28)	2.3
60	THINGT_MTF	-1.353	0.073	1.2	(0.89, 1.11)	3.3

The column headers are:

- Estimate column provides the item difficulty estimates for every item
- Error column provides the error of the item difficulty estimates
- MNSQ is the mean residual square between what is observed and what is expected.
- CI is the confidence interval of the MNSQ.
- T is the *t*-statistics that used to indicate the fitness of the item to the model.

Table A.4 The step threshold parameters of 38 good items

ID	item	b	d_1	d_2	d_3
1	ACRON_CR	0.745	-1.778	1.132	2.882
2	AIREV_CR	-0.277	-1.033	0.478	
3	ANIMW_CR	-1.208	-1.981	-0.435	
4	APPLR_CR	-0.227	-1.196	0.743	
5	MATCHEL_CR	-0.592	-1.367	0.184	
6	MATCHM_MC	-0.518	-1.293	0.258	
7	MATCHMA_CR	-0.005	-2.156	0.221	1.919
8	MATCHMB_CR	0.017	-2.362	0.317	2.098
9	CARBO_CR	0.957	0.754	1.161	
10	CARPA_CR	0.562	-3.348	2.428	2.605
11	CARGA_CR	-0.824	-2.792	-0.099	0.418
12	CONNL_CR	-0.583	-2.718	1.551	
13	CUTTR_CR	-0.031	-0.34	0.277	
14	DEERW_CR	0.986	0.4	1.573	
15	DIFEV_CR	-0.028	-1.265	-0.452	1.633
16	EATAP_CR	-0.038	-2.341	-1.025	3.251
17	EATBR_CR	0.299	-1.646	0.773	1.771
18	ENPLN_CR	-0.248	-1.838	-0.427	1.52
19	GIRLAB_CR	-0.084	-1.665	-0.824	2.238
20	GLOBM_CR	-0.445	-2.182	-0.273	1.119
21	GLOBH_CR	-0.978	-3.268	-0.807	1.139
22	GLUEG_CR	-1.103	-3.576	-1.278	1.544
23	GRAND_CR	0.703	-2.095	1.773	2.432
24	GRANP_CR	-0.009	-1.919	1.901	
25	GROWT_CR	0.915	-1.128	1.069	2.805
26	INFAN_CR	0.674	-0.893	2.241	
27	KLGSE_CR	0.486	-1.272	0.242	2.488
28	PLANG_CR	0.666	-1.563	2.895	
29	TREDEAB_CR	0.123	-2.131	0.899	1.603
30	TROPRA_CR	-1.138	-4.431	-0.805	1.822
31	WAXBUR_CR	1.317	0.959	1.674	
32	WTLOSS_CR	-0.665	-3.314	-0.605	1.925
33	ANIM_MTF	-0.089	-2.123	0.11	1.746
34	BODY_MTF	0.411	-1.442	2.264	
35	ENERPE_MTF	-0.599	-1.495	-1.088	0.784
36	GLOBM_MTF	-0.024	-1.38	1.333	
37	GLOBH_MTF	0.616	-1.414	-0.229	3.492
38	OCTAM_MTF	0.236	-0.565	0.584	0.688

The column headers are:

- b is the item difficulty parameter
- d_1 is the first item threshold parameter, it's the cutting point on the ability scale between score 1 and 2
- d_2 is the second item threshold parameter, it's the cutting point on the ability scale between score 2 and 3
- d_3 is the third item threshold parameter, it's the cutting point on the ability scale between score 3 and 4

Table A.5 Excluded items (misfit items and items that have thresholds not in the correct order)

ID		Item
1	ACORN_MC	misfit
2	THINT_CR	
3	AIREVE_MTF	
4	INFA_MTF	
5	POTATO_MTF	
6	THINGT_MTF	
7	AIRBO_CR	thresholds not in order
8	BODYE_CR	
9	BODYH_MC	
10	BODYH_CR	
11	BREAD_MC	
12	BREAD_CR	
13	DEERW_MC	
14	ECOSP_CR	
15	ENERP_CR	
16	LAMPE_CR	
17	OCTAM_CR	
18	TREDEC_CR	
19	ENERPL_MTF	
20	GIRLC_CR	
21	TROPRA_MC	
22	WTLOSS_MC	

Table A.6 Effective options of each MTF item

Item Name	Sub T or F questions	Effective option (X)
AIREVENT	PLANT GROWTH	X
	GIRL RUN	
	BURNING WOOD	X
	FOOD DECAY	
ANIMWINT	HEAT	X
	GAS/WATER	X
	WASTE	
	OTHER MATERIAL	
	DISAPPEAR	
BODYTEMP	WATER	
	FOOD	X
	AIR	
	EXERCISE	
ENERPEOP	WATER	X
	FOOD	
	NUTRIENT	X
	EXERCISE	X
	SUNLIGHT	X
	CO2	X
	O2	X
ENERPLNT	WATER	X
	LIGHT	
	AIR	X
	NUTRIENTS	X
	OWN ENERGY	
GLOBWARMM	TRUCK	X
	FOREST	X
	CANDLE	X
	HAIR SPRAY	X
GLOBWARMH	TRUCK	X
	CUT TREE	X
	REFRIGERATOR	X
	AEROSOL	X
	BEEF	X
INFANT	SUN	
	WATER	
	AIR	
	FOOD	
OCTMALE	CO2	X
	AIR POLLUTION	
	ENERGY	X
	DISAPPEAR	X

Table A.6 (Cont'd)

	HEAT	X
	WATER	X
THINGTREE	SUN	X
	SOIL	
	WATER	
	AIR	X

X indicates there are significant group differences in term of the paired CR score between students who selected T and how selected F to the question.

REFERENCES

REFERENCES

Ackerman, T. A. (1994a). Creating a test information profile in a two-dimensional latent space. *Applied Psychological Measurement*, *18*, 257–275.

Ackerman, T. A. (1994b). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *20*, 309–310.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*, 311–329.

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csake (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest, Hungary: Akademiai Kiado.

Allen, M.J., & Yen, W. M. (2002*). Introduction to Measurement Theory.* Long Grove, IL: Waveland Press.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5), 1–63. Retrieved August 30, 2005, from http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1008&context=jtla

Alonzo, A. C., & Steedle, J. T. (2008). Developing and assessing a force and motion learning progression. *Published online in Wiley InterScience.*

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anderson, C. W. (2010, March) Learning Progressions for Environmental Science Literacy. Paper *presented at NRC Science Framework Committee.*

Anderson, C. W., Alonzo, A. C., Smith, C., & Wilson, M. (2007, August). NAEP pilot learning progression framework. Report to the National Assessment Governing Board.

Anderson, C. W., Sheldon, T. H., & Dubay, J. (1990). The effects of instruction on college nonmajors' conceptions of respiration and photosynthesis. *Journal of Research in Science Teaching, 27,* 761-776

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

Boyes, E., & Stanisstreet, M. (1993). The greenhouse effect: Children's perceptions of causes, consequences and cures. International Journal of Science Education, 15 (5), 531-552.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345-370.

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

Champagne, A. B., Kouba, V. L., & Gentiluomo, L. (2008). Assessing science literacy using extended constructed-response items. In J. Coffey, R. Douglas, & C. Sterns (Eds.), *Assessing science learning: Perspectives from research and practice.* Arlington, VA: NSTA Press.

Chen, J., Anderson, C. W., Choi, J., Lee, Y., & Draney, D. (2010). *Assessing K-12 students' learning progression of carbon cycling using different types of items.* Paper presented at National Association for Research in science Teaching, Philadelphia, PA.

Coyle, K. (2005). Environmental literacy in America: What ten years of NEETF/Roper research and related studies say about environmental literacy in the U.S. Washington, DC: The National Environmental Education & Training Foundation.

Crowley, T.J. (2000). Causes of climate change over the past 1000 years. *Science, 289(270),* 270-277.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*, 14-24.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society,* Series, B, *39*, 1-38.

Dunham, M. L. (2007) An investigation of the multiple true-false item for nursing licensure and potential sources of construct-irrelevant difficulty. http://proquest.umi.com/pqdlink?did=1232396481&Fmt=7&clientId=79356&RQT=309&VName=PQD

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

Ebel, R. L. (1979). *Essentials of Educational Measurement* (3rd ed.). Englewood Cliffs, N.J.: Prentice Hall.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175–186.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35*, 137-154.

Falkowski, P., Scholes, R.J., Boyle, E., Canadell, J., Canfield, D., Elser, J., Gruber, N., Hibbard, K., Hogberg, P., Linder, S., Mackenzie, F.T., Moore III, B., Pederson, T., Rosenthal, Y., Seitzinger, S., Smetacek, V., Steffen, W. (2000). The global carbon cycle: A test of our knowledge of Earth as a system. *Science, 290(291)*, 291-296.

Fisher, Kathleen M. et al. (1986, February). Student Misconceptions and Teacher Assumptions in College Biology. *Journal of College Science Teaching, 15(4),* 276-80

Frederiksen, N. (1984). The real test bias: Influence of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice,11*(4), 21–26.

Grosse, M. E., & Wright, B. D. (1985). Validity and reliability of True-False tests. *Educational and Psychological Measurement, 45*(1), 1-13

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. Annals of Statistics 5: 1148-1169.

Hestenes, D.,Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30,* 141–151.

Hmelo-Silver, C.E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding in complex systems. *Journal of the Learning Sciences, 16(3)*, 307-331.

Hotinski, R. (2007). *Stabilization Wedges: A Concept & Game,* from http://www.princeton.edu/~cmi/resources/stabwedge.htm

IPCC, 2007: Summary for Policymakers. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jin, H., & Anderson, C. W. (2010, March). *Developing a long-term learning progression for energy in socio-ecological system.* Paper presented at National Association for Research in science Teaching, Philadelphia, PA.

Kansas Environmental Education Conference (KACEE) Newsletter. (2005). From http://www.kacee.org/About/newsletters/KACEE%20Newsletter%20Fall%202005.pdf

Keeling, C.D. & Whorf, T.P. (2005). *Atmospheric CO2 records from sites in the SIO air sampling network.* In Trends: A Compendium of Data on Global Change. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN.

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. Washington, D.C.: *ERIC Clearinghouse on Assessment and Evaluation,* ERIC Identifier: ED398237.

Kempton, W., Boster, J. S., and Hartley, J. A. (1995). *Environmental values and American culture*. Cambridge, MA: MIT Press.

Kennedy, C.A. (2005). *The BEAR Assessment System: A Brief Summary for the Classroom Context.* BEAR Technical Report Series 2005-03-01. Berkeley, CA: University of California, BEAR Center.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis, 21*(4), 345-363.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*(3), 6-14.

Lazarsfeld P.F, & Henry N.W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Lee, H. -S., Liu, O. L., & Linn, M. C. (2011). Validating Measurement of Knowledge Integration Science Using Multiple-Choice and Explanation Items. *Applied Measurement in Education, 24*(2), 115-136.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement, 41(3),* 205-237.

Lin, C.-Y., & Hu, R. (2003). Students' understanding of energy flow and matter cycling in the context of the food chain, photosynthesis, and respiration. International Journal of Science Education, 25(12), 1529-1544.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-173.

McNeill, K. L. & Krajcik, J. (2007). *Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations.* In Lovett, M & Shah, P (Eds.)

Thinking with data: The proceedings of the 33rd Carnegie symposium on cognition. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Merritt, J. & Krajcik, J. (2009, June). Developing a calibrated progress variable for the particle nature of matter. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Mislevy, R. J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*. 1(1): 3-62.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching*. 46 (6), 675-698.

Mohan, L., Chen, J., Baek, H., Anderson, C.W., Choi, J., & Lee, Y. (2009, April). *Validation of a multi-year carbon cycle learning progression.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73–90.

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

National Research Council. (2006). Systems for state science assessment. Washington, DC: The National Academies Press.

National Research Council. (2007). Taking science to school. Washington, DC: The National Academies Press.

NEETF & Roper Starch Worldwide. (2001). *Lessons from the Environment:The Ninth Annual National Report Card on Environmental Attitudes, Knowledge and Behavior.* Washington, DC: NEETF.

Novick, M.R. (1966) The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1), 1-18.

Pellegrino, J. W. (2009). The design of an assessment system for the race to the top: A learning sciences perspective on issues of growth and measurement. *Paper presented at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda.* http://www.k12center.org/rsc/pdf/PellegrinoPresenterSession1.pdf

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

Rao, C. R. & Sinharay, S. (2007). *Handbook of statistics,* Vol. 26: Psychometrics. Elsevier Science B.V.: The Netherlands, 2007.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Reckase, M. D. (1985). The difficulty of test items that measure more than on ability. *Applied Psychological Measurement*, *9*, 401–412.

Reckase, M. D. (1990, April). Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests. *Paper presented at the Annual Meeting of the American Educational Research Association.*

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Reckase, M. D., & Martineau, J. A. (2004). *The vertical scaling of science achievement tests.* Unpublished Report, Michigan State University, East Lansing, MI.

Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York: Springer.

Reckase, M. D. & Hirsch, T.M. (1991). Interpretation of number –correct scores when the true numbers of dimension assessed by a test is greater than two. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361–373.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461-464.

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

Songer, C. J., & Mintzes, J. J. (1994). Understanding cellular respiration: An analysis of conceptual change in college biology. *Journal of Research in Science Teaching 31,* 621-637.

Steedle, J. T. (2006, April). Seeking evidence supporting assumptions underlying the measurement of progress variable levels. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Stout, W., Douglas, B., Junker, B. & Roussos, L. (1999). DIMTEST [computer software]. The William Stout Institute for Measurement, Champaign, IL.

Stout, W., Froelich, A. G. & Gao, F. (2001). Using resampling to produce and improved

DIMTEST procedure. In Boomsma A, van Duijn MAJ, Snijders TAB (eds.) *Essays on item response theory* (pp. 357-375). Springer-Verlag, New York.

Thorndike, R. M. (2005). Measurement and evaluation in psychology and education. Prentice Hall, Upper Saddle River (NJ).

Wei, H. (2008). *Multidimensionality in the NAEP Science Assessment: Substantive perspectives, psychometric models, and task design.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No.[3307786])

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

Wilson, C. D., Anderson, C. W., Heidemann, M., Merrill, J., Merritt, B. W., Richmond, G., Sibley, D. F., & Parker, J. M. (2006). Assessing Students' Ability to Trace Matter in Dynamic Systems in Cell Biology. CBE—Life Sciences Education, 5(4), 323–331.

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

Wilson, M., & Wang, W.-C. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19*(1), 51-71.

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

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models.* Unpublished master's thesis, University of Melbourne.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modeling software*. Melbourne, Australia: Australian Council for Educational Research.

Yao, L., & Boughton, K. A. (2007). A Multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*, 1–23.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement*, *30*, 469–492.

Zhang, JM. & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika* 64: 213-249