Title

Investigating the Validity of a Scientific Argumentation Assessment Using Psychometric Methods

Author

Yao, Shih-Ying

Publication Date

2013

Investigating the Validity of a Scientific Argumentation Assessment Using Psychometric Methods

By

Shih-Ying Yao

A dissertation submitted in partial satisfaction of the requirements for the degree of
Doctor of Philosophy
in
Education
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Sophia Rabe-Hesketh
Professor Jonathan Osborne
Professor Alan Hubbard

Fall 2013

Abstract

Investigating the Validity of a Scientific Argumentation Assessment Using Psychometric Methods

By

Shih-Ying Yao

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

With the increasing focus on incorporating argumentation in the teaching and learning of science in recent years, how to assess students' scientific argumentation ability becomes an emerging challenge. There exists very little work on the development and validation of new kinds of assessments for scientific argumentation. In response to this challenge, a progression-based assessment was developed to assess scientific argumentation ability in middle school. This dissertation examined validity evidence for the scientific argumentation assessment with three separate yet related studies. The first study applied factor analysis and Rasch modeling to investigate the dimensionality and psychometric properties of the argumentation items. The second study applied latent class modeling to examine whether the hypothesized structure of the scientific argumentation learning progression, which is the backbone of the studied assessment, could be observed in the empirical response patterns. The third study applied a multidimensional item response modeling framework to investigate the differences, or similarities, in the function of the studied argumentation items and traditional content items. Collectively, these three studies contribute to the understanding of the performance of the assessment of interest, and inform future research on assessments of argumentation within the domain of science.

Dedication

To my father Ching-Huei Yao, my mother Hui-Fang Lan, my brother Chih-Wei Yao, and my husband Ee Hou Yong, for their endless love and support, this dissertation is dedicated.

Acknowledgements

I would like to first express my gratitude to my committee chair and academic advisor, Mark Wilson, for his sustained support during my years at Berkeley. In addition to academic mentoring, Mark has offered me great opportunities to participate in different research projects, which apply psychometric methods to the development and validation of educational assessments. Through these research projects, I have gained invaluable learning experiences. I thank Professor Sophia Rabe-Hesketh for being a wonderful role model in teaching and doing research, and for her guidance through my journey at Berkeley. I thank Professor Jonathan Osborne for his work on scientific argumentation and his support for the completion of this dissertation. I thank Professor Alan Hubbard for his insightful comments and kind help through the development of this dissertation.

I thank my dear friends in Taiwan and U.S. for their encouragement. Their friendship has helped me overcome all the frustration through my Ph.D. years. Special thanks go to my friends at Berkeley. I moved to Boston at the dissertation-writing stage, and had to fly back to Berkeley periodically. Without the kind and enduring support from my friends at Berkeley, the last stage of my Ph.D. study would have been even more difficult.

My deepest gratitude goes to my parents, my brother, and my husband, for their unconditional love and support. To them I dedicate this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

In recent years, there has been an increasing emphasis on incorporating scientific argumentation into practice of science education. Various research studies have reported that engaging in the process of argumentation led to enhanced understanding of scientific content knowledge (e.g.,Ford, 2008;Zohar & Nemet, 2002). Recently, the need for engaging students in argumentation is formally addressed in the national science standards. Specifically, A Framework for K–12 Science Education (National Research Council, 2012), along with the forthcoming Next Generation Science Standards, specifies "*engaging in argument from evidence*" as one of the eight essential science practices in grades K-12.

With the increasing focus on incorporating argumentation in the teaching and learning of science, how to assess students' scientific argumentation ability becomes an emerging challenge. As Osborne and colleagues (Osborne, Henderson, MacPherson, & Szu, 2012) pointed out, there exists very little work that attempts to operationalize scientific argumentation in a manner such that it can be assessed in a quantitative or even semi-quantitative way. Without valid and reliable assessments, science educators would not be able to readily diagnose students' argumentation ability. As a consequence, argumentation may not be incorporated adequately as an essential practice in science classrooms, even with the formal inclusion in national standards. Development and validation of new kinds of assessments for scientific argumentation is therefore needed.

In response to the call for research on scientific argumentation, a progression-based scientific argumentation assessment has been developed. The effort cited is part of a larger research project titled *Learning Progressions in Middle School Science Instruction and Assessment*, which is funded by Institute of Education Sciences and conducted from 2010 to 2014. In this project, researchers from University of California, Berkeley and Stanford University have been collaborating with middle-school science educators to develop learning progressions in student understanding of content and argumentation within a middle school science domain, Structure of Matter. Progression-based assessments have been developed correspondingly.

The primary objective of this dissertation is to apply different psychometric methodologies to investigate various pieces of validity evidence of a particular scientific argumentation assessment contextualized in Changes of State, which is a topic in the Structure of Matter domain, with three separate yet related studies. The development of the studied scientific argumentation assessment is based on a construct modeling approach (Brown & Wilson, 2011;Wilson, 2005), known as the BEAR Assessment System (BAS), which consists of four building blocks. The first building block, the core of the construct modeling approach, is a cognition model, which lays out a hypothesized student understanding of a concept from the lowest to the highest proficiency. The primary interest of measurement is to examine where a student stands on this continuum. To provide useful interpretations of student performance, a cognition model can be defined as distinguishable qualitative levels. Nevertheless, Wilson (2005) emphasized that, *"Although qualitative levels are definable, we assume that the respondents can be at any point in between--- that is, the underlying construct is continuous."* (pp.6-7). In the remaining text, this cognition model would be referred to as learning progression. The second building block is item development. Assessment items are designed with the aim of eliciting student responses indicative of progression levels. The third building block is the outcome space. The outcome space is a generalized approach for valuing responses across a range of items, which categorizes student responses and aligns each response category with a specific level on the learning progression. When this is made specifically to an individual item, we refer to it as a scoring guide. Measurement models, as the fourth building block, are applied subsequently by researchers to investigate the validity and reliability of an assessment. This procedure may iterate multiple times for the development of an assessment that measures the construct of interest with sufficient validity and reliability. For details, please see Wilson (2005). This chapter describes each of the four building blocks in the development of the studied scientific argumentation assessment and the data used in this dissertation.

1.2 A Scientific Argumentation Learning Progression

In response to the call of research on argumentation, Osborne and colleagues (2012) have proposed a hypothetical learning progression for argumentation in science. Walton (Walton, 1990) has pointed out that "*argument*" and "*reasoning*" are different in the sense that, reasoning is an ability used in the construction of arguments. Other research has also shown that a student's reasoning ability depends on domain-specific knowledge (e.g.,Perkins & Salomon, 1989). Following this conceptualization, Osborne and colleagues argued that,

> *"... an "argumentation progress map" represents a map of the reasoning component of the competency to construct and critique domain-specific arguments. Whilst the competency to argue is clearly dependent on knowledge as well, our initial interest lies in identifying a map for the reasoning component required in the performance of argumentation." (Osborne et al., 2012, p.3).*

Focusing on the reasoning component of argumentation competency, Osborne and colleagues adopted the well-known Toulmin's model of argument (Toulmin, 2003) as the basis for formulating their hypothesis of the development of reasoning ability.

1.2.1 Toulmin's Model of Argument

Toulmin's model of argument has been widely applied to analyze student discourse in science. In his book titled *The Uses of Argument* (Toulmin, 2003), Toulmin has identified six key elements for analyzing the pattern of an argument: claim, data, warrant, qualifier, rebuttal, and backing. An initial skeleton of a pattern for analyzing arguments is composed of the three main elements: claim, data, and warrant. Claim (C) is the *"conclusion whose merits we are seeking to establish"* and data (D) is the *"facts we appeal to as a foundation for the claim"* (Toulmin, 2003; p.90). While data presents what a specified claim is based on, a warrant acts as a bridge between data and claim. Warrants may be written very briefly, such as 'If D, then C,' or be made more explicit, such as 'Given data D, one may take it that C.'

Figure 1.1 lays out this fundamental pattern of an argument, along with an example in parenthesis. The relation between the data and the corresponding claim is symbolized with an arrow, while the warrant that authorizes us to take a step from the data to the specified claim is indicated immediately below the arrow (Toulmin, 2003).

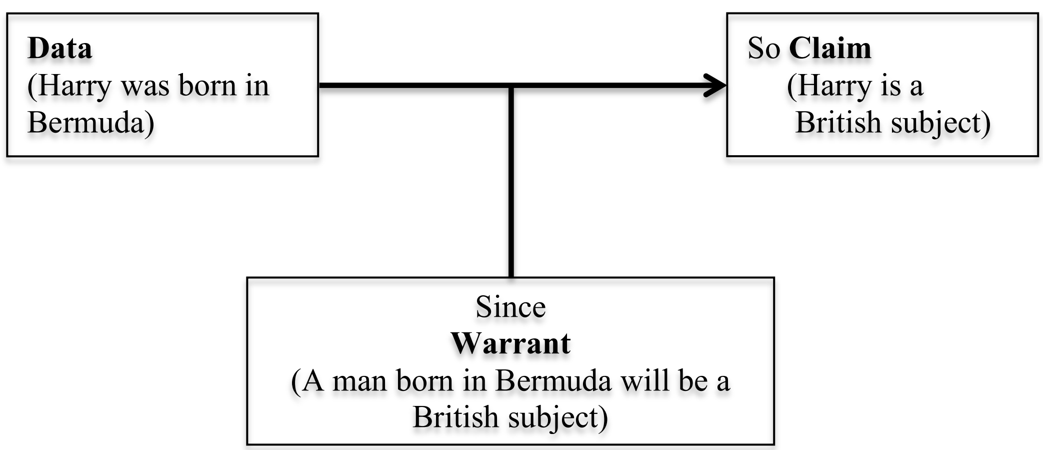


Figure 1.1. A Basic Skeleton of a Pattern for Analyzing Arguments (Toulmin, 2003)

It might not be sufficient to simply specify claim, data, and warrant to establish a sound argument. Specifically, Toulmin (2003) argued that, *"warrants are of different kinds, and may confer different degrees of force on the conclusions they justify."* It may thereby be necessary for one to provide an explicit reference to the degree of force which a particular piece or set of data confer on a specified claim in virtue of a warrant. Built upon the basic skeleton as shown in Figure 1.1, three additional elements can be added to compose a more complex pattern for analyzing arguments: qualifiers, rebuttals, and backing. Qualifiers indicate *"the strength conferred by the warrant"* on connecting the data to the claim, while conditions of rebuttal indicate *"circumstances in which the general authority of the warrant would have to be set aside"* (Toulmin, 2003; p.94).

Alternatively, backing presents statements that provide support for warrants. Figure 1.2 presents a typical complex skeleton of a pattern for analyzing arguments, which contains all six elements, with the example in parenthesis.

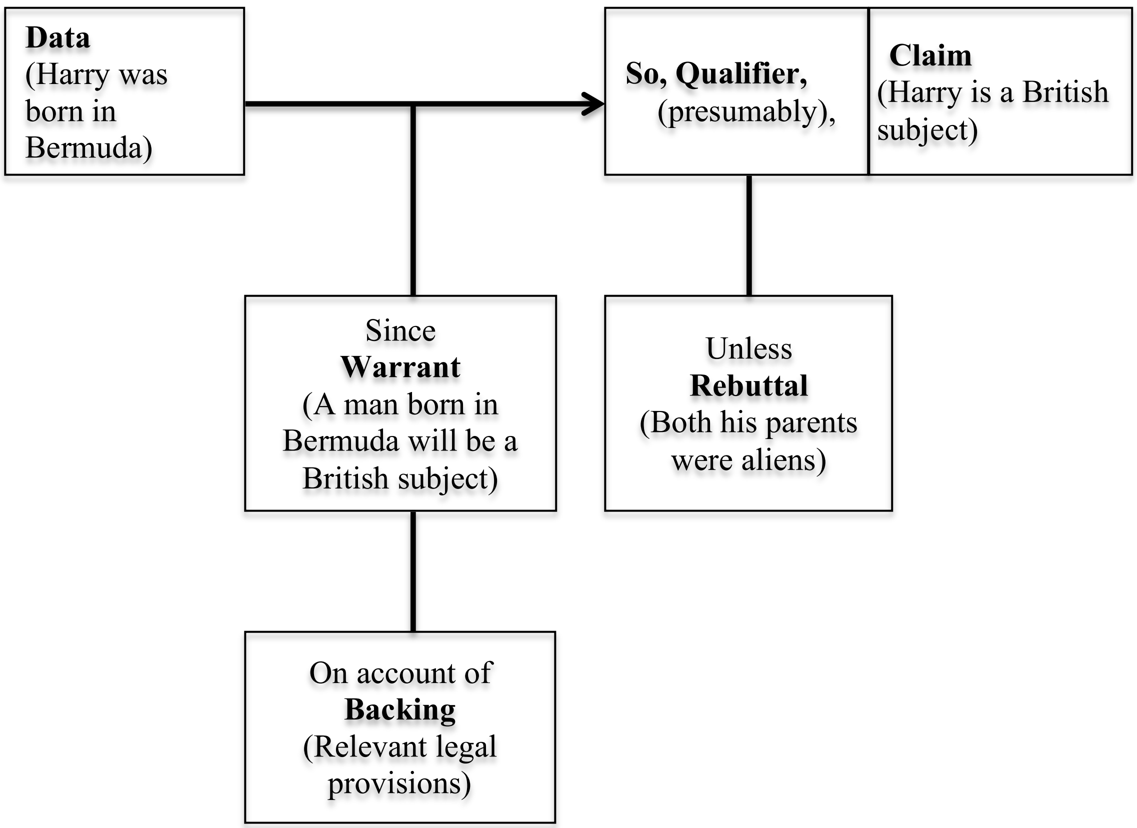


Figure 1.2. A Complex Skeleton of a Pattern for Analyzing Arguments (Toulmin, 2003)

1.2.2 Hypothetical Learning Progression for Scientific Argumentation

Based in Toulmin's model, Osborne and colleagues (2012) have proposed a learning progression for scientific argumentation, as shown in Table 1.1, which consists of three main levels with respect to the cognitive demand. A detailed description of the hypothetical learning progression is presented in Appendix A.

At the lowest level, Level 1, elements of argument are being identified or made, but none of these are being coordinated to form a complete argument. Students at Level 1a are capable of making a claim and students at Level 1b can correctly identify a claim. Level 1c requires the ability to first identify the claim and then the reasons that are being used to support the claim.

Level 2 builds on Level 1, as it requires a student to construct a relationship between a claim and the reasons that support it. At Level 2a, students are able to construct an argument with data or reasons. This level requires the ability to make a claim, select reasons or data that might support that argument, and then construct a synthesis between the claim and the warrant. Alternatively, Level 2b requires the ability to identify why a given argument is flawed and construct a rebuttal.

Level 3 is most advanced and requires the coordination of two or more arguments. Students at Level 3a can make a comparative argument. This ability relies on being able to identify why other arguments are potentially flawed, and requires a process of evaluating which is the most likely or best explanation. Moreover, not all evidence is equally important. Level 3b requires the ability to identify how each piece of evidence may or may not support an argument, to judge the quality of each piece of evidence, and to evaluate the relative importance of one piece of evidence with another. Last, Level 3c requires the ability to construct a counter claim. This is the most proficient level, since it requires all of the capabilities. Specifically, it requires the individual to construct an alternative explanatory hypothesis, compare it with the existing argument, and evaluate why it is superior. The latter requires a reason for why it offers greater explanatory coherence and to refute alternative theories or arguments.

Table 1.1. The Hypothetical Learning Progression for Scientific Argumentation (Osborne et al., 2012)

Level	**Constructing Arguments**	**Critiquing Arguments**
1a	Making a claim	
1b		Identifying a claim
1c		Identifying data or a reason supporting a claim
2a	Constructing an argument with data or reasons	
2b	Constructing a rebuttal	(Identifying weak arguments)
3a	Making a comparative argument	
3b		Identifying relative significance of several pieces of evidence
3c	Constructing a counter claim with justification	

1.3 Item Development

To empirically investigate the validity of the hypothesized scientific argumentation learning progression, a set of assessment items were designed to elicit student responses indicative of the hypothesized progression levels. The set of assessment items studied in this dissertation are contextualized in a specific content topic within the Structure of Matter domain, i.e., Changes of State. As described previously, it is widely recognized in the literature that, reasoning ability is dependent on domain-specific knowledge. A major challenge in item development is thereby how to assess a student's reasoning ability separately from his or her content knowledge. As Osborne and colleagues described,

> *"Our attempted solution was to write items in the domain specific area of "Structure of Matter," that were content lean in that they did not require a student to recall specific elements of appropriate domain knowledge. Granted, it is impossible to write items that are totally content-independent for even the ability to read text requires appropriate background knowledge. Nevertheless, by seeking to minimize the domain-specific knowledge by providing much of the knowledge either in the opening scenario or the stem of the question, we sought to emphasize students' ability to engage in reasoning in a scientific context." (Osborne et al., 2012, p.5).*

Like other performance assessments attempting to measure higher-order thinking skills, the argumentation assessment in this study consists of different tasks, for which a setting is established and multiple questions related to the setting are presented subsequently. The nature of questions ranges from identifying a claim to appraising competing arguments. Figure 1.3 presents one example of the argumentation items, called *Sugar in the Water*. This item presents a scenario that is contextualized in the Changes of State content area. Specifically, three hypothetical students in the scenario make different claims about what happens to sugar grains when the sugar is added to a cup of hot water. Five pieces of evidence that may or may not support these claims are provided in the prompt. Following the prompt, four questions are presented, which ask students to identify what evidence best supports or challenges a particular claim given in the prompt and justify their answers. Rosenbaum (1988) introduced the term "*item bundle*" to denote a subset of items that share a common stimulus. In the remainder of the text, items sharing the same scenario will be referred to as item bundle.

While each of the three primary levels in the hypothetical learning progression contains multiple sub levels, e.g., Level 1 contains sub levels 1a, 1b, an 1c, currently the efforts of assessment development and progression validation focus on the primary levels. That is, each argumentation item is designed to assess one of the three main levels, i.e., Level 1, 2, or 3, in the hypothesized scientific argumentation learning progression.

To evaluate whether the developed assessment items were clear and capable of eliciting relevant student responses as desired, researchers conducted think-aloud interviews (Wilson, 2005) with a group of grade eight students in a local district. Students were asked to explicitly describe their mental procedure of answering each item. Feedback on item design, such as clarity of the text and figures, was also collected. In addition, teacher feedback on the clarity and utility of the items was also obtained. Item revision was made subsequently according to interview results and teacher feedback.

Item A2. What Happens to Sugar in the Water?

Three students are discussing what happens to sugar grains when the sugar is added to a cup of hot water.

Paul says: *The sugar breaks up and disappears for good.*

Mary says: *The sugar breaks up but is still there and has not changed.*

Laura says: *The sugar has changed to make a new substance.*

They have several pieces of evidence for their argument.

1. The water tastes sweet.
2. After stirring the sugar can no longer be seen.
3. They leave the cup for several days. All the liquid evaporates leaving a sticky, solid substance at the bottom of the cup. This tastes sweet but is one large lump rather than separate grains.
4. They weigh the cup of water and the sugar separately. The weight of the cup of water and the sugar added together is the same as the weight of the cup of water with the sugar added.
5. They have been told that you cannot destroy or make new matter.

5a. Which evidence do you think best supports Paul's argument?
(a) Number ____
(b) Explain why you think this evidence supports Paul's argument.

5b.Which evidence challenges Paul's argument?
(a) Number____
(b) Explain why you think this evidence challenges Paul's argument.

5c. Which piece of evidence best supports Mary's argument?
(a) Number ____
(b) Explain why you think this evidence supports Mary's argument.

5d. Laura claims that the evidence available is not enough to decide whether she or Mary is right. Do you agree? Why?

Figure 1.3. The "Sugar in the Water" Argumentation Item Bundle

1.4 Outcome Space

To differentiate the quality of student responses with respect to the argumentation proficiency demonstrated, an iterative procedure was employed to develop a scoring guide for each item. Scoring guides were first drafted, which described and characterized expected student responses with respect to the proficiency level they suggested. These scoring guides were then moderated by researchers and employed to score a sample of empirical student responses for each item from trials. Revision was undertaken subsequently if a scoring guide was unclear or difficult to use empirically.

As an example, Figure 1.4 presents the final scoring guide of the item 5b in the "*Sugar in the Water*" item bundle (see Figure 1.3). This item asks students to identify a piece of evidence that challenges the claim made by a hypothetical student, Paul, in the prompt, and provides a justification for how the selected piece of evidence challenges Paul's claim. As described previously, each argumentation item is designed to assess one of the three main levels, i.e., Level 1, 2, or 3, in the hypothesized scientific argumentation learning progression. This particular item is developed to assess the Level 2 ability in the hypothesized argumentation learning progression, which is to identify evidence that contradicts a given claim and build an alternative argument about why that claim may be flawed. Note that partial credit was given to responses, where reasoning is largely yet not entirely correct. Responses with full credit, i.e., code 2 in the current example, are regarded as fully demonstrating the proficiency level an item is designed to assess, i.e., Level 2 in this example. Alternatively, responses with partial credit, i.e., code 1 in the current example, are considered as demonstrating the proficiency level partially. This logic applies to the scoring of all items.

Code	**Descriptions**	**Mock Student Responses**	**Empirical Student Responses**
2	Correct answer with appropriate reasoning. *A correct piece of counter-evidence is identified and reasoning is provided as to how it counters Paul's claim.*	"Number [anything but 2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any piece of evidence besides #2] suggests that something remains."	[PICKED 5] "Paul said the sugar was gone for good but due to the law of conservation, the sugar is still there just that it has dissolved." (Note that this response tied this evidence back to Paul's claim)
1	Reasoning is largely correct, i.e., correct answer with inadequate reasoning or good reasoning that is used to justify an incorrect answer. *A correct piece of counter-evidence is identified, but reasoning is not provided as to how it counters Paul's claim.* *<u>Alternatively,</u>* *A correct piece of counter-evidence is not identified, but at least reasoning is provided as to how the student thinks it counters Paul's claim.*	"Number [anything but 2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that does not justify why the selected evidence does not support Paul]." *<u>Alternatively,</u>* "Number [2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that tries to justify why the selected evidence does not support Paul]."	[PICKED 5] "Because you cannot destroy or make new matter." (Note that this response did not tie this back to Paul's claim) [PICKED 1] "Followed by what Paul said, the sugar breaks up and mixed with the water." [PICKED 4] "Because number 4 said that the sugar is still there."
0	Answer fails to address the question. *The student not only fails to identify a piece of evidence that challenges Paul's argument, but no justification is provided for why in fact they believe the piece of they chose counters Paul.*	"Number [2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that does not justify why the selected evidence does not support Paul]."	[PICKED 2] "Because Paul says the sugar breaks apart and disappears, not stirring the water." [PICKED 2] "Tells more details, and makes his statement sort of clear."
B	No Response		

Figure 1.4. The Scoring Guide of the Item 5b in the "Sugar in the Water" Item Bundle

1.5 Measurement Model

The fourth building block is represented by measurement models. Through a measurement model, scored student responses were calibrated and related back to the proposed cognition model (Wilson, 2005). There have been many measurement models proposed and used in the field of educational and psychological testing. In describing the philosophy of the construct modeling approach (i.e., the BAS approach), Wilson (2005) focused on the special role the Rasch models (Rasch, 1960) play in understanding a construct of interest. For a detailed account of Rasch models, please see Chapter 2.

In addition to the Rasch models, other types of measurement models (or more generally, latent variable models) were adopted to investigate the validity of the scientific argumentation assessment in this dissertation. Specifically, Chapter 2 adopted Rasch and factor models, Chapter 3 adopted latent class models, and Chapter 4 adopted multidimensional item response models. For a detailed account of each type of model, please refer to each chapter.

1.6 Validity

Among other psychometric properties, evidence on the validity of the investigated scientific argumentation assessment is of particular interest. As defined in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999, p.9), "*validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests*." The *Standards* has described various kinds of evidence that might be collected to validate the proposed interpretation of a test score in light of the purpose of testing, and also to provide implications for test development. From Chapter 2 to Chapter 4, various measurement models were applied to examine various sources of validity evidence for the usage of the studied scientific argumentation assessment.

Internal Structure Validity. According to the *Standards*, analyses of the internal structure of a test provide empirical evidence on *"the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based*" (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999, p.13). The argumentation items were designed to measure a unidimensional scientific argumentation construct. It is of interest to investigate if there is evidence for unidimensionality, and the degree to which the empirical pattern of item difficulties matches with the hypothesis.

External Validity. External validity evaluates the degree to which the relationships of performance on the investigated test to variables external to the test are consistent with the hypothesis drawn on a theory or past research. Different kinds of external variables may be used, such as tests measuring related or different constructs, measures of some criteria the test is expected to predict, and so on (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999, p.13). As Walton (1990) suggested, the competency to argue is dependent on knowledge. It is thereby expected that performance on the investigated scientific argumentation and content assessments are two distinct yet correlated variables.

1.7 Data

1.7.1 Instrument

In this data set, 20 argumentation items that were embedded in four scenarios contextualized in the Changes of State topic in grade eight physical science were investigated. Table 1.2 provides a summary of the information about the argumentation items. Appendix B presents all argumentation items used in the current study, and Appendix C presents the corresponding outcome space of these argumentation items. To investigate the relationship between content knowledge and argumentation ability, 12 traditional items measuring Changes of State content knowledge were administered along with the argumentation items to the same sample.

Table 1.2. Information of the Argumentation Items

Item #	Bundle	Item Name	Progression Level
1	A1.	A1-3a	1
2	Desalinating Water	A1-3b	1
3		A1-3c	1
4		A1-3d	1
5		A1-3e	3
6	A2.	A2-5a*	1
7	Sugar in the Water	A2-5b*	2
8		A2-5c*	1
9		A2-5d	3
10	A3.	A3-14aa	1
11	Ice to Water Vapor	A3-14ab	1
12		A3-14ac	2
13		A3-14ba	1
14		A3-14bb	1
15		A3-14bc	2
16		A3-14c	2
17	A4.	A4-7a	2
18	Something in the Air	A4-7b	2
19		A4-7c*	1
20		A4-7d	2

Note. Items with * are those with two parts: the first part asks students to select one of the choices given in the item prompt and the second part asks students to provide an explanation. One overall score was given and used in the subsequent analysis.

As described in the previous section, partial credits were given for most of the argumentation items, and thus the original data are mixed of dichotomous and polytomous items. For Chapters 2 and 4, the original polytomous data were used. However, the analysis in Chapter 3 requires that each score category should align with a particular learning progression level, and thus does not allow the existence of partial-credit categories that represent responses “partially” demonstrating performance expected at a progression level. The data thus have to be dichotomized for Chapter 3. For details of this procedure, please refer to Chapter 3.

1.7.2 Sample

The 20 argumentation items were administered in spring 2011. The data used in this dissertation were collected from 347 grade eight students in a local school district, who attempted at least one item in each bundle. Table 1.3 presents background information of the 342 students in the sample. Information about five students is unavailable.

Table 1.3. Background Information of the Sample (n=342)

Variable	**Frequency**	**Percentage**
Gender		
Female	188	54.97
Male	154	45.03
2011 California Standards Test: English-Language Arts Proficiency		
Far Below Basic	6	1.75
Below Basic	15	4.39
Basic	64	18.71
Proficient	86	25.15
Advanced	171	50.00
2011 California Standards Test: Science Proficiency		
Far Below Basic	9	2.63
Below Basic	20	5.85
Basic	57	16.67
Proficient	61	17.84
Advanced	195	57.02

Chapter 2: Validity Evidence from Factor Analysis and Rasch Modeling

The purpose of this chapter is to obtain evidence of internal structure validity for the investigated argumentation items using exploratory factor analysis and Rasch modeling approach. The argumentation items were designed to measure a unidimensional scientific argumentation construct. It is of interest to investigate if there is evidence for unidimensionality, and the degree to which the empirical pattern of item thresholds matches with the hypothesis. Other psychometric properties, such as test reliability and item discrimination, were also examined to evaluate the function of the items.

2.1 Exploratory Factor Analysis

2.1.1 Theoretical Background

To investigate whether the argumentation assessment contextualized in the Changes of State content area is indeed unidimensional, exploratory factor analysis (EFA) was performed to evaluate the dimensionality of the assessment. The main purpose of EFA is to discover and identify a set of latent common factors (i.e., dimensions) that account for the correlations among the manifest variables (Mulaik, 1987). EFA can be performed with various procedures. Numerous criteria have also been proposed for determining the number of factors to retain in the EFA framework. In reality, an assessment is never strictly unidimensional. The interest of this study is to identify major and salient factors. In a study of dimensionality assessment of polytomous items, Timmerman and Lorenzo-Seva (2011) suggested that the parallel analysis with minimum rank factor analysis performed consistently well at identifying the number of major factors, when minor factors were present in the simulated data. Hence, in the current study, the factor models were estimated with the minimum rank factor analysis (ten Berge & Kiers, 1991), and the parallel analysis was used as the main technique to determine the number of factors to retain.

In a common factor analysis, the decomposition of the covariance of the responses is as below,

$$\mathbf{\Sigma} = \mathbf{\Sigma_c} + \mathbf{\Psi} \tag{2.1}$$

, where $\mathbf{\Sigma}$ is the covariance matrix of the responses, $\mathbf{\Sigma_c}$ is the variance-covariance matrix due to the common factors, and $\mathbf{\Psi}$ is a diagonal matrix of unique variances. The ideal of factor analysis is to find a decomposition of $\mathbf{\Sigma}$ with $\mathbf{\Sigma_c}$ of low rank r. For more details of factor analysis, please see Skrondal and Rabe-Hesketh (2004). Built upon the notion that the ideal of low reduced rank is hardly attained in practice (Shapiro, 1982), the minimum rank factor analysis (MRFA; ten Berge & Kiers, 1991;Shapiro & Ten Berge, 2002) decomposes the covariance matrix of the responses as below,

$$\mathbf{\Sigma} = \mathbf{FF}' + (\mathbf{\Sigma_c} - \mathbf{FF}') + \mathbf{\Psi} \tag{2.2}$$

, where variances of responses (diagonal elements of $\mathbf{\Sigma}$) are decomposed into explained common variances (diagonal elements of $\mathbf{FF}'$), unexplained common variances (diagonal elements of $\mathbf{\Sigma_c} - \mathbf{FF}'$), and unique variances (diagonal elements of $\mathbf{\Psi}$). Subject to the constraints that $\mathbf{\Sigma_c} - \mathbf{FF}'$ and $\mathbf{\Psi}$ are both positive semidefinite, MRFA minimizes the unexplained common variance for any fixed number of factors r.

After EFA is performed, the subsequent question is how many factor(s) should be retained. Popular criteria include Kaiser's eigenvalue-over-one criterion (Kaiser, 1960), the scree plot (Cattell, 1966), the goodness-of-fit measures (Browne & Cudeck, 1992), and parallel analysis (Horn, 1965). In simulation studies that included major and minor factors, parallel analysis has been shown to perform well at indicating the number of major factors (e.g., Humphreys & Montanelli, 1975; Zwick & Velicer, 1986; Timmerman & Lorenzo-Seva, 2011). The original rationale of the parallel analysis is that the eigenvalues of the true factors extracted from empirical data in an EFA should be larger than the eigenvalues of the factors derived from simulated random data (Horn, 1965). That is, only factors with empirical eigenvalues larger than a given threshold in the distribution of random eigenvalues should be retained. The mean and the 95th percentile are both common cut-off thresholds. In the PA-MRFA procedure (Timmerman & Lorenzo-Seva, 2011) used here, the proportions of explained common variance, rather than eigenvalues, of successive factors extracted from empirical data are compared to that obtained from randomly generated data.

2.1.2 The EFA Procedure

In the current study, the PA-MRFA procedure described by Timmerman and Lorenzo-Seva (2011) was performed with the software FACTOR (Lorenzo-Seva & Ferrando, 2006). The number of factors to be retained is indicated by the number of factors with the empirical proportions of explained common variance larger than a given threshold in the distribution of the simulated proportions of explained common variance. The mean threshold appeared to yield a tendency to overestimate the dimensionality, and thus the use of a more stringent criterion, such as the 95 percentile, was advocated in the literature (e.g., Buja & Eyuboglu, 1992). The goal of the current study is to identify salient factors, and thus the 95 percentile was used to determine the number of factors. Permutation of the raw data has been recommended in the literature to obtain sampling distributions of eigenvalues (Buja & Eyuboglu, 1992). 500 random Pearson correlation matrices were generated with the permutation of the sample values. To evaluate the number of factors, the scree plot (Cattell, 1966) was also examined.

2.1.3 Results from the Exploratory Factor Analysis

Figure 2.1 is the scree plot of the eigenvalues of the reduced correlation matrix. Please see Timmerman and Lorenzo-Seva (2011) for details. The plot shows that a sharp bend occurs around the second and the third eigenvalue. Although the first two factors both have eigenvalues larger than one, the eigenvalue of the first factor is about three times as large as the eigenvalue of the second factor and the eigenvalue of the second factor is only slightly larger than one. These evidence point to the presence of one salient factor.

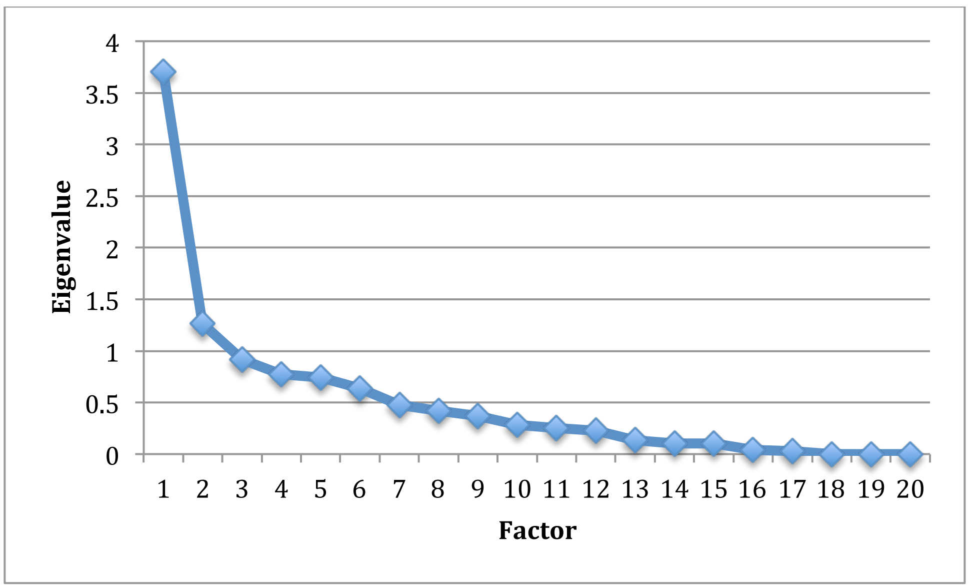


Figure 2.1. Scree Plot

The presence of one salient factor is further observed in the results of parallel analysis (Timmerman & Lorenzo-Seva, 2011). Following the procedure described in Timmerman and Lorenzo-Seva (2011), Figure 2.2 compares the proportions of explained common variance of successive common factors of empirical data with the 95 percentile in the proportion of explain common variance distributions obtained from randomly generated data. The number of factors indicated is the number of factors with the proportion of explained common variance from the empirical data larger than the 95 percentile from the randomly generated data. The comparison points to the one-factor solution.

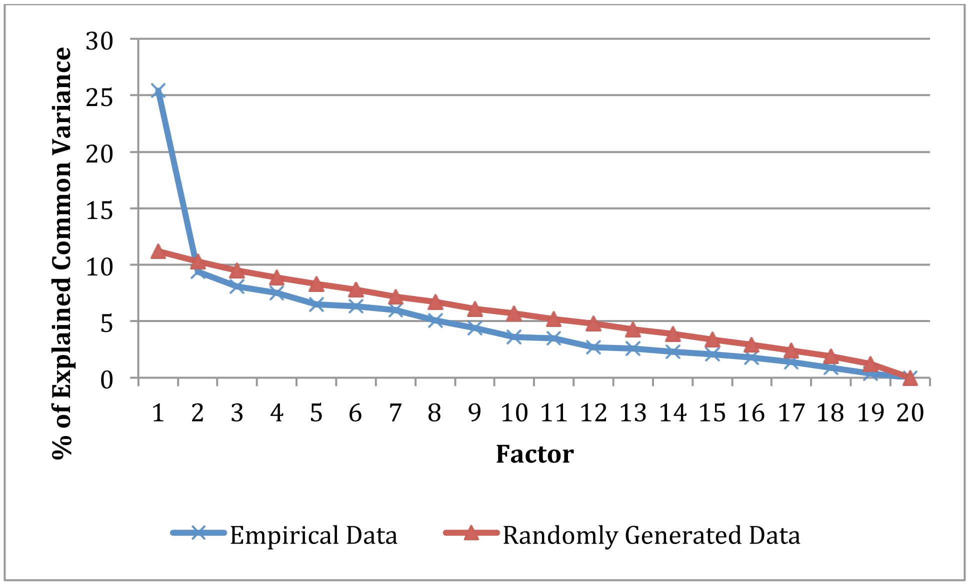


Figure 2.2. Results of Parallel Analysis

The value of the root mean square of residuals was further checked to evaluate the fit of the one-factor model. The values of the root mean square residuals should be below 0.08, with lower values indicating better model fit (Hu & Bentler, 1999). The one-factor model has the value of the root mean square of residuals equal to 0.08, which suggests reasonable fit.

The aforementioned results provide statistical evidence for the structural validity of the investigated argumentation items that, these items measured one salient latent construct as intended. Table 2.1 presents the pattern of factor loadings for the 20 argumentation items. As a common criterion, 0.3 is used to determine the strength of the relationship between an item and a factor. To make the factor loading pattern more readable and interpretable, factor loadings with absolute values lower than 0.3 are reported in gray scale. As shown in Table 2.1, all of the argumentation items load positively on the factor as desired. Seven out of the 20 argumentation items have loadings smaller than 0.3.

Table 2.1. Factor Loadings Obtained from the One-Factor Model

Variable	Factor Loading
A1_3a	0.23
A1_3b	0.28
A1_3c	**0.39**
A1_3d	**0.53**
A1_3e	**0.42**
A2_5a	**0.59**
A2_5b	**0.50**
A2_5c	0.29
A2_5d	0.22
A3_14aa	**0.64**
A3_14ab	**0.49**
A3_14ac	**0.54**
A3_14ba	**0.41**
A3_14bb	0.13
A3_14bc	**0.54**
A3_14c	0.28
A4_7a	**0.30**
A4_7b	0.28
A4_7c	**0.52**
A4_7d	**0.54**

2.2 Rasch Analysis

2.2.1 Theoretical Background

Rasch models are commonly used as confirmatory tools to examine the measurement properties of an instrument. Rasch models provide information of the extent to which items meet the fundamental measurement criteria: (a) items in a test measure one latent construct (i.e., assumption of unidimensionality), and (b) response to an item is independent of the response to any other item in a test, after the latent ability the test attempts to measure is controlled for (i.e., assumption of local item independence). Item statistics derived from a Rasch model are useful to detect items that do not function as desired.

Rasch analysis is chosen in this study because of its desired property of specific objectivity. In Rasch analysis, a raw score is a sufficient statistic for the ability estimate. If a Rasch model fits the data well, the calibrated measurement scale is analogous to a linear ruler that is invariant across examinees. The ability estimate derived from a Rasch model is thereby easier to interpret and understand for test users without a psychometric background, compared to two-parameter or three-parameter measurement models.

2.2.2 Model Formulation

In this study, the data is obtained from a mix of dichotomous and polytomous items. Thus, the Rasch partial credit model (Masters, 1982) was used as a confirmatory model to evaluate the measurement properties of the instrument as a whole and of each individual item. Under the Rasch partial credit model, the probability of student p scoring x on item i can be formulated as follows:

$$P_{pix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x}(\theta_p - \delta_i - \tau_{ij})\right]}{\sum_{r=0}^{m_i}\left[\exp\sum_{j=0}^{r}(\theta_p - \delta_i - \tau_{ij})\right]}, x = 0,\ldots,m_i \text{, where} \sum_{j=0}^{0}(\theta_p - \delta_i - \tau_{ij}) \equiv 0 . \qquad (2.3)$$

θ_p stands for the latent ability of student p. δ_i represents the overall difficulty of item i, and τ_{ij} refers to the additional step parameter associated with category j of item i.

The partial credit model was estimated using the software ConQuest (Wu, Adams, Wilson, & Haldane, 2007) using the marginal maximum likelihood estimator (Gauss-Hermite Quadrature with 15 nodes), where ability parameter θ_p is assumed to be normally distributed as $N(\mu_\theta, \sigma_\theta^2)$. Item parameters, δ_i and τ_{ij}, are assumed to be fixed parameters. For model identification, the mean of the ability distribution is fixed as 0 and $\sum_j \tau_{ij} = 0$ for each item i.

2.2.3 Results from the Rasch Analysis

Cronbach's Alpha was estimated as 0.78. The maximum likelihood estimate (MLE) person separation reliability was estimated as 0.81, and the Expected A Posteriori (EAP) reliability was estimated as 0.82. These evidence suggested reasonable internal consistency among the items.

Item Fit Statistic

There is more than one method to investigate whether the mathematical model of choice is appropriate. The weighted mean square statistic was used to evaluate the fit between the partial credit model and data on each item, which measures the degree of discrepancy between the expected and observed item scores. The weighted mean square statistic has an expected value of one. Items with values of the weighted mean square statistic larger than one are those with observed variance greater than the expected, and contribute less toward the overall estimation of the latent variables. Statistically, the weighted mean square statistic is a chi-square statistic divided by degree of freedom. Conventionally, the weighted t statistic is used to test the statistical significance of the mean square (Wright & Masters, 1981). Items that have a weighted mean square statistic outside the range of 0.75 and 1.33 and have the absolute value of the weighted t statistic larger than 1.96 should be considered as problematic (Wilson, 2005; p.129). All of the investigated argumentation items have weighted mean square statistics within the range of 0.75 and 1.33, which provides support for the use of the partial credit model.

Item Discrimination

Students' performance on a particular item is expected to be consistent with their performance on the instrument as a whole. That is, students that demonstrate higher proficiency level on the latent construct should also score higher on each item than their counterparts. To check this consistency between each single item and the whole instrument, Wilson (2005) recommended examining the mean ability of students within each score category for a given item. Specifically, the mean ability of students within each score category is expected to increase as the score increases. All of the studied argumentation items met this expectation, except for item A2_5d. Figure 1.3 in Chapter 1 presents the prompt of the item bundle A2, and Table 2.2 below presents the information about each score category and the average ability estimate of the students within each score category. As shown in the table, the average ability estimate increases as the score increases for item A2_5d, except for the category of score 2. For this item, categories of score 1 and score 2 are responses that earned partial credits. According to the scoring guide, responses as described in the category of score 2 (i.e., recognize uncertainty but DO NOT provide reasoning) got a higher score than responses as described in the category of score 1 (i.e., provide reasoning but DO NOT recognize uncertainty), because recognizing uncertainty in evidence is considered to be more sophisticated than providing reasoning to back a claim on the hypothetical argumentation progress map. The empirical evidence, however, shows that the average ability of the students that scored 2 (n=80) on this item is lower than the average ability of the students that scored 1 (n=82). The reason why item A2_5d did not show as displaying large misfit statistic is probably because there are only 80 students in the misbehaving response category (i.e., score 2). Nevertheless, the disordering in the average ability estimates of the students across response categories manifested on the weighted mean square statistic. Specifically, the weighted mean square statistic of item A2_5d is 1.21, which is closer to the upper boundary of the acceptable range (i.e., 1.33). When the value of the weighted mean square statistic is larger than one, it suggests that the observed residual variance is greater than the expected. This is the only argumentation item that assessed students' ability to recognize uncertainty in evidence. More items of this kind are needed in the future to investigate whether recognizing uncertainty is indeed more sophisticated than providing reasoning to back a claim as hypothesized.

Table 2.2. Information about Each Score Category of Item A2_5d

Score	Descriptions	Mean Ability (logit)	Number of Students
3	Recognize uncertainty AND provide reasoning. *The student recognizes that the evidence is not conclusive and makes a case for why this uncertainty does not allow them to side with either Laura or Mary*	0.38	28
2	Recognize uncertainty but DO NOT provide reasoning. *The student recognizes that the evidence is not conclusive, but does not make a case for why this uncertainty does not allow them to side with either Laura or Mary.*	0.05	80
1	Provide reasoning but DO NOT recognize uncertainty. *Here students, failing to recognize the uncertainty in the evidence, conclusively side with either Laura or Mary, but at least they provide a plausible reason why they chose to side with Laura or Mary.*	0.30	82
0	DO NOT provide reasoning/ DO NOT recognize uncertainty. *The student not only fails to see the uncertainty in the evidence, but they are unable to correctly reason for the claim (i.e., side with either Laura or Mary) in which they feel the evidence is indeed sufficient.*	-0.27	157

The software ConQuest provides a discrimination index for each item, which is the product-moment correlation between the proportion of the maximum score that student p achieved on item i (i.e.,p_{pi}) and the sum of the proportions of the maximum score achieved by student p (i.e.,$p_p = \sum_i p_{pi}$) for all students (Wu et al., 2007;Adams & Wu, 2002). In the case of dichotomous items, this discrimination index is the usual point-biserial correlation. The values of this more traditional discrimination index were also checked to examine the function of each item. Items with values of the discrimination index larger than 0.40 are regarded as having good discriminating power, while items with values smaller than 0.20 are regarded as having poor discrimination power (Ebel, 1965). Only item A3_14bb has a value of the discrimination index below 0.20. Referring back to the pattern of factor loadings in Table 2.1, we can find that this item also has the smallest factor loading. These two pieces of evidence suggest that item A3_14bb performed poorly on measuring the latent argumentation proficiency of interest.

Wright Map

Figure 2.3 presents a Wright map, which displays the student ability distribution and item thresholds on the same logit scale. The second column in the figure represents the distribution of the estimated student ability from most able at the top to least able at the bottom of the map, with mean equal to 0 and variance equal to 0.39. Each "x" represents approximately 2.1 students and the location of "x" indicates the level of ability estimate. The third to the eighth column present the item Thurstone thresholds (Wilson, 2005), which are arranged with respect to the progression level an item is designed to measure. As described in Chapter 1, partial credit was awarded to responses that were regarded as indicating partial performance at a targeted progression level. As an example, Figure 1.4 in Chapter 1 presents the final outcome space of the item 5b in the "*Sugar in the Water*" item bundle (see Figure 1.3). This particular item is developed to assess Level 2 ability in the hypothesized argumentation learning progression. Responses with full credit, i.e., code 2 in the current example, are regarded as fully demonstrating the proficiency level an item is designed to assess, i.e., Level 2 in this example. Note that responses with partial credit, i.e., code 1 in the current example, are considered as demonstrating Level-2 performance partially and should not be interpreted as demonstrating performance at Level 1. While the rationale of this scoring scheme is common in practice, response categories awarded partial credit do not represent performance at a particular progression level, and only the item thresholds of the full-credit response categories are clearly interpretable with respect to the learning progression level.

Figure 2.4 presents the Wright map with only Thurstone thresholds of the full-credit response category for each item, which were extracted from Figure 2.3. The jth Thurstone threshold represents the point, where the probability of scoring below j is equal to the probability of scoring j and above. In the current case, all item thresholds shown on the Wright map are the full-credit Thurstone thresholds. If the ability estimate of a student and the full-credit Thurstone threshold of an item are located at the same position on the map, the probability of this student obtaining full credit on the item is 0.5. More information about a student's proficiency can be obtained if locations of student ability and item thresholds are close. If items are too easy or too difficult for the student, very little information is gained. As a result, an ideal measurement situation is that the item threshold distribution covers the span of the student ability distribution, so that appropriate ability estimates for students of all performance levels can be obtained. The map suggests that there is a lack of items at the lower end and easier items are needed to avoid the issue of a floor effect.

Each item is hypothesized to measure one specific learning progression level. Items measuring a lower progression level are expected to have lower difficulty levels, compared to items measuring a higher progression level. Ideally, item thresholds tapping the same progression level will locate at similar positions on the Wright map and form a band, which represents the difficulty level of the progression level. The empirical ordering of the progression levels can then be compared to the theoretical ordering of the hypothesized learning progression. Examining Figure 2.4, the expected ordinal structure is clearly not present.

logit	Ability	Item Thresholds					
		Partially Level 1	Level 1	Partially Level 2	Level 2	Partially Level 3	Level 3
			A3_14bb.2 A1_3c.2				
2							
			A2_5c.2				
					A3_14c.2		
							A2_5d.3
				A3_14c.1			
	X				A4_7d.2		
	X				A3_14bc.2		
	X		A1_3a.2				
1	XX	A3_14bb.1					
	XX		A1_3b.2				
	XXX				A3_14ac.3		
	XXXXX						
	XXXXXX				A2_5b.2		
	XXXXXXXX	A1_3b.1					
	XXXXXXX					A2_5d.2	
	XXXXXXXXX						
	XXXXXXXX		A3_14ab.2				
	XXXXXXXXX			A3_14ac.2			
	XXXXXXXXX				A4_7b.2		
	XXXXXXXX		A4_7c.3 A2_5a.2				A1_3e.2
	XXXXXXXXXX			A3_14bc.1			
0	XXXXXXXXXX		A3_14aa.2	A3_14ac.1			
	XXXXXXXX	A4_7c.1 A4_7c.2	A1_3d.2		A4_7a.2	A2_5d.1	
	XXXXXXXX	A3_14aa.1					
	XXXXXXXX						
	XXXXXXX						
	XXXXXXXX	A3_14ab.1	A3_14ba				
	XXXXXX			A4_7d.1			
	XX						
	XXX						
	XXX						
	XXX						
	XX						
-1	XX	A1_3c.1		A4_7a.1		A1_3e.1	
	XX	A1_3d.1					
	X	A1_3a.1					
	X			A4_7b.1			
	X	A2_5a.1					
	X			A2_5b.1			
	X						
	X						
-2							
	X						
		A2_5c.1					

Figure 2.3. Wright Map with All Item Thresholds

logit	Ability	Item Thresholds		
		Level 1	Level 2	Level 3
		A3_14bb.2 A1_3c.2		
2				
		A2_5c.2		
			A3_14c.2	
				A2_5d.3
	X		A4_7d.2	
	X		A3_14bc.2	
	X	A1_3a.2		
1	XX			
	XX	A1_3b.2		
	XXX		A3_14ac.3	
	XXXXX			
	XXXXXX		A2_5b.2	
	XXXXXXXX			
	XXXXXXX			
	XXXXXXXXX			
	XXXXXXXX	A3_14ab.2		
	XXXXXXXXX			
	XXXXXXXXX		A4_7b.2	
	XXXXXXXX	A4_7c.3 A2_5a.2		A1_3e.2
	XXXXXXXXXX			
0	XXXXXXXXXX	A3_14aa.2		
	XXXXXXXX	A1_3d.2	A4_7a.2	
	XXXXXXXX			
	XXXXXXXX			
	XXXXXXX			
	XXXXXXXX	A3_14ba		
	XXXXXX			
	XX			
	XXX			
	XXX			
	XXX			
	XX			
-1	XX			
	XX			
	X			
	X			
	X			
	X			
	X			
	X			
-2				
	X			

Figure 2.4. Wright Map with Item Thresholds of the Full-Credit Response Categories

logit	Ability	Item Thresholds Level 1	Item Thresholds Level 2	Item Thresholds Level 3
		A1_3c.2		
2				
		A2_5c.2		
				A2_5d.3
	X		A4_7d.2	
	X		A3_14bc.2	
	X	**A1_3a.2**		
1	XX			
	XX			
	XXX		A3_14ac.3	
	XXXXX			
	XXXXXX		A2_5b.2	
	XXXXXXXX			
	XXXXXXX			
	XXXXXXXXX			
	XXXXXXXX	A3_14ab.2		
	XXXXXXXXX			
	XXXXXXXXX			
	XXXXXXXX	A4_7c.3 A2_5a.2		**A1_3e.2**
	XXXXXXXXXX			
0	XXXXXXXXXX	A3_14aa.2		
	XXXXXXXX	A1_3d.2		
	XXXXXXXX			
	XXXXXXXX			
	XXXXXXX			
	XXXXXXXX	A3_14ba		
	XXXXXX			
	XX			
	XXX			
	XXX			
	XXX			
	XX			
-1	XX			
	XX			
	X			
	X			
	X			
	X			
	X			
	X			
-2				
	X			

Figure 2.5. Wright Map with Item Thresholds of the Full-Credit Response Categories (Version Two)

One reason for this lack of clear structure may be that, some items did not perform well on measuring the common latent argumentation proficiency, and thereby brought "noise" due to construct vagueness into the Wright map in Figure 2.4. Five of the items that do not conform to the banding have small factor loadings (i.e., values of loadings no larger than 0.3): items A1_3b, A3_14c, A3_14bb, A4_7a, and A4_7b. Figure 2.5 presents the Wright map based on the smaller set of items, where the five items with small factor loadings are excluded. Although not perfect, an ordinal structure among locations of item thresholds associated with different progression levels emerges in Figure 2.5.

Level-1 items

First, all Level-1 item thresholds fall in the range between -0.47 logit and 0.41 logit on the Wight map, except for the thresholds of the following three items: items A1_3c, A2_5c, and A1_3a. Getting full credit on these three items appears to be much more difficult than getting full credit on the other Level-1 items.

An examination of item A1_3c reveals a potential influence of language on the difficulty level of the item. To illustrate, Figure 2.6 shows the scenario of the bundle A1, along with items A1_3c and A1_3d. Bundle A1 establishes a scenario with three paragraphs: the first paragraph presents the issue of California being short of fresh water, and each of the remaining two paragraphs presents a method of desalination to obtain fresh water from the oceans. Hypothetically, items A1_3c and A1_3d are expected to have similar difficulty levels, since they both ask students to identify an argument against a particular method of desalination. However, item A1_3c appears to be much more difficult than item A1_3d in the empirical data. One possible reason is the ambiguity of the language. Item A1_3d indicates clearly that the method students should identify an argument against is the one "*using membranes*," which helps students to locate the critical paragraph, i.e., the third paragraph, correctly. In contrast, item A1_3c refers the method students should identify an argument against "*the standard method of desalination*," the corresponding paragraph of which is the second paragraph. Through the text, however, "*the standard method of desalination*," is not specified explicitly.

An examination of item A1_3a provides another implication on the assessment of scientific reasoning. This item asks, "*what is the main claim that is being made in this piece.*" According to the scoring guide, students got full credit only if their responses referred to both (a) water shortage and (b) how the water shortage is to be addressed. 55% of the students mentioned only one of the two pieces in their responses, and were awarded partial credit. These pieces of evidence suggest that, identifying parts of a claim may be easy as hypothesized yet it is difficult to identify *all* pieces in a given claim. This finding calls for reflection on two general assessment issues: (a) what level of sophistication student responses should demonstrate to be considered as demonstrating performance at a particular learning progression level, and (b) how item design can be improved to prompt students to demonstrate their understanding and proficiency level to the fullest.

Osborne and colleagues (2012) have provided a discussion about why item A2_5c was surprisingly difficult. As shown in Figure 1.3 in Chapter 1, item A2_5c asks students to identify a piece of evidence that best supports the claim the hypothetical student Mary made and provide an explanation accordingly. Item A2_5a is identical to item A2_5c, except that item A2_5a is about identifying evidence for the claim the other hypothetical student Paul made. While by design these two items were expected to have similar difficulty levels, empirically it is much more difficult to get full credit on item A2_5c than on item A2_5a. An examination of the question shows that, only one of the five pieces of evidence given in the prompt could support Paul's claim, while four pieces of evidence could support Mary's claim. Osborne and colleagues (2012, p.11) argued that, "*in the case of Paul, the question merely demanded an assessment of which single element was the most significant piece [of] evidence for his claim. However, in the case of Mary [,] the question demanded the assessment of the relative significance of multiple pieces of evidence in relation to her claim.*" Osborne and colleagues (2012) then suggested that, *degree(s) of coordination*, might be an additional confound when one evaluated the difficulty of an argumentation task. Processing more pieces of information requires a higher degree of coordination to identify and link critical elements of an argument, and thus imposes an increasing cognitive demand.

Level-2 items

As to Level-2 item thresholds, the four Level-2 item thresholds cluster between the range of 0.70 logit and 1.30 logit, and locate higher on the Wright map than most of the Level-1 item thresholds do. This pattern matches with the expectation that, item thresholds targeting the same progression level have similar difficulty levels and item thresholds of a higher progression level should locate higher on the Wright map than those of a lower progression level. More Level-2 items will be needed in the future to obtain more evidence about this pattern.

Level-3 items

For Level 3, item A1_3e is surprisingly easy. Following the prompt in Figure 2.6, item A1_3e asks students to *"using only the information given above, which method do you think is the best one? Explain why you think this is the best method."* It was hypothesized that this item assessed a student's ability to evaluate and compare competing arguments. However, according to the scoring guide, students get full credit if they successfully give a claim, i.e., choose one of the two desalination methods, and a reason provided in the prompt. Cognitively, succeeding on this item then only requires the ability to identify a claim and a relevant piece of supporting evidence (Level 1c). This item is a good example of how the way an item is written may affect the cognitive demand it requires. Also, more Level-3 items will be needed for a further examination of the "banding" and average difficulty level of the Level-3 items.

Item A1. Desalinating Water

California is increasingly short of fresh water. Lack of fresh water limits the amount of food that can be grown. Clearly, there is a need to better manage this increasingly valuable resource. 97% of the world's water is in the oceans. The trouble is that it is salty and cannot be drunk. The process of removing salt from water is called *desalination*.

One method of desalination is to use energy to heat seawater so that it evaporates leaving the unwanted salt behind. The water vapor is then passed over a cold object where it turns back to pure water. A great deal of energy is lost during this process.

Another method involves pushing the seawater through a sheet with tiny holes in it called a membrane. Only the water particles pass through leaving all the salt on the other side. This uses half the energy but the membranes get clogged and have to be cleaned regularly. All the sludge produced then has to be taken away.

3c. What is the main argument against the standard method of desalinating water?
3d. What is the main argument against using membranes?

Figure 2.6. Items A1_3c and A1_3d

2.3 Conclusions

The main purpose of this chapter is to obtain evidence of internal structure validity for the investigated argumentation items using exploratory factor analysis and the Rasch modeling approach. Evidence collected and implications for future development are discussed below.

Test Dimensionality. The result of the exploratory factor analysis provides statistical evidence that, there is no violation of the unidimensionality and the investigated argumentation items measured one salient latent construct as intended.

Empirical Pattern of Item Thresholds. The Wright map obtained from the Rasch partial credit model serves as a graphical representation of the locations of item thresholds. Over half of the items have locations of thresholds that conform to the expected pattern: (a) item thresholds of a lower progression level have lower difficulty levels than item thresholds of a higher progression level, and (b) item thresholds of the same progression level locate at similar positions on the Wright map. The ordinal structure of these item thresholds shown on the Wright map (i.e., Figure 2.5) provides support for the hypothetical structure of the three learning progression levels.

Implications for Test Development. In addition to collecting validity evidence on the internal structure, the validation process also provides several implications for assessment development. First, the cognitive demand of an item may depend on how the item is written. In the case of this study, it is found that the amount of information a student needs to process to solve an argumentation item influences the difficulty of the item. Between two items that were designed to measure the same learning progression level, the item that requires students to process more pieces of information may impose an increasing cognitive demand and thus appears to be more difficult than the other item. Second, ambiguity of the language can also contribute to an unwanted increase in item difficulty level. Third, in the current study student responses that are regarded as being partially correct are considered to demonstrate partial performance at the targeted progression level. The distinction between performance "at" and performance "partially at" a progression level tends to be vague, and calls for reflections on two questions: (a) what level of sophistication student responses should be observed to be considered as fully demonstrating performance at a particular learning progression level, and (b) how item design can be improved to prompt students to demonstrate their understanding and proficiency level to the fullest.

Chapter 3: Validating the Argumentation Learning Progression with Latent Class Analysis

3.1 Background

Learning progressions have been advocated and used, both within the research community and in mainstream accountability systems, as means to foster coherence of assessment design (Alonzo & Gotwals, 2012). The main interest of using learning-progression-based assessments is to diagnose where students stand in their progression towards a sophisticated level of knowledge and/or skill. However, interpreting student performance on assessment tasks with respect to levels of a learning progression is challenging. As West and colleagues (West et al., 2012; p.257) pointed out, "*What is needed is a methodology that can be used to map assessment performance onto the levels, to combine information across multiple tasks measuring similar and related KSAs [knowledge, skills, and/or abilities]", to support inferences abut students, and to study how well actual data exhibit the relationships posited by the LP [learning progression]*." Despite the strong call for its use, the body of research on empirical validation of learning progressions is limited.

The goal of this study is twofold. Substantively, this study aims to examine whether student groups exhibited performance levels as defined in the hypothesized learning progression in scientific argumentation ability (See Chapter 1 for details of the learning progression). Specifically, this study employed latent class analysis (Lazarsfeld & Henry, 1968) to investigate whether hypothesized progression levels could be observed in student responses. The interest is to investigate if response patterns of students in each latent class resemble performance levels as defined in the progression. Methodologically, it is of interest to explore the utility, and challenges, of latent class analysis for validation of learning progressions. Note that the focus of this study is on learning progressions that describe student thinking at a specific time point, rather than longitudinal learning trajectories.

With limited research on validation of learning progressions using psychometric methods, methods and lessons learned from this study could inform research methodologies applied to the validation of learning progressions. Findings of this study will also contribute to literature on assessments of scientific argumentation ability.

3.2 A Latent Class Analysis Approach

3.2.1 Theoretical Background

Suppose the latent variable c is categorical with C classes (c=1,…,C). The unconstrained latent class model for dichotomous items[1]is formulated below (Lazarsfeld & Henry, 1968):

$$\Pr(\mathbf{y}_p) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} \pi_{i|c}^{y_{pi}} (1-\pi_{i|c})^{1-y_{pi}} \tag{3.1}$$

, where $\Pr(\mathbf{y}_p)$ denotes the probability of the response vector $\mathbf{y}_p$, π_c is the proportion of persons in the population in latent class c, and $\pi_{i|c}$ is the conditional probability of a person scoring 1 on item i given that the person belongs to class c.

Unlike the unconstrained latent class model, the ordered latent class model assumes that latent classes can be ranked as increasing progression levels of a latent trait. The formulation of the ordered latent class model (Croon, 1990) is the same as defined in Equation 3.1, with as a set of restrictions on class-specific response probabilities as below,

$$\pi_{i|c} < \pi_{i|c'} \text{, for } c < c' \text{ and all i} \tag{3.2}$$

, where persons in a latent class c' have a higher probability of a correct response than persons in a latent class c, and the probability of a correct response is equal for persons within the same latent class.

Ordered latent class analysis has recently been advocated and extended as an alternative approach for validity studies of learning progressions. A learning progression could be conceptualized as an ordered set of latent classes (Wilson, 2009), and this ordinal structure of the latent classes could be modeled explicitly with the ordered latent class model (Diakow, Irribarra, & Wilson, 2011). In addition to the structure of a single learning progression, in practice there may be further assumptions with respect to the relations among levels of two or multiple dimensions of learning progressions. There have been ongoing efforts to develop models that are capable of testing hypotheses of relations among progression levels. Such models are called Structured Construct Models (SCM; Wilson, 2009). The current study focuses on the validation of a learning progression with just a single dimension, and thus SCM is beyond the scope of the current study.

---

[1] All models in this study can be extended for the analysis of polytomous items.

3.2.2 Alignment between Latent Classes and Progression Levels

A prerequisite for applying latent class models to validating a learning progression is that each response category should correspond to a specific progression level. If a response category is not aligned with a specific learning progression level, the subsequent interpretation of latent classes with respect to learning progression levels will be jeopardized. In the original data, as described in Chapters 1 and 2, partial credit was given to responses that were regarded as indicating partial performance at a targeted progression level. Recall that responses that were awarded partial credit in this data could not be interpreted as demonstrating a specific performance at a level lower than the targeted level; instead, they were judged on the basis that they had only partly achieved the level. While the rationale of this scoring scheme is common in practice, response categories given this sort of partial credit do not represent performance at a specific progression level, and thus the originally-coded data are not suitable for the subsequent procedure of latent class analysis. To prepare for the latent class analysis, the data used in this chapter were dichotomized, such that each response category is representative of a specific progression level. For details of this data dichotomization, please refer to the *Data Preparation* section below (i.e., Section 3.3.1).

In the limited body of research on validation of learning progressions with latent class analysis, three means have been proposed to examine the association between empirically identified latent classes and hypothetical progression levels: (1) conditional probabilities, (2) expected item scores, and (3) ability distribution within each latent class. As will be described subsequently, expected item scores are essentially derived from the conditional probabilities. If all items are dichotomous, as the case of the data used in this chapter, the use of expected item scores is fundamentally not different from the use of conditional probabilities. Hence, in the subsequent sections only the conditional probabilities will be reported. In this chapter, I also investigated the use of Wright map to examine the characteristics of each latent class. In sum, the following three results were used to investigate characteristics of latent classes with respect to learning progression levels in this chapter: (1) conditional probabilities, (2) ability distribution within each latent class, and (3) Wright map.

1. Conditional Probabilities

Usually assessments associated with learning progressions are designed, such that each item response category corresponds to a hypothetical progression level. If a latent class resembles a group of students at a particular hypothetical learning progression level, students in the latent class are expected to give responses indicative of that progression level with high probabilities consistently in the whole test (Steedle & Shavelson, 2009). There is no established rule of thumb in the literature, however, with respect to the question of how large the conditional probability of giving a response representative of a progression level should be for a latent class to be regarded as resembling students at that level.

Building upon the use of conditional probabilities, Diakow and colleagues (2011) proposed the use of expected item score to examine the meaning of the ordered latent classes. For each item i in latent class c, the expected item score is calculated as follows:

$$E[y_{pi} \mid c] = \sum_k k\pi_{ik|c} \tag{3.3}$$

, where $\pi_{ik|c}$ is the probability of responding at score level k on item i for person p being in class c. Suppose responses to a dichotomous item suggest the following two types of performance: (1) incorrect responses, and (2) responses indicative of Level-2 performance. Then these two response categories will be scored as 0 and 2 respectively if expected item scores are to be calculated. Since each item score represents a level, the expected score could be directly compared to a progression level. The average expected score is then calculated among items that have the same maximum possible score. Note that if all items are dichotomous, the average expected score of items targeting the same progression level are essentially equal to the average of the conditional probabilities of the items targeting the same progression level multiplied by that level. That is, the average expected score of the Level-2 items of a latent class is essentially the average conditional probability of the Level-2 items multiplied by 2. Hence, as in the case of the data used in this chapter, the use of expected item scores is fundamentally not different from the use of conditional probabilities when data are dichotomous. The discussion of the use of expected item scores in the case of polytomous data is beyond the scope of this chapter. For a detailed explanation of the calculation and interpretation of expected item scores in the case of polytomous data, please refer to Diakow, Irribarra, and Wilson (2011).

2. Ability Distribution within Each Latent Class

Compared to models assuming that the latent variable of interest is continuous and providing proficiency estimates on a continuous scale, one advantage of latent class analysis is that it provides direct diagnostic information by grouping individuals into latent classes. Nevertheless, it is a common interest in the literature to assess how well latent classes separate persons with respect to a continuous proficiency scale. For this purpose, researchers would examine the distribution of proficiency estimates from an item response theory (IRT) model fitted to the same data within each latent class. In the validation of learning progressions, it would be expected that (a) ability distributions would be well separated across latent classes, and (b) there would be an increasing order in average abilities from a latent class resembling low progression level to a latent class resembling high progression level (e.g., Diakow *et al*., 2011).

3. Wright Map

As described in Chapter 2, the Wright map is a graphical representation tool that allows a direct comparison of person and item parameter estimates. In the context of learning-progression-based assessments, items are designed to elicit student responses indicative of hypothetical progression levels. Due to this design, each single item threshold can be linked to a hypothetical progression level.

In the past research on the validation of learning-progression-based assessments, the Wright map was most commonly used to investigate whether the empirical pattern of item thresholds provides evidence that supports the hypothetical structure of learning progression levels (e.g., Rivet & Kastens, 2012). Chapter 2 provides a full account of this investigation for the studied argumentation assessment.

The fundamental purpose of learning-progression-based assessments is to provide diagnostic information about a student's performance level. If the empirical pattern of item thresholds on the Wright map shows the hypothesized ordinal structure of the progression levels, one further step that can be taken is to set the cut-off points between the progression levels. Specifically, item thresholds and ability estimates obtained from the Rasch models are calibrated on the same scale and thus can be compared directly. Since item thresholds are aligned with specific progression levels, cut-off points between two adjacent progression levels can be set with the locations of the corresponding item thresholds. Then, students with ability estimates falling within the "zone" of an empirically-set learning progression level are classified into that level (Kennedy & Wilson, 2007). Details of how to set cut-off points with the item thresholds and how to interpret the classification with respect to progression levels were illustrated explicitly with the analysis output in the section of the results. Note that students with ability estimates falling in a particular learning-progression-level zone have an average probability of 0.5 of getting items targeting the same progression level correct, since ability levels of students and difficulty levels of items at the same progression level are regarded as equivalent (M. Wilson, personal communication, June, 2013). Examining the classification of students' progression levels from Wright map within each latent class provides another piece of information about students' proficiency level with respect to a continuous ability scale.

3.3 Methods

3.3.1 Data Preparation

As described in Chapter 1, the data were collected from 347 grade eight students in a local school district on 20 argumentation items, each of which is aligned with a particular level on the hypothesized scientific argumentation learning progression.

To meet the requirement that each response category represents a particular progression level, the following procedure was undertaken to prepare the data for latent class analysis. Using the results from Chapter 2, where the partial credit model was fitted, the average IRT ability estimate of students in the partial-credit response category was then compared to the average IRT ability of students in the full-credit category and also to the average IRT ability of students in the incorrect- response category—if the partial-credit ability estimate is close to one or the other of the other scores, then that might be taken as evidence that the partial credit category could be collapsed into the nearest one.

As an example, Table 3.1 shows the coding of response categories of the item A3_14ab and the resulted average ability estimate of the students in each response category. According to the original outcome space, this item has three response categories. The column "Code" presents the original codes of these categories, which range from 0 to 2. The third column, Score: PCM, shows the scores assigned to these response categories in fitting the partial credit model. In the current case, the scores assigned for the partial credit model are identical to the original codes. The fourth column, Average Ability: PCM, shows the average ability estimate of students in each response category obtained from the partial credit model. In the current example, the average ability of students in the partial-credit response category is very close to that in the full-credit category. This evidence provides support for collapsing the partial-credit and full-credit categories for this item. For further assurance, another model was fitted, in which partial-credit and full-credit categories were assigned the same score. This model is essentially an ordered partition model (Wilson, 1992) because two (or more) categories are assigned the same score (Wu et al., 2007). The "Score: OPM" and "Average Ability: OPM" columns show the score for each response category and average ability of students in each response category from the ordered partition model in the case of item A3_14ab. The result from the ordered partition model also suggests that the average ability of students in the partial-credit response category is very close to that in the full-credit category. Hence, the partial-credit and full-credit categories in item A3_14ab were collapsed for the subsequent analyses. The same logic and practice were applied to all items with partial-credit categories that do not represent a specific progression level. See Appendix D for information of average ability of students within each score category for all polytomous items used in this study. If the average ability of students in a partial-credit category was closer to that in the incorrect-response category, the partial-credit category was collapsed with the incorrect-response category.

Table 3.1 Coding of Response Categories: Item A3_14ab

Code	**# of Students**	**Score: PCM**	**Average Ability: PCM**	**Score: OPM**	**Average Ability: OPM**
0	121	0	-0.42	0	-0.72
1	96	1	0.19	1	0.40
2	130	2	0.23	1	0.42

19 out of the 20 argumentation items were polytomous in the original scoring scheme. 16 polytomous items have only one partial-credit category. Based on a comparison of the average IRT ability estimate of students across score categories for each polytomous item, for 15 items the partial-credit and full-credit response categories were combined and for one item the partial-credit and incorrect-response categories were

combined. In addition, three polytomous items have two partial-credit response categories. For two out of the three items, the partial-credit category adjacent to the full-credit category was combined with the full-credit category, and the partial-credit category adjacent to the incorrect category was combined with the incorrect category. For one item, item A2_5d, the partial-credit category adjacent to the full-credit category was combined with the incorrect-response category, and the partial-credit category adjacent to the incorrect category was combined with the full-credit category, See Chapter 2 for a discussion of item A2_5d.

After the aforementioned data preparation, in the subsequent analysis all items are dichotomous: incorrect responses are coded as 0 and responses indicative of the progression level an item is designed to measure are coded as 1.

3.3.2 Analysis

The goal was to (a) identify the number of latent classes, and (b) investigate the association between each latent class and hypothesized progression levels. A series of ordered latent class models with different numbers of classes were fitted. Three criteria were used in model selection: (a) model fit statistics, (b) classification error, and (c) interpretability of latent classes. First, a better fitting model is one with a lower value of the Bayesian Information Criterion (BIC). BIC has been reported to perform better than other information criteria on deciding the number of classes in mixture modeling (Nylund, Asparouhov, & Muthén, 2007), and is defined as below,

$$\text{BIC} = -2 \text{ log-likelihood} + (\log N) \text{ npar} \tag{3.4}$$

, where N denotes the sample size and npar denotes the number of parameters. If two or more models have similar values of BIC, the model with a smaller value of classification error and better class interpretability would be selected.

All latent class models were estimated with the software Latent GOLD (Vermunt & Magidson, 2005a). The estimation of parameters in Latent GOLD is based on the maximum likelihood method. In order to prevent boundary solutions, priors were used in the estimation procedure, which turns the estimation method to Posterior Mode (PM) estimation. In the context of the current study, the boundary problem is that multinomial probabilities may become zero. This problem is circumvented by using Dirichlet priors for the conditional response probabilities (Vermunt & Magidson, 2005b). The influence of the default priors on the parameter estimates is negligible with moderate sample sizes. Latent GOLD uses both the EM and the Newton-Raphson algorithm in estimation to find estimates of model parameters. A well-known problem in the latent class analysis is the occurrence of local maxima (Hagenaars & McCutcheon, 2002; McLachlan & Peel, 2000). To avoid this problem, 10 random sets of start values were used. Students were assigned to the latent class having the highest posterior membership probability.

Chapter 3: Latent Class Analysis

One of the fundamental assumptions in standard latent class models is local independence, that is, the latent variable of interest fully accounts for the relations between any two manifest variables (Hagenaars & McCutcheon, 2002). Violation of this assumption can cause lack of fit of a latent class model. In such cases, one way to proceed in the model-building procedure is to increase the number of classes until a model with an acceptable fit is obtained. The interest of the current study is in the substantive interpretability of each latent class. Increasing the number of latent classes due to the presence of local dependence might lead to a number of latent classes that were not meaningful substantively and was thus undesired.

In this study, an alternative model-fitting strategy was adopted, which has been expounded in the literature on latent class analysis to relax the local independence assumption (Hagenaars, 1988;Vermunt & Magidson, 2005b). Specifically, bivariate residuals, which provide direct checks of the local independence assumption, indicate how similar the estimated and observed bivariate associations are. These indices can be interpreted as lower bound estimates for the improvement in fit (-2 log-likelihood) if the corresponding local independence constraints were relaxed (Vermunt & Magidson, 2005b). In general, a bivariate residual larger than 3.84 indicates that the model has not adequately explain the correlation between the two indicators (Vermunt & Magidson, 2005a). In the case of a large bivariate residual, direct association between the two indicators was incorporated in the model. Suppose we would like to relax one local independence assumption and assume that y_{p1} and y_{p2} are directly related. Latent GOLD would model the response probabilities as below (Vermunt & Magidson, 2005b),

$$\Pr(y_{p1} = m_1, y_{p2} = m_2) = \sum_{c=1}^{C} \Pr(c) \Pr(y_{p1} = m_1, y_{p2} = m_2 \mid c) \tag{3.5}$$

, where $\Pr(y_{p1} = m_1, y_{p2} = m_2)$ denotes the probability of person p scoring m_1 on item 1 and scoring m_2 on item 2, and $\Pr(c)$ denotes the probability of being in latent class c. With direct association allowed for y_{p1} and y_{p2}, y_{p1} and y_{p2} now served as a joint dependent variable and $\Pr(y_{p1} = m_1, y_{p2} = m_2 \mid c)$ denotes the joint probability. When users include local dependencies using information on bivariate residuals, Latent GOLD automatically sets up the correct and most parsimonious probability structure for the situation concerned (Vermunt & Magidson, 2005b). The result of the latent class analysis was interpreted as usual.

Two sets of latent class analyses were performed. The first set of latent class analyses was performed with the full set of 20 argumentation items. As discussed in Chapter 2, five items were found to measure the latent argumentation construct poorly and did not function as desired. Also, another six[2] items were found to not conform to the

[2] Four items that did not conform to the ordinal structure of item thresholds across the progression levels were already flagged in Chapter 2. After data dichotomization as described in the section of data preparation, the data was re-analyzed with the Rasch

expected ordinal structure of the item thresholds across the progression levels. Considering that the presence of malfunctioning items might falsely influence the identification and classification of latent classes, a second set of latent class analyses was performed with the set of nine well-performing items.

As mentioned above, an examination of IRT ability estimates within each latent class and information of empirical performance levels set using Wright map might assist the interpretation of latent classes. For this purpose, results obtained from the Rasch model was investigated with respect to each latent class. Specifically, Expected A Posteriori (EAP) ability estimates were obtained for individual students. Also, the Wright map was used to set cut-off points between empirical performance levels.

3.4 Results

3.4.1 Number of Latent Classes

Table 3.2 summarizes model fit statistics of the ordered and unconstrained latent class models fitted to the data on the 20 argumentation items. According to the BIC, the ordered latent class analysis clearly points to the two-class solution, which is consistent with the result of the unconstrained latent class models. Figure 3.1 presents the pattern of the correct-response probabilities of the 20 argumentation items within each of the two latent classes. As revealed in Figure 3.1, with the incorporation of the items identified as malfunctioning in the Rasch and factor analyses, the latent class analysis of the 20 items failed to differentiate students at the intermediate level from students at high or low proficiency level. The identified two latent classes were also not readily interpretable with respect to learning progression levels.

Table 3.2. Model Fit Statistics of the Ordered and Unconstrained Latent Class Models (20 Items)

Ordered Latent Class Models				**Unconstrained Latent Class Models**			
Classes	**npar**	**Log likelihood**	**BIC**	**Classes**	**npar**	**Log likelihood**	**BIC**
1	35	-3492.00	7188.72	1	35	-3492.00	7188.72
2	56	-3330.53	6988.62	2	56	-3330.53	6988.62
3	72	-3297.78	7016.71	3	77	-3297.00	7044.40
4	85	-3286.99	7071.16	4	98	-3274.75	7122.74
5	91	-3275.59	7083.47	5	119	-3253.38	7202.84
6	93	-3274.15	7092.28	6	140	-3234.96	7288.82

model for this chapter, and the corresponding Wright map was examined. Two more items that did not follow the expected pattern were further identified.

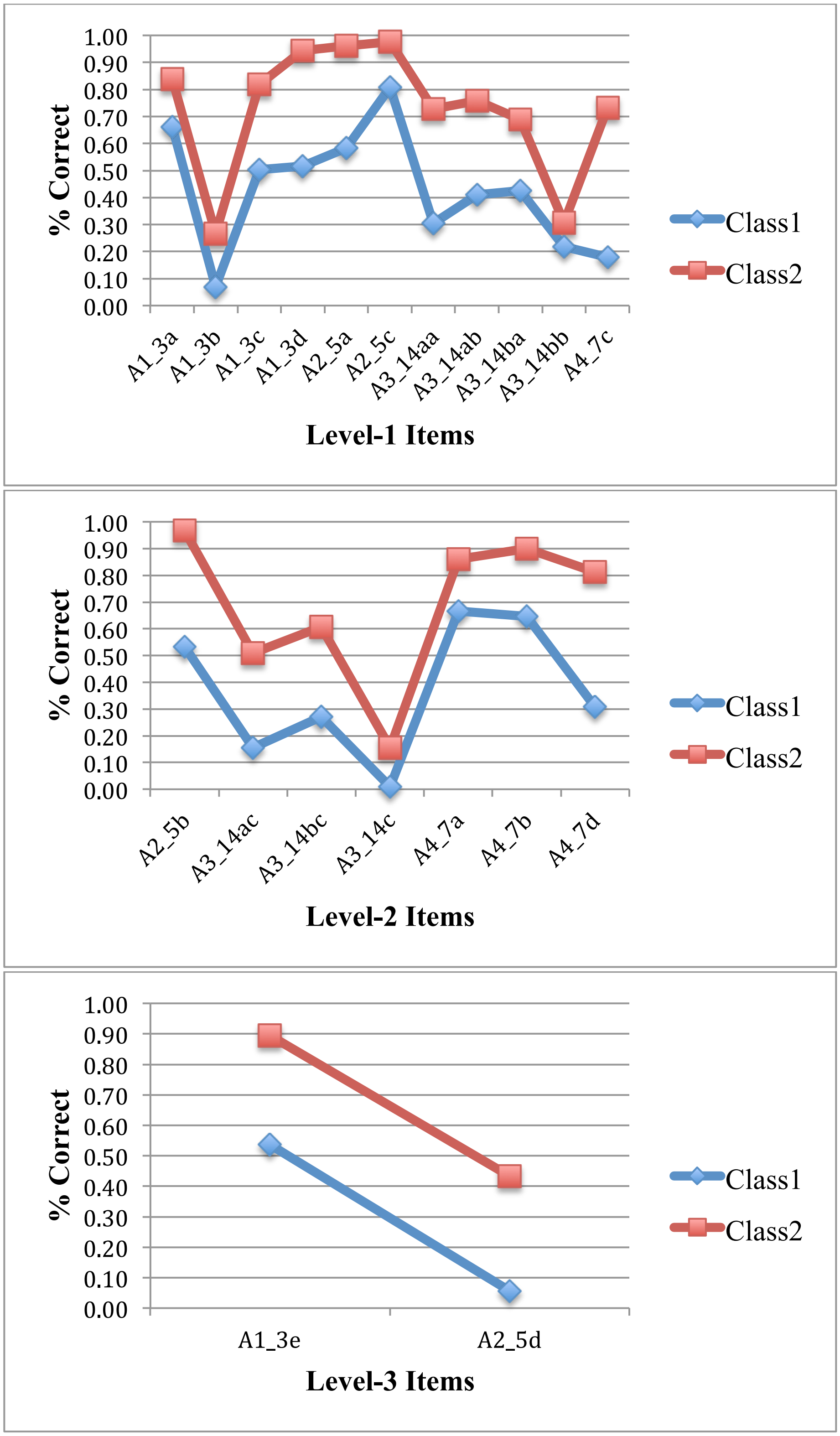


Figure 3.1. Correct-Response Probabilities across Latent Classes (20 Items)

Considering the influence of the malfunctioning items on class classification, a second set of the latent class analyses was performed with the nine argumentation items that were identified as well-behaving in the Rasch and factor analyses. This set of latent class analyses is the focus of the remaining part of the chapter. Table 3.3 presents the model fit statistics of the ordered and unconstrained latent class modeled fitted to the nine argumentation items. According to the BIC, the ordered latent class analysis points to the three-class solution, which is consistent with the result of the unconstrained latent class models.

Table 3.3. Model Fit Statistics of the Ordered and Unconstrained Latent Class Models (Nine Items)

Ordered Latent Class Models				**Unconstrained Latent Class Models**			
Classes	**npar**	**Log likelihood**	**BIC**	**Classes**	**npar**	**Log likelihood**	**BIC**
1	10	-1910.01	3878.51	1	10	-1910.01	3878.51
2	20	-1731.17	3579.33	2	20	-1731.17	3579.33
3	29	-1699.03	3567.68	3	30	-1699.03	3573.53
4	35	-1681.70	3568.12	4	40	-1673.28	3580.54
5	39	-1677.31	3582.73	5	50	-1660.65	3613.76
6	44	-1676.50	3610.37	6	60	-1652.46	3655.89

Although the three-class ordered latent class model has the smallest value of BIC, the BIC values of the three-class and four-class ordered latent class models are close. Thus, the values of classification error were further examined. The results support the selection of the three-class solution. Specifically, the classification error of the four-class ordered latent class model (0.19) is much higher than that of the three-class ordered latent class model (0.09). This is also the case for the unconstrained latent class model. The four-class unconstrained latent class model has a classification error (0.16) much higher than the three-class unconstrained latent class model (0.09) does. Figure 3.2 further presents the correct-response probabilities for the four-class ordered latent class model. Note that, with the large classification error (0.19), a student's class classification in the four-class model is doubtful, which in turn makes the interpretation of the four-class model less meaningful. The figure also reveals that, latent classes in the four-class ordered latent class model are not readily interpretable with respect to learning progression levels. These pieces of evidence all support the selection of the three-class ordered latent class model, which thereby would be the focus of the remaining sections.

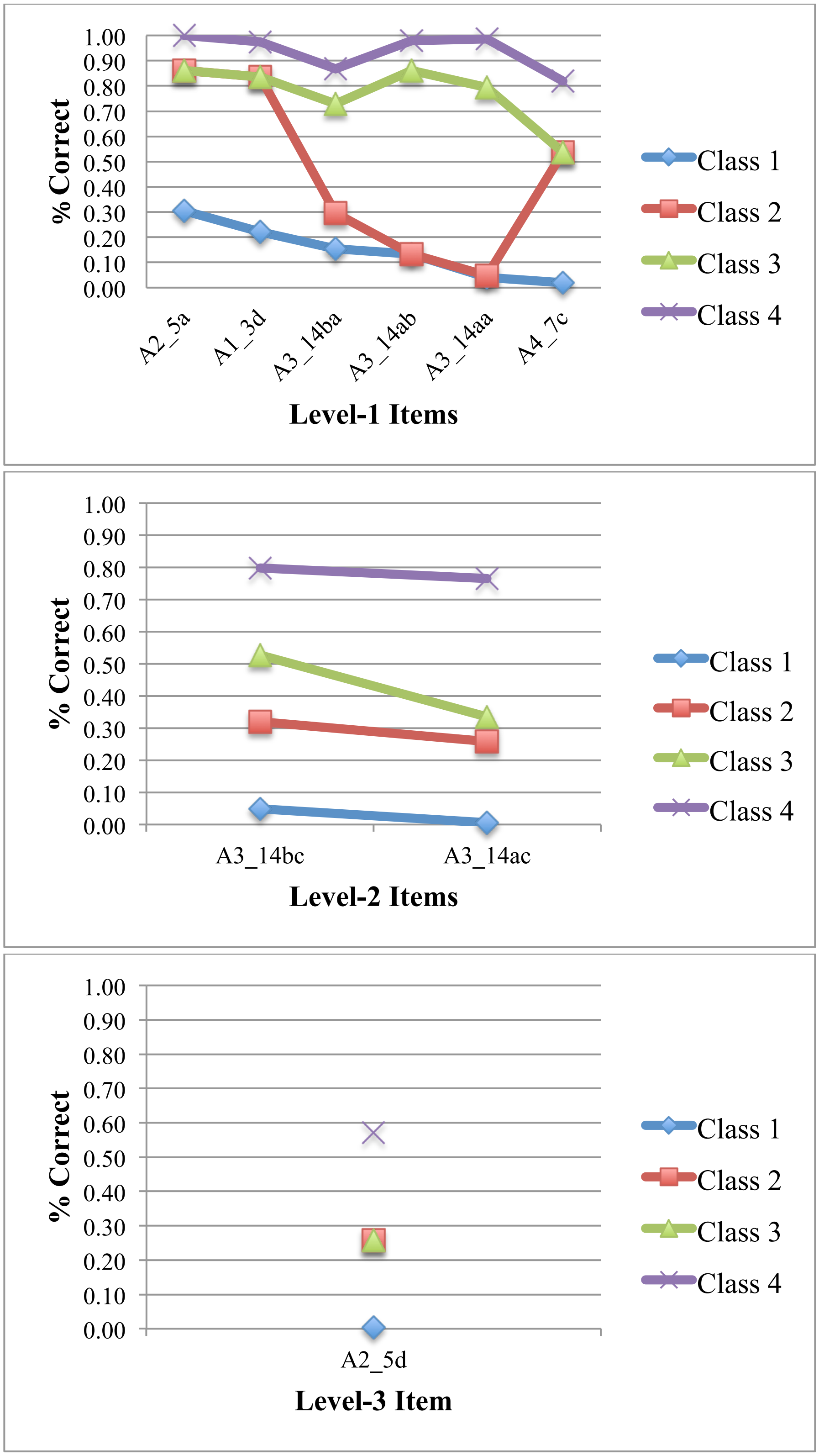


Figure 3.2. Correct-Response Probabilities (Nine Items, Four Classes)

3.4.2 Interpretation of the Latent Classes

As mentioned above, the following three methods were used to investigate characteristics of latent classes with respect to learning progression levels: (1) conditional probabilities, (2) Rasch ability distribution within each latent class, and (3) Wright map. Specifically, to examine the association between latent classes and the hypothesized progression levels, it is of interest to investigate whether patterns of student performance on items at different levels support the hypothesized ordinal structure of the three progression levels. In addition, how well students across latent classes were separated with respect to their abilities on a continuous scale calibrated from the Rasch model was examined, and also a comparison between empirical performance levels set using the Wright map and latent classes was made.

1. Conditional Probabilities

Figure 3.3 first presents correct-response probabilities of the six Level-1 items across the three latent classes. Overall, it is apparent that the three latent classes represent three groups of students differing in ability levels. Class 1 has low correct-response probabilities over all Level-1 items, and appears to be the low-performing group. In contrast, Class 3 has high correct-response probabilities on all Level-1 items. Among the classes, probabilities vary most across items in Class 2. A closer examination of probability patterns reveals that Class 2 performed much worse on the following three items: items A3_14aa, A3_14ab, and A3_14ba. Note that the rationale of dichotomization of these three items was the same as that of the majority of the items. Specifically, item A3_14ba is a dichotomous item in the original scoring scheme. Items A3_14aa and A3_14ab only have one partial-credit response category. In the process of dichotomization, the partial-credit and full-credit response categories were combined in the case of both items. This pattern is interesting, since these three items require a response format that is very different from the other items. While other items mostly ask students to either choose an option or compose a response, these three items ask students to “underline” sentences directly in a given text as their responses. The finding of the latent class analysis here provides evidence for the non-ignorable impact of response format on student performance.

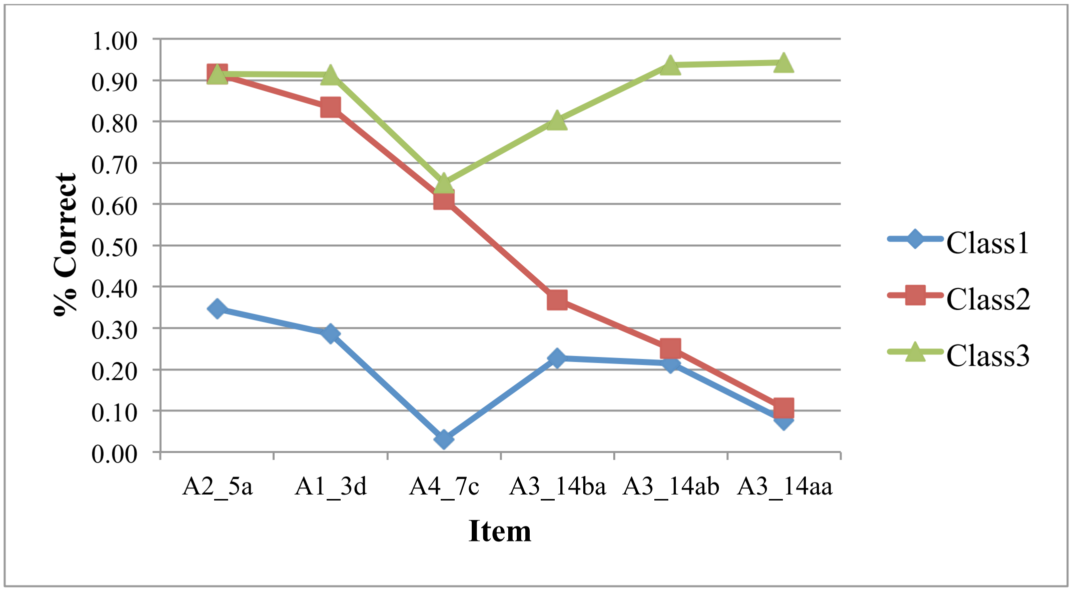


Figure 3.3. Correct-Response Probabilities of the Level-1 Items

Figure 3.4 presents the correct-response probabilities of the Level-2 items. Class 3 is the highest-performing group among the three classes, while Class 1 remains the lowest performing group. The performance of each class on Level 2 items is clearly worse than their performance on Level 1 items. This phenomenon is expected, since Level 2 items are meant to be more difficult than Level 1 items by design. Classes 2 and 3 appear to resemble students at intermediate levels. Note that only two Level-2 items were identified as well-behaving in the Rasch and factor analyses, and hence retained in this set of the latent class analysis. To obtain more information of student performance at or beyond the intermediate level, more Level-2 items should be developed in the future.

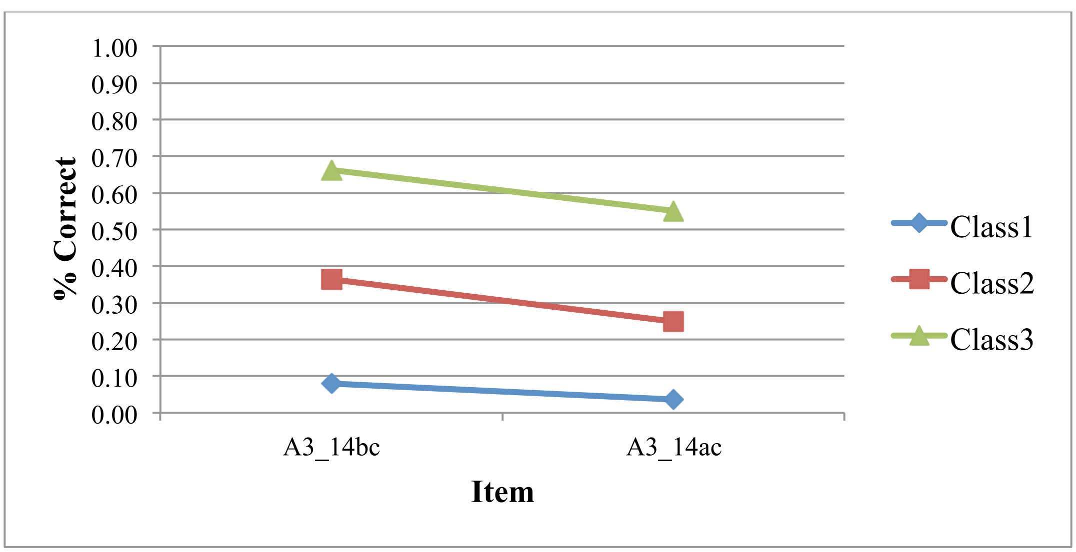


Figure 3.4. Correct-Response Probabilities of the Level-2 Items

There is only one Level-3 item in the current data. As expected, the Level-3 item is more difficult than the Level-1 or Level-2 items consistently across latent classes. The correct-response probability of this item is only 0.38 in the high-performing class, Class 3, 0.33 in Class 2, and 0.01 in Class 1. More Level-3 items will be needed to obtain information of student performance at the top level.

Table 3.4 shows the average correct-response probability of items within each progress level across the three latent classes. The ordered latent class models constrain persons in latent class c' to have a higher probability of correct response than persons in a higher latent class c, consistently across items. It is of the first interest to examine whether, for the successive latent classes, items targeting a higher progression level indeed appear to be more difficult than items targeting a lower progression level in average. As shown in Table 3.4, within Classes 1 and 3, the average correct-response probability is the largest for Level-1 items and the smallest for Level-3 items as desired. In Class 2, Level-1 items also have the largest average correct-response probability as expected, but Level-2 and Level-3 items have very similar average correct-response probabilities (actually, they are in the opposite order from what expected, but the actual probabilities are quite close, so I will not interpret that observation.). Considering the limited number of the Level-2 and Level-3 items in this set of analysis, the development of more items is necessary for a further examination of the response pattern across latent classes.

Table 3.4. Average Correct-Response Probability

	Learning Progression Levels Items Target		
Latent Class	**Level 1**	**Level 2**	**Level 3**
1	0.20	0.06	0.01
2	0.51	0.31	0.33
3	0.86	0.61	0.38

It is also of interest to align each latent class identified in empirical data with a hypothetical learning progression level. However, there is no established definition of "being at a progression level" with respect to the magnitude of or consistency in response probabilities in the literature of research on learning progressions. In a study that employed latent class analysis to evaluate whether empirical response patterns on diagnostic multiple-choice items afforded valid interpretations of hypothetical learning progression levels, Steedle and Shavelson (2009) adopted the rule that, students of a given progression level are expected to select responses indicative of that level with a probability of 0.70. Diakow, Irribarra, and Wilson (2011) did not use a cut-off value but only stated that, if a latent class represents a hypothesized progression level, the average expected score for that level or below should be near their possible maximum. Suppose all items are dichotomous and designed to measure a specific progression level; following what Diakow and colleagues suggested, if a latent class represents a hypothesized progression level, the average correct-response probability for items measuring that level or below should be reasonably close to 1.

Fluctuation in student performance on items measuring the same level is expected. Hence, lack of a well-accepted rule in the literature of learning progression research for determining how consistent student responses should be for a group of students to be considered as demonstrating performance at a specific progression level imposes challenges on the alignment of latent classes and progression levels. Note that research on assessments of mastery, or criterion-referenced assessments, is not of interest here. In the context of educational testing, the focus of mastery or criterion-referenced assessments is on setting standards and cut-scores, such that each examinee can be classified into a discrete achievement level, the simplest being pass or fail. The process of standard setting requires complex inputs, such as expert judgment, statistical expertise, and stakeholder interests (Berk, 1986; Macready & Dayton, 1977; Tiffin-Richards, Pant, & Köller, 2013), which is beyond the scope of the current study.

Given that there is no established criterion in the literature of learning progression or latent class analysis about how to assign a student into a progression level based on response probabilities, the following rules-of-thumb were proposed (M. Wilson, personal communication, June, 2013). Specifically, a latent class is regarded:

(a) to have mastered a learning progression level if the average correct-response probability of items targeting the level is larger than 0.75,
(b) to be actively learning knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is between 0.50 and 0.75,
(c) to be beginning learning knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is between 0.25 and 0.50, or
(d) to be struggling with knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is below 0.25[3].

Although one may argue that these rules are arbitrary, the set of rules proposed here do offer a start point to evaluate the performance of each latent class with respect to learning progression levels. Following these rules, the interpretation of the performance of each class, as shown in Table 3.4, would be as follows. Broadly speaking, Class 1 has the average correct-response probabilities of the Level-1, Level-2, and Level-3 items all below 0.25, and hence can be interpreted as struggling with *all* levels. Second, Class 2 has an average correct-response probability for the Level-1 items in the range of 0.50 and 0.75, and also has the average correct-response probabilities of the Level-2 and Level-3 items in the range between 0.25 and 0.50. Following the rule, Class 2 can be interpreted as actively learning Level 1 and beginning learning Level 2 and Level 3. Last, Class 3 has an average correct-response probability for the Level-1 items that is larger than 0.75, an

[3] Note that a coarser categorization, into three categories, would combine the second and third categories into one that might be labeled simply as "learning."

average correct-response probability for the Level-2 items falling in the range of 0.50 to 0.75, and an average correct-response probability of the Level-3 items falling in the range between 0.25 and 0.50, Thus, Class 3 can be interpreted as having mastered Level 1, be actively learning at Level 2, and beginning learning at Level 3.

2. IRT Ability Estimates within Each Latent Class

Figure 3.5 shows the distribution of the Rasch EAP ability estimates within each latent class, which were obtained from the Rasch model fitted to the nine items. As shown in the figure, the *average* ability increases from the lowest-performing class (Class 1) to the highest-performing class (Class 3) as desired: specifically, the mean is -2.09 logit for Class 1, -0.67 logit for Class 2, and 0.74 logit for Class 3. This is consistent with the results in the previous section. However, the ability distributions of the three classes are clearly overlapping on the Rasch scale. Recall that an important assumption of the latent class analysis is the homogeneity of students *within* each latent class. The ability distribution shown in Figure 3.5 calls for attention to the possible heterogeneity of the students within the classes: specifically, the standard deviation of the Rasch scale estimates for Class 1 is 0.57 logit, for Class 2 is 0.59 logit, and for Class 3 is 0.76 logit. Another way to gauge this range is to note that the mean of Class 2 occurs at approximately the 97th percentile for Class 1, and the mean of Class 3 occurs at approximately the 99th percentile for Class 2.

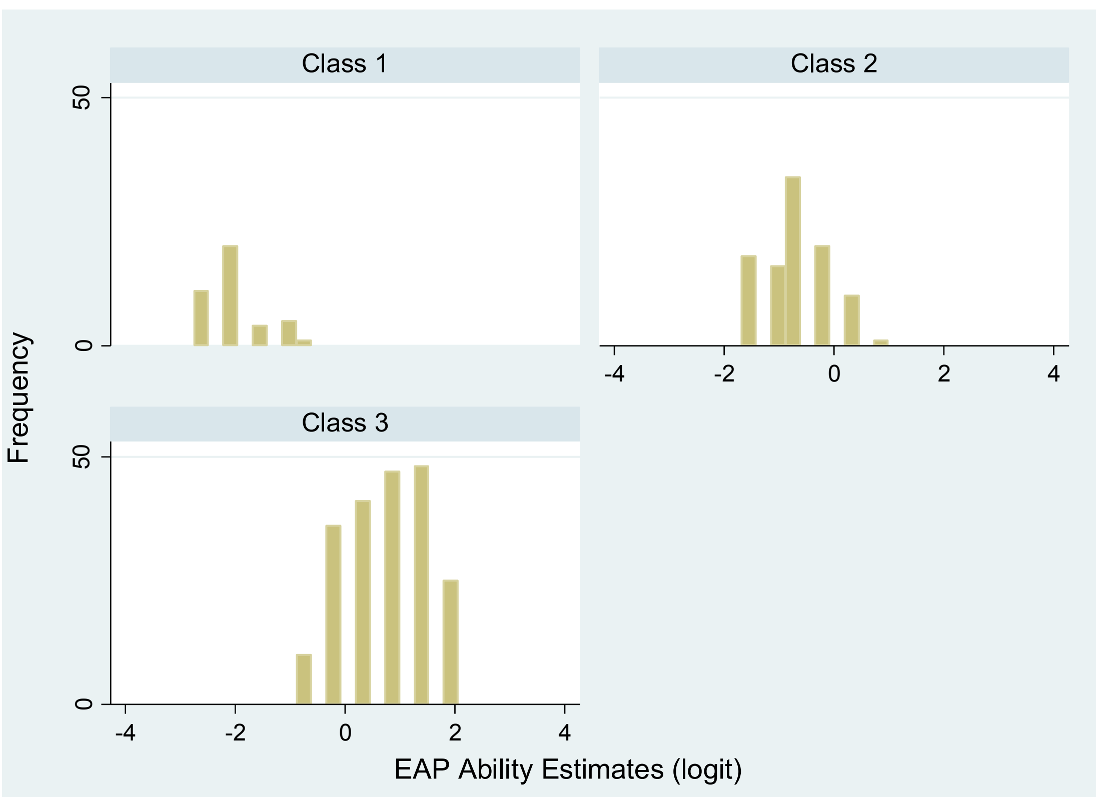


Figure 3.5. Expected A Posteriori Ability Estimates within Each Latent Class

3. Performance Levels Set Using Wright Map

Figure 3.6 shows the Wright map of the nine argumentation items obtained from the Rasch model. Please refer to Chapter 2 for a detailed account of the interpretation of the Wright map. The following cut-off points were used to set zones of learning progression levels on the Wright map. First, the cut-off point between Level-2 and Level-3 zones is the mid-point between item A2_5d and item A3_14ac. Second, the cut-off point between Level-1 and Level-2 zones is the mid-point between item A3_14bc and item A4_7c. Last, the cut-off point between Level-0 and Level-1 zones is item A2_5a. Specifically, students with EAP ability estimates equal to or larger than 0.81 logit were classified into the Level-3 zone. Students with EAP ability estimates ranging from -0.17 logit to 0.81 logit were classified into the Level-2 zone. Students with EAP ability estimates ranging from -2.25 logit to -0.17 logit were classified into the Level-1 zone. Last, students with EAP ability estimates lower than -2.25 logit were classified into the Level-0 zone. Note that students with EAP ability estimates falling in a particular zone were regarded as "learning" the corresponding learning progression level. For example, students classified into the Level-3 zone have ability estimates at the same level as the difficulty level of Level-3 items. Thus, the average probability of students at Level 3 getting a Level-3 item correct is 0.5. Following the rules described in the discussion of the average correct-response probabilities, students classified into the Level-3 zone on the Wright map are interpreted to be learning at Level 3.

Table 3.5 shows the pattern of performance levels set using Wright map within each of the three latent classes. First of all, 27% of the students in Class 1 were diagnosed as learning Level 0, while 73% of the students in Class 1 were classified as learning Level 1. Second, as observed in the pattern of average response probabilities, Class 2 was found to consist of students at the intermediate levels. Specifically, 89% of the students in Class 2 were diagnosed as learning Level 1, while 11% of the students in Class 2 were classified as learning Level 2. Third, Class 3 was the largest and the most heterogeneous class among the three latent classes. The diagnosis of the students in Class 3 ranged from learning Level 1 to learning Level 3.

Table 3.5. Pattern of Performance Levels Set Using Wright Map within Each Latent Class

	Performance Levels Set Using Wright Map				
Latent Class	**Level-0 Zone**	**Level-1 Zone**	**Level-2 Zone**	**Level-3 Zone**	**Total**
1	11 (27%)	30 (73%)	0 (0%)	0 (0%)	41
2	0 (0%)	88 (89%)	11 (11%)	0 (0%)	99
3	0 (0%)	46 (22%)	88 (43%)	73 (35%)	207
Total	11	164	99	73	347

Note. Percentage in the parentheses represents the percentage of students within each latent class that were classified into each zone.

It's important to recognize that fundamentally classifying students into a progression level using Wright map is different from the latent class analysis approach. Underlying the Wright map approach, the latent construct is assumed to be continuous and performance levels are set using corresponding item thresholds as cut-off points. Students classified into a progression level are assumed to have a probability of 0.5 of getting items at the same progression level right, and are thus regarded as "learning" that level. Underlying the latent class analysis approach, the latent construct is assumed to be discrete. The alignment of a latent class and a progression level is done in an initial exploratory manner with respect to the pattern of conditional probabilities of items targeting different progression levels. Recognizing the differences between the Wright map and latent class analysis approaches, the main purpose of examining students' performance levels set using Wright map within each latent class is to examine the variability in performance of students within a latent class with respect to a continuous proficiency scale. It is desired, under the latent class model approach, that, students within each latent class are homogeneous, and hence one would expect that the performance of students in each latent class would be well separated and distinguished with respect to a continuous latent ability scale. In an ideal scenario, the number of latent classes will be the same as the number of hypothetical progression levels, and most of (if not all) students would be in the diagonal cells in a cross-tabulation of latent classes and learning progression levels set with Wright map, as shown in Table 3.5.

Table 3.5 provides another piece of evidence that suggests the problem of heterogeneity of latent classes. Observing Table 3.5, we can see that students classified into Level-1 zone on the Wright map are presented in all three latent class classes, and students classified into Level-2 zone on the Wright map are presented in two out of the three latent classes. In contrast, students classified into the Level-0 zone are only presented in Class 1, and students classified into the Level-3 zone are only presented in Class 3. Note again that the number of the Level-2 and Level-3 items is very small in this set of data, which might lead to imprecision in the diagnosis of a student's learning progression level for those levels. Nevertheless, the findings reported here provide implications for research on learning progressions. Specifically, the results suggested the complexity of classifying students with middle levels of proficiency, which in turn led to difficulties of interpreting results of latent class analyses with respect to learning progression levels. This phenomenon has been reported in past research (e.g., West et al., 2012). Students' middle knowledge is referred to as the "*messy middle*" (Gotwals & Songer, 2010, p.277). Past researchers have pointed out that, it is most straightforward to define the endpoints of a learning progression. While the lowest level represents a novice state where students do not have any knowledge or skill of interest at all, the highest level refers to an expert state where students posses the knowledge and/or skill of interest. Defining the intermediate levels that students may experience to progress from the novice to the expert state is the most difficult, since numerous pathways to knowledge and/or skill acquisition are possible (Gotwals & Songer, 2010; West et al., 2012). In addition, since students with middle knowledge might not have all the pieces of knowledge and/or skills, they might be able to demonstrate a certain level of knowledge and/or skill on one task but not on another (e.g., Gotwals & Songer, 2010; Gotwals, Songer, & Bullard, 2012). In the context of latent class analysis, the inconsistent response behaviors students at intermediate levels might demonstrate would impose difficulties on classifying these intermediate-level students into homogenous latent classes. Without distinctive separation of students at different proficiency levels and identification of homogenous latent classes, interpreting latent classes with respect to learning progression levels would be hard to justify.

logit	Ability	Level-1 Items	Level-2 Items	Level-3 Item
	X			
3	XX			
	XX			
	X			
	XX			
	XX			
2	XXXXX			
	XXXX			
	XXXXX			
	XXXXXX			
	XXXXX			
1	XXXXX			A2_5d
	XXXXXX			
	XXXXXXX			
	XXXXXXXX		A3_14ac	
	XXXXXXXXXX			
	XXXXXXXXX			
0	XXXXXXXXX		A3_14bc	
	XXXXXXXXX	A4_7c		
	XXXXXXXX	A3_14aa		
	XXXXXXXX	A3_14ba		
	XXXXXXXX	A3_14ab		
-1	XXXXXXX			
	XXXXXX			
	XXX			
	XXX			
	XXXX			
-2	XXX	A1_3d		
	XX			
	XX	A2_5a		
	XX			
	X			
-3	X			
	X			
	X			
	X			

Figure 3.6. Wright Map (Nine Items)

3.5 Discussion

This chapter investigated the use of latent class analysis to obtain diagnostic information about students' performance level with respect to the hypothetical argumentation learning progression. Substantive implication for the future work on the studied learning progression and methodological lessons with respect to the latent class analysis of learning-progression-based assessments were discussed below.

3.5.1 Substantive Implication for the Future Development of the Learning Progression

The hypothetical learning progression consists of three performance levels. An examination of the latent class analysis result suggests that, none of the latent classes appears to fully attain the highest level of hypothesis. Among other reasons, lack of appropriate instruction may contribute to this outcome. Past research on learning progressions has clearly pointed out that, students' progression along the defined levels of a learning progression is crucially dependent on the instructional activities designed to further that progression (Corcoran, Mosher, & Rogat, 2009; Wilson, 2012). The studied argumentation learning progression, and the corresponding learning-progression-based items, were developed and piloted to obtain information of students' argumentation proficiency levels at the stage of middle school. It was hypothesized that students would learn about composing or critiquing an argument from other sources (J. Osborne, personal communication, January, 2012). Given that no instructional activities were provided, however, students were neither introduced to the critical elements of arguments nor educated about the level of sophistication expected in performance on an argumentation task systematically. This lack of instructional activities may also contribute to the inconsistency in performance across tasks. The instructional aspect of the learning progression (Wilson, 2012) should thus be addressed in the future work to guide students' progression.

3.5.2 Methodological Implications for the Analysis of Learning-Progression-Based Assessments

As West and colleagues pointed out (West et al., 2012, p.257), "*A central challenge in using learning progressions (LPs) in practice is modeling the relationships that link student performance on assessment tasks to students' levels on the LP*." Below is a discussion of the issues encountered in establishing the link between a latent class and a hypothetical progression level in this study, the implications of which may be generalized to other quantitative analyses of learning-progression-based assessments.

Chapter 3: Latent Class Analysis

Issue 1. How to handle responses that do not represent a specific progression level

In learning-progression-based assessments, deciding how to interpret and appropriately code students' partial understandings or skills is critical. One of the main purposes of basing assessment design in learning progressions is to understand where students stand in their progression to the expert state of knowledge or skills, instead of simply determining whether students succeed or fail on a task. For this purpose, in the past research on learning progression, coding schemes were developed based on the levels of the hypothetical learning progression, such that students' responses to or performance on a task could be directly mapped to progression levels (Alonzo & Steedle, 2009; Gotwals, Songer, & Bullard, 2012).

Nevertheless, mapping students' partial understandings to a particular progression level could be a challenge. In this study, student responses demonstrating partial understandings were found to indicate incomplete or imperfect performance at a hypothetical progression level, and could not be mapped to a specific progression level. This type of "*messy middle*" was reported in other research on learning progressions too (e.g., Gotwals & Songer, 2010; Gotwals, Songer, & Bullard, 2012). Students with middle knowledge might not have all the pieces of knowledge or skills to succeed on a task, and provide vague or discrepant responses that did not fit neatly into progression-level-based coding schemes.

In traditional educational testing, it is a common practice to give partial credits to student responses that are partially correct. The prerequisite of applying the aforementioned latent class models to validating a learning progression is that, each response category should correspond to a specific progression level. In the framework of latent class analysis, the interpretation of latent classes is based on the characteristics of their response probability patterns. Lack of a direct correspondence between a response category and a particular progression level jeopardizes the subsequent interpretation of latent classes with respect to learning progression levels (i.e., what progression level the performance of a particular latent class resembles). In this study, we evaluated the average ability estimates of students across response categories to determine whether a partial-credit response category should be combined with the full-credit or the incorrect-response category for each item. Nevertheless, the decision regarding how to deal with partial-credit response categories that do not represent performance at a progression level involves subjective judgment and requires an established criterion. Thus, the best strategy is to always design partial credit coding so that they correspond to a level of the learning progression. This may be realized by either (a) expanding the number of levels (and enhancing the criteria for) the levels and dimensions of the learning progression, or (b) changing the ways that questions are asked and responses coded, or (c) both.

Issue 2. Heterogeneity of students within each latent class

An important assumption of the latent class analysis is the homogeneity of students within each latent class. In the literature on the application of latent class analysis to investigate a hypothetical learning progression, the logic of mapping a latent class to a particular progression level is built upon an examination of the response probability patterns. If a latent class is not homogeneous, as assumed by the latent class approach, the response probabilities of a latent class are obtained from a group of students with a range of different proficiency levels. In such a case, using response probabilities to align the latent class with a specific progression level and classifying all of the students within the latent class into that progression level would be hard to justify.

In this study, Rasch analysis was used to examine how well latent classes separate persons with respect to a continuous proficiency scale. Examining the distribution of the ability estimates obtained from the Rasch model within each latent class, it was found that the variation in student abilities within each latent class was large. Also, the ability distributions of the students across the three latent classes were not well separated. The progression level each student demonstrated according to Wright map was further examined within each latent class, which also provided evidence of student heterogeneity within each latent class. These findings suggest the importance of checking the homogeneity assumption when using latent class analysis. In future research, models that allow heterogeneity within each class, such as Rasch mixture models, should be considered as well as traditional latent classes.

Issue 3. Alignment between a latent class and a progression level

Among the five components of learning progressions specified in the literature (Corcoran et al., 2009), one component is to specify operational definitions of what knowledge and/or skills would look like at each learning progression level. In the current learning progression research, an operational definition of a progression level was mainly provided as what knowledge and/or skills students were expected to demonstrate at that level. However, students' performance on items targeting the same level may vary. Past learning progression research (e.g., Gotwals & Songer, 2010; Gotwals, Songer, & Bullard, 2012) has reported that, students might be able to demonstrate a certain level of knowledge and/or skill on one task but not on another. The variation in student performance calls for reflection on what response behaviors students should demonstrate to be classified into a learning progression level.

In the framework of latent class analysis, if a latent class resembles a group of students at a particular hypothetical learning progression level, students in the latent class are expected to give responses indicative of that progression level with high probabilities consistently in the whole test. There are no established rules of thumb in the literature, however, with respect to what those conditional probabilities of giving responses representative of a progression level should be for a latent class to be regarded as resembling students at the particular progression level. In this study, I adopted a set of rules (M. Wilson, personal communication, June, 2013) to interpret the performance of a latent class with respect to learning progression levels using conditional response probabilities. Specifically, a latent class was regarded as (a) having mastered a learning progression level if the average correct-response probability of items targeting the level is larger than 0.75, (b) actively learning knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is between 0.50 and 0.75, (c) beginning learning knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is between 0.25 and 0.50, or (d) struggling with knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is below 0.25. While these rules may seem arbitrary, this set of rules can serve as a start point for future researchers to consider what "being at a learning progression level" means (e.g., does it mean that students have mastered the level?) and how to interpret the performance of a latent class with respect to a progression level.

Chapter 4: Comparing the Function of Content and Argumentation Items in a Science Test: A Multidimensional Item Response Theory Approach

4.1 Background

The new vision of science education calls for the development of science items that measure more than science content knowledge. The *Framework* states that science assessments should be capable of assessing a student's ability to integrate and apply knowledge to reason a scientific problem, as opposed to memorization of facts devoid of context (Gotwals, Songer, & Bullard, 2012). This new view provides a fairly multidimensional view of science assessments, and requires development of items that assess more than single content knowledge or process. For example, an item that asks students to reason about a problem in a given scientific context would require both content knowledge and reasoning skill. In response to this call, research on the design and performance of items that demand higher cognitive levels within the domain of science, as opposed to lower-level cognitive demands of recall and comprehension, is needed. Also, with inclusion of items that may assess more than single content knowledge or process ability, the use of multidimensional models to investigate the validity of construct(s) assessed in a science test is necessary.

The purpose of this research is to demonstrate the use of a multidimensional item response theory (IRT) framework, the multidimensional random coefficient multinomial logit (MRCML) framework, as a framework to investigate the construct validity of a science test. Specifically, the test consists of the following two types of items: (a) content items that measure a student's knowledge of a grade eight physics topic, Changes of State, and (b) argumentation items that are contextualized in the Changes of State topic and designed to measure a student's reasoning ability. With both content and argumentation items presented in the investigated science test, it is of interest to investigate the following two research questions:

- By design, the content and argumentation items are expected to measure two separate yet correlated constructs, i.e., the content knowledge and argumentation ability respectively. It is of interest to investigate if argumentation items do measure a construct distinguishable from the content construct as desired, and if so, what the relationship of the content and argumentation constructs is.

- Researchers in science education have suggested that literacy is the essential element of learning and doing science. Whilst the studied items are designed to assess a student's content and argumentation proficiency within the domain of science, succeeding on an item requires a certain level of English literacy. English literacy can be regarded as a *nuisance* dimension measured by these science items. It is expected that the demands of reading and writing skills may vary across items. It is of interest to investigate what content and argumentation items function differently from the other items and favor students with advanced English literacy level.

This study presents the use of two analyses in the MRCML framework to investigate each of the two research questions: (1) a dimensionality analysis, which examines the dimensional structure of the studied test, and (2) a differential item functioning (DIF) analysis, which investigates what content and argumentation items function differently from the other items and favor students with advanced English literacy level.

4.2 The Multidimensional Random Coefficient Multinomial Logit Framework

The formulation and potential usefulness of the multidimensional random coefficient multinomial logit (MRCML) framework have been previously introduced and discussed in the literature (Adams, Wilson, & Wang, 1997;Wang, Wilson, & Adams, 1997). Despite of these efforts, the MRCML framework has not been widely applied in practice. In the multidimensional framework, assume D latent traits underlie individuals' responses. These D latent traits define a D-dimensional latent space. The position of examinees in the D-dimensional latent space can then be represented with vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_D)'$. Suppose items are indexed as $i = 1,..., I$, and each item has $K_i + 1$ categories (0,…,K_i). The MRCML framework models the probability of a response in category k of item i as follows (Wang et al., 1997):

$$\Pr(X_{ik} = 1; \mathbf{A}, \mathbf{B}, \xi \mid \theta) = \frac{\exp(\mathbf{b}'_{\mathbf{ik}}\theta - \mathbf{a}'_{\mathbf{ik}}\xi)}{\sum_{u=1}^{K_i} \exp(\mathbf{b}'_{\mathbf{iu}}\theta - \mathbf{a}'_{\mathbf{iu}}\xi)} \tag{4.1}$$

where X_{ik} = 1 if response to item i is in category k, and X_{ik} =0 otherwise. Items are characterized by a vector $\boldsymbol{\xi} = (\xi_1, ..., \xi_p)'$ of p parameters. In a response probability model, linear combinations of these item parameters are used to describe the empirical characteristics of the response categories of each item. These linear combinations are defined by design vectors $\mathbf{a_{ik}}$ ($i = 1,..., I$, and $k = 1,..., K_i$), each of which is length p. Design vectors can be further collected to form a design matrix

$\mathbf{A} = (\mathbf{a}_{11},...,\mathbf{a}_{1K_1},\mathbf{a}_{21},...,\mathbf{a}_{2K_2},...,\mathbf{a}_{IK_I})'$. A response in category k in dimension d of item i is scored b_{ikd}. The scores can be collected across D dimensions into a column vector $\mathbf{b}_{\mathbf{ik}} = (b_{ik1},...,b_{ikD})'$, then be further collected into a scoring sub-matrix for item i as $\mathbf{B}_{\mathbf{i}} = (\mathbf{b}_{\mathbf{i1}},...,\mathbf{b}_{\mathbf{iD}})'$, and finally be collected into a scoring matrix for the whole test as $\mathbf{B} = (\mathbf{B}_1',...,\mathbf{B}_I')'$.

The MRCML framework is flexible and permits generalization to a variety of Rasch models (see Adams, Wilson, & Wang, 1997 for details). Through the manipulation of the scoring matrix, **B**, and the design matrix, **A**, users are allowed to form customized models in the MRCML framework. The item parameters and population mean(s) and variance(s) in the MRCML model are estimated with the marginal maximum likelihood technique. A detailed description of the parameter estimation in the MRCML framework can be found in Adams, Wilson and Wang (1997).

4.3 Dimensionality Analysis

The test of interest is intended to measure a grade eight student's proficiency on a physics topic, Changes of State. Whilst the content items are designed to measure a student's knowledge of the topic, the argumentation items are meant to assess a student's competency to construct and critique arguments in the selected content domain. In the literature, constructing and critiquing arguments is conceived as a competency that is dependent on both knowledge and a competency to reason. That is, reasoning is a component of the competency to argue (Walton, 1990;Osborne et al., 2012). Whilst the performance of argumentation inevitably depends on a person's knowledge of a topic, the argumentation items in the current test focuses on the reasoning component of the competency to argue. For this purpose, content knowledge deemed relevant was presented in the prompts of argumentation items, and a student's performance on argumentation items was evaluated solely with respect to his/her competency to construct or critique arguments.

Built upon the rationale of the development of content and argumentation items, it is of first interest to investigate (a) whether the argumentation items measure a separate argumentation construct, which is distinguishable from the Changes of State content construct, as intended in test design, and if so, (b) how the content and argumentation constructs are correlated. Von Aufschnaiter and colleagues (von Aufschnaiter, Erduran, Osborne, & Simon, 2008) have suggested that, "*students' ability to undertake high-level argumentation depends on whether or not the content of the argument relates to students' prior experiences*" (p.127). In classroom case studies, they observed that students did not make use of information in the argumentation tasks that is not at their level of prior knowledge, regardless of whether this knowledge related to a scientific concept or was based on their everyday experiences. Although the argumentation items are not intended to directly measure the knowledge of the Changes of State concept as the content items do, it is expected that performance on the argumentation items will be positively correlated with performance on the content items.

Figure 4.1 presents the graphical representations of three competing models to be tested: the unidimensional model, the multidimensional between-item model, and the multidimensional within-item model (Wang et al., 1997). Underlying the multidimensional between-item model, the content and argumentation items are assumed to measure content and argumentation constructs respectively. The two constructs are allowed to be correlated with each other in model formulation. Underlying the multidimensional within-item model, all content items are assumed to measure only the content construct and the argumentation items in the item bundle A1 are assumed to measure only the argumentation construct, while the other argumentation items in the item bundles A2 through A4 are assumed to measure both content and argumentation constructs. While the multidimensional between-item model was formulated according to the rationale of the item design (see Osborne et al., 2012), the alternative hypothesis of the multidimensional within-item model was proposed by a content specialist in a recent discussion (B. Henderson, personal communication, March, 2013). See Table 1 in Appendix E for details. Suppose the argumentation items fail to measure an additional higher-level reasoning ability and do not function differently from content items, then the test would be unidimensional with all items assessing a composite of content knowledge and argumentation proficiency.

Note that the models presented in Figure 4.1 assume conditional independence, as also known as local independence, among item responses. Underlying the assumption of local independence (Rosenbaum, 1984), the response pattern probability is essentially equal to the product of the individual item probabilities, controlling for the relevant person and item parameters. As described in Chapter 1, the investigated argumentation assessment is composed of four scenarios, each of which is followed with a set of items. It has been well documented in the literature of educational and psychological testing that items sharing a common stimulus might be locally dependent (e.g., Rosenbaum, 1988; Wilson & Adams, 1995; Embretson & Reise, 2000). If the assumption of local independence is violated, aspects of the model will be mis-estimated by the IRT models presented in Figure 4.1. In particular, the standard errors of item parameters tend to be under-estimated.

Rosenbaum (1988) introduced the term "*item bundle*" to denote a subset of items that share a common stimulus, and suggested that one should formulate the independence requirement between the item bundles instead of individual items. Following this concept, a variety of strategies to account for dependence between items in the same bundle have been proposed. The focus here is on two modeling strategies that can be formulated in the MRCML framework. The first strategy focuses on the total score on a bundle, i.e., the sum of item scores within a bundle. Using this bundle score strategy, each bundle is modeled as an ordered polytomous item. Wilson (1988) has applied this strategy using the partial credit model (Masters, 1982). As discussed by Wilson (1988), this strategy ignored response vectors in the same bundle and resulted in loss of information. The second strategy is to take individual items in a bundle into account and model response vectors explicitly. Wilson and Adams (1995) have adopted this strategy and proposed the item bundle model. A saturated bundle model includes a different parameter for each response vector within a bundle, except for the reference response vector. While the saturated bundle model has the advantage of utilizing information of response vectors within each bundle, the number of parameters increases quickly as the number of items within a bundle or the number of item response category increases. Wilson and Adams (1995) have also discussed strategies for designing models that combined the advantages of the partial credit model approach and the saturated bundle approach.

Chapter 4: Multidimensional Models

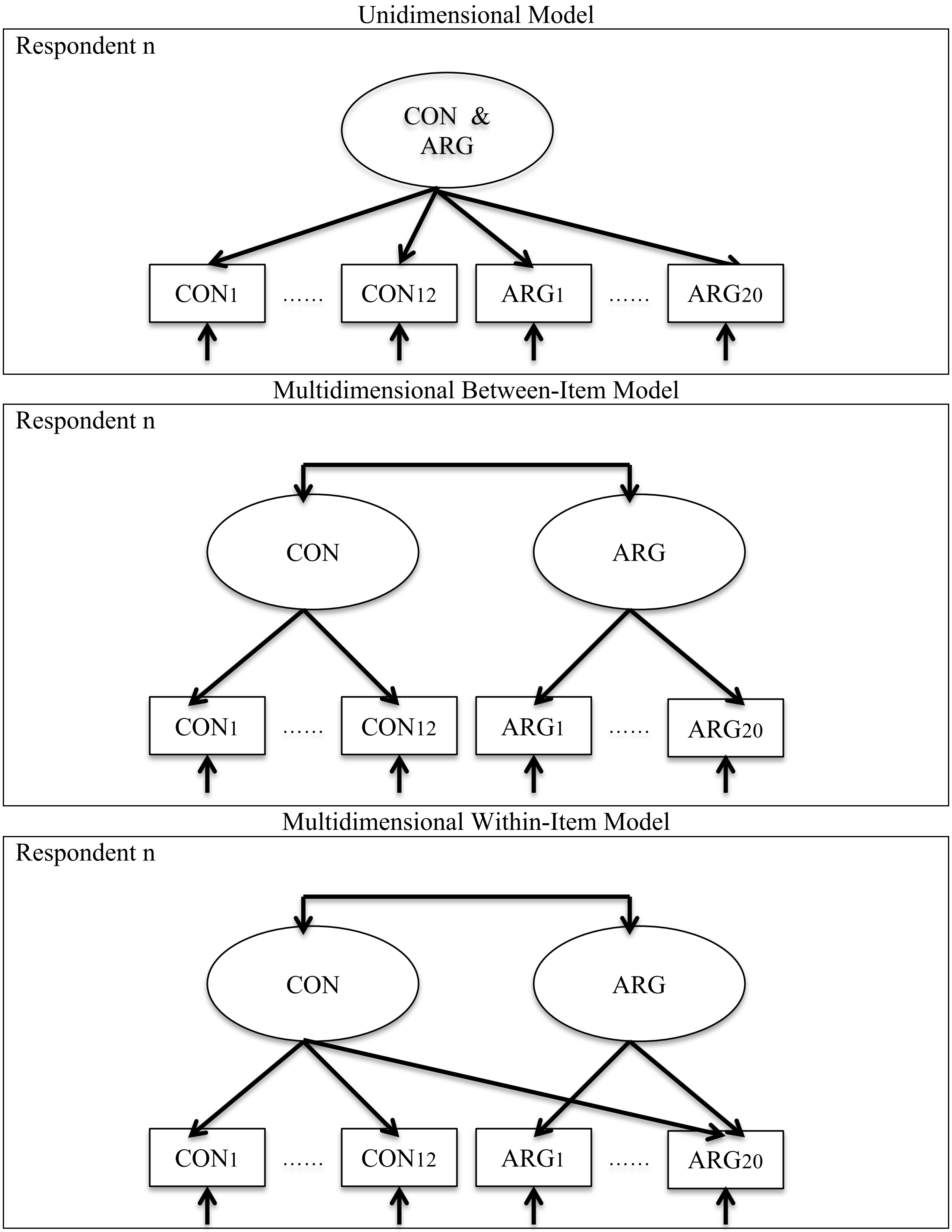


Figure 4.1. Graphical Representations of the Unidimensional and Multidimensional Models (CON=Content and ARG=Argumentation)

Chapter 4: Multidimensional Models

Considering that model misfit might result from the potential local dependence between the investigated argumentation items, argumentation items sharing a common stimulus were modeled as an item bundle in the dimensionality analysis. This leads to four item bundles, A1 to A4. Each argumentation item bundle contains four to seven items, and most of the items have three or larger response categories. The resulting number of valid response vectors for bundle A1 to A4 is thereby large. Specifically, bundle A1 has 101 valid response vectors, bundle A2 has 62 response vectors, bundle A3 has 166 response vectors, and bundle A4 has 66 response vectors.[4] Due to the large number of response vectors, the saturated item bundle model (Wilson & Adams, 1995) could not be estimated. Alternatively, each of the four argumentation item bundles could be readily modeled as an ordered polytomous item (Wilson, 1988) in the dimensionality analysis. Note that the focus here is to compare the fit of the aforementioned unidimensional and multidimensional models, and the estimation of response vector parameters is not of central interest. In spite of the loss of information of response vectors, modeling each bundle as an item accounts for potential local dependence between items in the same bundle and suffices the current interest.

For the purpose of model identification, either item difficulty or person ability has to be centered on zero for each dimension in the MRCML model. This constraint is necessary in other multidimensional modeling approaches. Estimates for different dimensions will need to be calibrated on a common scale before these estimates can be compared directly. In the current study, the Delta Dimensional Alignment technique (Schwartz & Ayers, 2011) was employed to align the content and argumentation dimensions when fitting a multidimensional model to data. Specifically, the Delta Dimensional Alignment technique consists of the following three steps. Suppose polytomous items are presented in a test. At the first step, all items are assumed to measure a single dimension and item difficulty estimates are obtained from a unidimensional analysis. The mean (μ_{uni}) and standard deviation (σ_{uni}) of item difficulty estimates are then computed for each subset of items by dimension. At the second stage, a multidimensional between-item model is fitted to obtain another set of item difficulty estimates. Note that for model identification the mean of item difficulties is constrained to be zero for each dimension. Thus only the standard deviation (σ_{multi}) of item difficulty estimates needs to be computed for each dimension. Using the estimates obtained from these first two steps, item difficulty and step estimates obtained from the multidimensional analysis are transformed as follows:

Item difficulty:

$$\delta_{id(transformed)} = \delta_{id(multi)} \left(\frac{\sigma_{d(uni)}}{\sigma_{d(multi)}} \right) + \mu_{d(uni)} \tag{4.2}$$

[4] The number of response vectors for each item bundle reported here was calculated using the R code provided by Michelle LaMar.

Step difficulty:

$$\tau_{ikd(transformed)} = \tau_{ikd(multi)} \left(\frac{\sigma_{d(uni)}}{\sigma_{d(multi)}} \right) \tag{4.3}$$

The last step is to run a multidimensional analysis with these transformed item difficulty and step estimates as anchored values, where estimates for different dimensions are now aligned and can be compared directly.

4.4 Differential Item Functioning Analysis

Researchers in science education have pointed out that literacy is the essential element of learning and doing science (e.g., Anderson, 1999;Norris & Phillips, 2003). As Norris and Phillips (2003) argued, for individuals to be able to comprehend, interpret, analyze, and evaluate any text in science requires scientific literacy in its fundamental sense (e.g., reading and writing skills). Without appropriate literacy skill, science content knowledge presented in a text will be conceived as isolated facts, or cannot even be comprehended appropriately. In the case of the current study, whilst the studied science items are designed to measure a student's proficiency in the Changes of State content and in the scientific argumentation, succeeding on these items inevitably requires a certain level of English literacy. Without appropriate English literacy proficiency, students won't be able to comprehend an item and express their thoughts in responses. It is of interest to investigate whether students with higher English proficiency have a higher probability of succeeding on a particular item than other students, after conditioning on students' latent ability of the construct(s) of interest. Note that the latent ability used as the ability conditioning variable could be different, depending on the model of choice. A more detailed account of this was presented later in the text. To identify content and argumentation items that favor students with advanced English literacy proficiency, differential item functioning (DIF) analysis was employed.

4.4.1 Background of DIF

DIF is believed to exist, if the probability of giving a particular response to an item is not simply a function of a student's ability but also depends on a student's group membership (Meredith & Millsap, 1992; Holland & Wainer, 1993). Since the 1960s, DIF analysis has been undertaken as a crucial step in test development to identify items that unfairly disadvantage a certain demographic group. In addition to detection of DIF items, in recent years there is an increasing emphasis on identifying substantive causes for the occurrence of DIF. It has been reported that the general cause of DIF is essentially the presence of the multidimensionality (Lord, 1980). That is, DIF would occur in items that measure at least one secondary construct, in addition to the primary construct a test is intended to measure. Built upon this supposition, Shealy and Stout (1993) proposed a multidimensionality-based DIF paradigm, which provides a mathematical definition of how multidimensionality could cause DIF. Specifically, the authors argued that occurrence of DIF was due to the presence of two factors: (a) the item in investigation is sensitive not only to the primary construct of interest, but also to a secondary construct, and (b) a difference exists between two groups of examinees of interest in their conditional distributions of the secondary construct given values of the primary construct. In the case of the current study, the primary construct of interest is a student's proficiency in Changes of State content and argumentation, and the secondary construct is a student's English literacy level. If the secondary construct that results in the occurrence of DIF is undesired and regarded as a threat to test validity, DIF is referred to as adverse DIF. However, if the secondary construct inducing DIF is regarded as an auxiliary construct, the DIF is viewed as benign DIF.

Traditionally, DIF detection procedures have focused on individual items. In recent years, DIF procedures have been extended to test simultaneously whether a cluster of items favor a particular group of examinees. Douglas and colleagues (Douglas, Roussos, & Stout, 1996) introduced the concept of differential bundle functioning, where groups of items were examined for DIF simultaneously. Note that in their definition bundles were clusters of items that were not necessarily adjacent items but chosen according to some organizational principle. Using a multidimensional-IRT-model-based approach, they have demonstrated how differential bundle functioning could be used to identify potential sources of DIF. Wainer and colleagues (Wainer, Sireci, & Thissen, 1991) have introduced another term, differential testlet functioning, for a multidimensional approach to detect DIF simultaneously for a group of items. A testlet is an interrelated group of items that are always presented as a single unit in a test (Wainer & Kiely, 1987). In their approach, a testlet was modeled as a polytomous item, and the likelihood ratio test was conducted for DIF analysis.

As mentioned above, the investigated argumentation assessment consists of four scenarios, each of which is followed with a group of items. For these argumentation items, the DIF detection procedure would be performed at the bundle level for the following reasons. First, items in the same bundle were administered as a unit in the test. Analyzing them simultaneously as a unit matches the model to the design of the test (Wainer et al., 1991). Second, it is possible that DIF is not detectable at the individual item level but manifests at the bundle level. That is, testing DIF simultaneously for items in the bundle provides increased statistical power of detecting DIF at the bundle level (Wainer, Sireci, & Thissen, 1991; Douglas, Roussos, & Stout, 1996). Thus, each argumentation bundle is to be scored as an ordered polytomous item in the DIF analysis, as in the case of dimensionality analysis.

The MRCML framework is flexible in formulating different kinds of DIF models. In this study, all of the DIF models are formulated in the MRCML framework and estimated with a specialized software ConQuest (Wu et al., 2007). Information from the English-language arts (ELA) test in the 2011 California Standards Test (CST) was used as an indicator of a student's reading and writing proficiency. Students classified into the advanced ELA proficiency level in 2011 CST were in the focal group of advanced English literacy proficiency, and all of the other students were in the reference group.

4.4.2 The Unidimensional DIF Model in the MRCML Framework

Suppose the content and argumentation items in the studied test essentially measure one common latent construct, which is a weighted composite of content knowledge and argumentation proficiency. The DIF model would be formulated as a unidimensional model. The random coefficient multinomial logit (RCML) DIF model, (adapted from the formulation used by Santelices & Wilson, 2012) could be employed to investigate DIF for a dichotomous item as follows:

$$\log\left(\frac{\Pr(X_{ig}=1)}{\Pr(X_{ig}=0)}\right)=\theta_{pg}-\delta_i+\Delta_g-\gamma_{ig} \tag{4.4}$$

, where X_{ig} represents the binary response of a person in group g to item i . θ_{pg} is the latent ability of person p in group g, which is assumed to follow a normal distribution. δ_i is the difficulty parameter of item i . Δ_g represents the mean ability difference between students in the reference and the focal group, where g denotes the group membership. γ_{ig} refers to the DIF parameter for item i , which is modeled as the interaction between a given item i and a student's group membership. For model identification, either the mean of the ability distribution, or the sum of item difficulty parameters, needs to be fixed as zero. Also, in ConQuest $\sum_g \Delta_g = 0$ and $\sum_g \gamma_{ig} = 0$. For more details, see Paek (2002).

The RCML DIF model can be extended to polytomous items as follows:

$$\log\left(\frac{\Pr(X_{ikg}=1)}{\Pr(X_{i(k-1)g}=1)}\right)=\theta_{pg}-\delta_i-\tau_{ik}+\Delta_g-\gamma_{ig} \tag{4.5}$$

, where $\Pr(X_{ikg})$ refers to the probability of the response of a person p in group g to item i is in category *k*, τ_{ik} is the additional step difficulty parameter associated with category *k* of item i . The interpretations of all the other parameters are identical to those in the case of dichotomous data. Note that in this model it is assumed that the DIF effect occurs at the item level and manifests as the difference in the overall item difficulty parameters between the reference and focal groups. That is, the group-by-item interaction term is introduced only for item difficulty parameter. It is possible to add step-specific interaction terms in the model. Yet, in the current sample, the number of students in a specific score category in a given group is small, and thus adding an interaction term at the step level would be inappropriate. Hence, only the group-by-item interaction effect is considered in this study. In addition to the aforementioned constraints for model identification, one more constraint, $\sum_k \tau_{ik} = 0$, is required.

Note that ConQuest consists of two parts in the modeling of item responses (Wu et al., 2007; Santelices & Wilson, 2012): (a) a latent regression model (also referred to as the *population model*) that accounts for the formulation and estimation of latent trait(s), and (b) an item response model that accounts for the formulation and estimation of the item parameters. The main effect of the group membership can be modeled either in the item response model or in a latent regression model in ConQuest. In Equations 4.4 and 4.5, the main group effect is modeled at the item side as Δ_g . If the main group effect is modeled in the latent regression model, the RCML DIF model for polytomous data would be formulated as follows (adapted from the formulation used by Paek, 2002 and Huang, 2010):

$$\log\left(\frac{\Pr(X_{ikg}=1)}{\Pr(X_{i(k-1)g}=1)}\right) = \theta_{pg} - \delta_i - \tau_{ik} - \gamma_{ig} \tag{4.6}$$

, where $\theta_{pg} = \alpha_0 + \alpha_1 d_{pg} + \varepsilon_p$. $d_{pg} = 1$ if person p is in the focal group and zero otherwise. α_0 is the estimated average latent ability of the reference group, α_1 represents the difference in the average latent ability between the reference group and the focal group, and ε_p is the deviation of a student's latent ability from the mean. ε_p is assumed to follow a normal distribution with mean equal to zero.

Figure 4.2 is the graphical representation of the unidimensional DIF model. In this unidimensional model, all of the content and argumentation items are assumed to measure a common construct, which is a composite of content and argumentation proficiency. Note that DIF detection was performed at the bundle level for argumentation items. Each argumentation bundle was modeled as an ordered polytomous item. Advanced English literacy is the grouping variable in this study, which indicates whether a student is classified in the advanced category in the 2011 CST: ELA test. The link between the latent construct (i.e., CON&ARG) and grouping variable (i.e., Advanced English Literacy) represents the difference in the mean of the latent ability between the two groups. The link between each item and the grouping variable represents the DIF effect, because it is the direct effect of group on an item after allowing for the indirect group effect via the latent variable. ConQuest tests the DIF effect for each item consecutively, assuming the rest of the items are free of DIF. For more details, see Wang (2004).

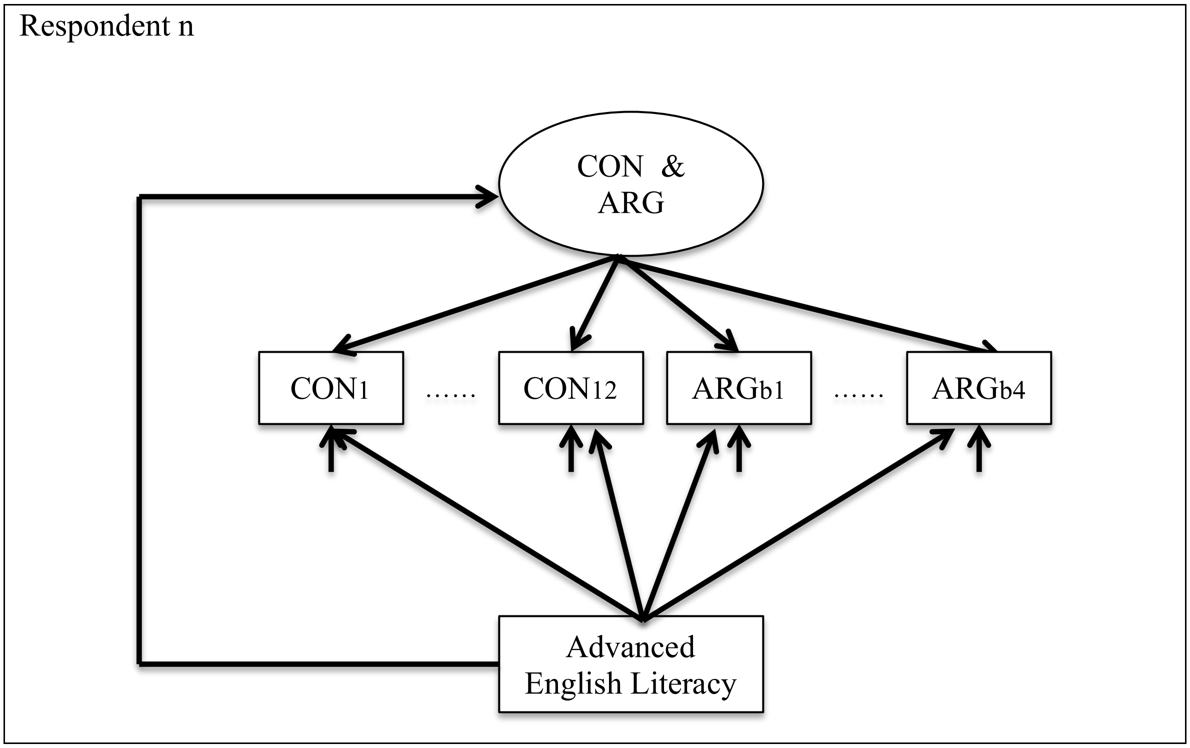


Figure 4.2. Unidimensional DIF Model in the MRCML Framework (CON=Content and ARG=Argumentation)

4.4.3 The Multidimensional DIF Model in the MRCML Framework

Suppose the content and argumentation items measure two separate and distinct constructs as designed, the use of a unidimensional model in DIF analysis would be inappropriate. Past research has reported that, if the multidimensionality of a test was ignored and a DIF analysis was performed under the assumption of unidimensionality, the result of DIF detection could be false (Ackerman, 1992; Huang, 2010). Nevertheless, the multidimensional DIF detection procedure has not been widely applied in practice.

The RCML DIF model can be easily adapted to a multidimensional DIF model in the MRCML framework, i.e., the MRCML DIF model. For polytomous items, the MRCML DIF model is formulated as follows (Paek, 2002; Huang, 2010):

$$\log\left(\frac{\Pr(X_{ikg}=1)}{\Pr(X_{i(k-1)g}=1)}\right)=\theta_{pd}-\delta_i-\tau_{ik}-\gamma_{ig} \tag{4.7}$$

, where X_{ikg} =1 means that the response of a person in group g to item i is in category k. θ_{pd} represents the latent ability of person p in dimension d. Within each dimension d, θ_{pd} can be formulated as $\theta_{pd} = \alpha_{0d} + \alpha_{1d} d_{pg} + \varepsilon_{pd}$, where α_{0d} represents the mean latent ability of the reference group in dimension d, α_{1d} represents the difference in the mean latent ability in dimension d between the reference and focal groups, and ε_{pd} is the deviance of a student's latent ability from the mean latent ability in dimension d. ε_{pd} is assumed to follow a normal distribution with mean equal to zero. The definitions of the other parameters in the model are the same as those in the RCML DIF model. For model identification, $\sum_i \delta_i = 0$, $\sum_k \tau_{ik} = 0$, and $\sum_g \gamma_{ig} = 0$.

Figure 4.3 is the graphical representation of the MRCML DIF model, where the content and argumentation items are assumed to measure two separate yet correlated constructs, i.e., content (CON) and argumentation (ARG) constructs. As in the case of unidimensional DIF analysis, advanced English literacy is the grouping variable here, which indicates whether a student is classified in the advanced category in the 2011 CST: ELA test. The link between each latent construct and grouping variable represents the difference in the average ability of each latent construct between the two groups. The link between each item and the grouping variable represents the DIF effect. ConQuest tests the DIF effect for each item consecutively, assuming the rest of the items are free of DIF.

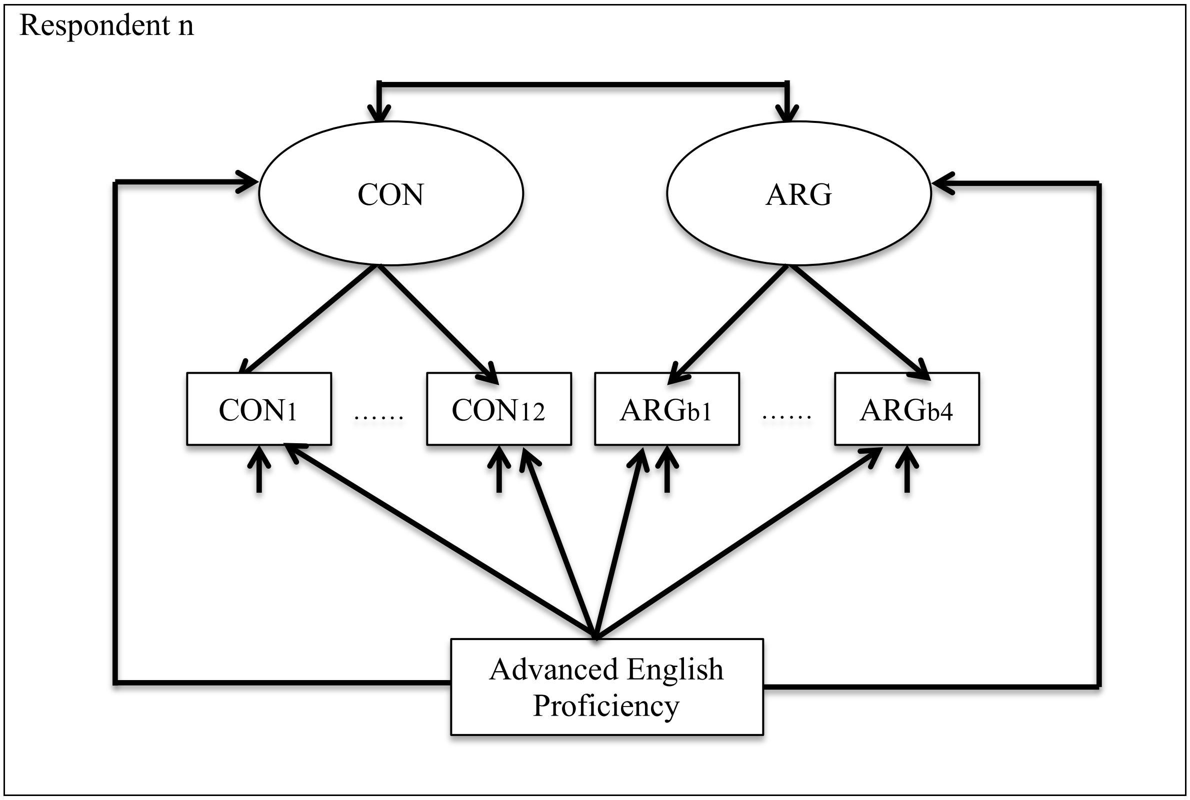


Figure 4.3. Graphical Representation of the Multidimensional DIF Model (CON=Content and ARG=Argumentation)

4.5 Analysis

4.5.1 Data

The sample of this study is identical to the sample described in Chapter 1. For DIF analyses, the grouping variable is based on the 2011 CST: ELA test. 342 out of the 347 grade eight students with information of 2011 CST: ELA test were used in the DIF analyses. 171 students classified in the advanced category in the 2011 CST: ELA test were grouped as the advanced English proficiency group, and modeled as focal group, while the other 171 students were in the reference group. Data from item responses on 20 argumentation items and 12 content items was analyzed in the dimensionality and DIF analyses. Among the 12 content items, eight items were in multiple-choice format and four items were in open-ended format.

4.5.2 Dimensionality Analysis

To assess the underlying dimensionality of the investigated science test, the software ConQuest (Wu et al., 2007) was used to fit the unidimensional model, the multidimensional between-item model, and the multidimensional within-item model. The unidimensional model was nested in each of the two multidimensional models. Thus, the relative fit of the unidimensional and each multidimensional model was commonly evaluated by comparing the difference in the deviance statistics between the two models with a chi-squared test where the degree of freedom is equal to the difference in the number of parameters estimated by each model. Note that the p value obtained from the chi-squared test is conservative (i.e., larger than the true p value), because the perfect correlation coefficient of one is on the border of parameter space. So if the p value is smaller than the level of significance (i.e., .05 in the current study), the result can still be interpreted as significant because the true p value will be even smaller. However, if the p value is larger than the level of significance (i.e., .05 in the current study), it'll be difficult to say whether the result is significant or non-significant (S. Rabe-Hesketh, personal communication, July, 2013). In addition, the Bayesian Information Criterion (BIC; see Equation 3.4), and the Akaike Information Criterion (AIC) were reported. The AIC is defined as below,

$$\text{AIC} = -2 \text{ log-likelihood} + 2 \text{ npar} \tag{4.8}$$

, where npar denotes the number of parameters. A better-fitting model is the one with a lower value of BIC or AIC. If a multidimensional model was selected, the Delta Dimensional Alignment technique would be applied to align the two dimensions, so that items have similar metrics (Schwartz & Ayers, 2011).

4.5.3 DIF Analysis in the MRCML Framework

DIF analyses in the two aforementioned formulations were performed with the software ConQuest (Wu et al., 2007) to evaluate what content items and argumentation bundles function differently from the other items/bundles and favor students with advanced English proficiency level as follows: (a) the RCML DIF model, and (b) the MRCML DIF model. In both the unidimensional and multidimensional models, the main group effect was modeled in the latent regression model of the latent ability as formulated in Equations 4.6 and 4.7.

The statistical significance of the DIF effect was examined using the Wald test. Under the Wald test, the maximum likelihood estimate of the parameter of interest is compared with the proposed value, and the difference between the two is assumed to follow a normal distribution asymptotically. Here the test statistic is the ratio of the estimated group-by-item interaction term (γ_{ig}) to its estimated standard error obtained from ConQuest for each item, which was compared to a standard normal distribution under the null hypothesis. .05 level of significance was used, that is, an item with the absolute value of this test statistic larger than 1.96 is regarded as displaying significant DIF.

For items or bundles displaying significant DIF, the effect size of the DIF effects was evaluated subsequently using the guidelines developed by Paek (2002). ETS has established classification rules of DIF effect size for the Mantel-Haenszel D-DIF statistic (Longford, Holland, & Thayer, 1993). Based on these ETS classification rules, Paek (Paek, 2002) has derived the following rules to classify the effect size of the DIF effects estimated in the MRCML framework with ConQuest: an item is said to display negligible DIF (Class A) if $2|\gamma_{ig}|<0.426$, intermediate DIF (Class B) if $0.426\leq 2|\gamma_{ig}|<0.638$, and large DIF (Class C) if $0.638\leq 2|\gamma_{ig}|$.

4.6 Results

4.6.1 Dimensionality Analysis

Table 4.1 below presents the fit statistics and reliability indices of the unidimensional and multidimensional models, based on the 12 content items and 20 argumentation items. Note that for this analysis the 20 argumentation items were assumed to be locally independent. As shown in the table, the multidimensional between-item model is the best-fitting model, which fits significantly better than the unidimensional model (chi-square statistic=55.29, df=2, p<.001). This provides evidence that the argumentation items collectively measure a latent construct distinguishable from the construct measured by the content items. In the multidimensional between-item model, the maximum likelihood estimate (MLE) person separation reliability of the content dimension is 0.71 and the MLE person separation reliability of the argumentation dimension is 0.81. The correlation between the content and argumentation dimensions is 0.76.

Table 4.1. Comparison of Unidimensional and Multidimensional Models (Assuming Local Independence; 12 Content and 20 Argumentation Items)

Model	**npar**	**Deviance**	**AIC**	**BIC**	**Reliability**	
					Content	**Argumentation**
Unidimensional	64	17,685.20	17,813.20	18,059.56	0.84	
Multidimensional Within-Item	66	17,755.56	17,887.56	18,141.62	0.62	0.54
Multidimensional Between-Item	66	17,629.91	17,761.91	18,015.97	0.71	0.81

As mentioned above, considering the potential threat of local dependence between argumentation items sharing a common stimulus, each argumentation item bundle was scored as an ordered polytomous item. Table 4.2 below presents the fit statistics and reliability indices of the unidimensional and multidimensional models, based on the 12 content items and four argumentation items. As shown in the table, the multidimensional between-item model remains the best-fitting model and fits significantly better than the unidimensional model (chi-square statistic=89.68, df=2, $p<.001$). Underlying the multidimensional between-item model, the correlation between the content and argumentation dimensions is 0.83. Comparing Table 4.1 and Table 4.2, we can see that, the same conclusion about model fit is revealed after potential local dependence among the argumentation items were accounted for by the bundling.

Table 4.2. Comparison of Unidimensional and Multidimensional Models: Each Argumentation Bundle Was Scored as One Single Partial Credit Item (12 Content Items and Four Argumentation Items)

Model	**Npar**	**Deviance**	**AIC**	**BIC**	**Reliability**	
					Content	**Argumentation**
Unidimensional	61	11,395.10	11,517.10	11,751.91	0.80	
Multidimensional Within-Item	63	11,475.07	11,601.07	11,843.58	0.60	0.24
Multidimensional Between-Item	63	11,305.42	11,431.42	11,673.93	0.71	0.67

Figure 4.4 presents the Wright map obtained from the multidimensional between-item model, where each argumentation bundle was scored as an ordered polytomous item. The Delta Dimensional Alignment technique was used to align the content and argumentation dimensions. The reliability indices of and correlation between the two dimensions after dimensional alignment are close to those before dimensional alignment. Specifically, the MLE person separation reliability is 0.70 for the content dimension and 0.69 for the argumentation dimension. The correlation between the two dimensions is 0.82. This outcome is expected since the Delta technique simply transforms the estimates of different dimensions onto a common scale. Now that the dimensions are aligned, the ability distributions and item difficulty estimates on the Wright map can be compared directly between the two dimensions. As to the ability distribution, the content dimension has a mean of -0.85 and variance of 0.63, while the argumentation dimension has a mean of 0.12 and variance of 0.16. As to the item thresholds, in general thresholds of argumentation bundles appear to locate higher than most of the content item thresholds on the Wright map. That is, it was more difficult to perform well on the argumentation bundles than on most of the content items. A further examination showed that, the content items with item thresholds that are located on the top of the Wright map, i.e., item 225b, item 219b and item 113, are all in open-ended format. As shown in the Wright map, achieving high scores on these open-ended content items appears to be as challenging as performing well on the argumentation bundles.

	Content		Argumentation	
logit	Ability	Item Thresholds	Ability	Item Thresholds
		225b.4		A2.9 A3.12
2				
				A1.9 A2.8
				A3.11
				A4.9
				A1.8
		219b.3		A3.10
1			X	A2.7
			X	
			XX	
	X		XXX	A1.7 A3.9 A4.8
	X		XXXXXX	
	X		XXXXXXXX	A2.6 A3.8
	XX	113.3	XXXXXX	A1.6 A3.7 A4.7
	XX		XXXXXXXX	A4.6
	XXX		XXXXXXXXX	A3.6
0	XX		XXXXXXXX	A2.5 A3.5
	XXXX		XXXXXXXXX	A3.4 A4.5
	XXX		XXXXXXX	A1.5
	XXX		XXXXXX	A2.4 A3.3 A4.4
	XXX		XXX	A1.4 A3.2
	XXXX	6	XX	A4.3
	XXXXX		X	A1.3 A3.1
	XXXX	225b.3	XX	A2.3
	XXX		X	A1.2 A4.2
	XXX			
-1	XXXXXX			A2.2
	XXXXX	222		A1.1 A4.1
	XXXXX	219b.2 225b.2 286_7.2		
	XXXX	117		
	XXXX			
	XXX	286_7.1		
	XXX			A2.1
	XX			
	XX	205		
-2	XX	225b.1		
	X			
	X			
	X	219b.1 5b		
	X			
		291		
		113.2		
-3		2a.2		
		2a.1		
		113.1		
-4		2c		

Figure 4.4. Wright Map of the Multidimensional Between-Item Model, with Dimensions Aligned with Delta Dimensional Alignment Technique (Argumentation Bundle Was Scored as One Single Partial Credit Item)

4.6.2 DIF Analysis

RCML DIF Model

As mentioned above, the unidimensional DIF model was fitted to the data, based on the assumption that the content and argumentation items measured a common latent construct. The main group effect was modeled in the latent regression model of the latent ability. The result from ConQuest shows that the mean latent ability of the reference group (i.e., the students in all but the advanced categories in the 2011 CST: ELA test) is 0.20 logit (standard error= 0.04), and the mean latent ability of the focal group (i.e., students classified in the advanced category in the 2011 CST: ELA test) is 1.06 logit (standard error=0.05) higher than that of the reference group. The variance of the latent dimension is 0.24.

Table 4.3 below presents information about the items identified as displaying non-ignorable DIF effects in the unidimensional DIF model. Please see Table 2 in Appendix E for a complete list of items and corresponding DIF results.

Table 4.3. Information of the DIF Items Identified in the Unidimensional DIF Model: Each Argumentation Bundle Was Scored as One Single Partial Credit Item (12 Content Items and Four Argumentation Items)

Item	**Type**	**Interaction (γ_{ig})***	**Standard Error**	**Effect Size**
291	Content	0.38	0.03	Large
2c	Content	0.56	0.03	Large
205	Content	0.30	0.03	Intermediate
225b	Content	-0.25	0.03	Intermediate
2a	Content	-0.30	0.03	Intermediate
A2	Argumentation	-0.25	0.03	Intermediate
A4	Argumentation	-0.29	0.12	Intermediate

* The group-by-item interaction term reported in this table is for the reference group.

MRCML DIF Model

The multidimensional DIF model was fitted subsequently to the data, where the content items were assumed to measure a content construct and argumentation items were assumed to measure a separate argumentation construct. The two constructs were allowed to be correlated in model estimation. Again, the main group effect was modeled in the latent regression model of the latent ability in each dimension. The focal group was found to have higher mean ability than the reference group in each dimension as expected. Specifically, the estimated mean ability of the reference group in the content dimension is 0.54 logit (standard error= 0.05), and the estimated mean ability of the focal group is 0.97 logit (standard error= 0.07) higher than that of the reference group. As to the argumentation dimension, the estimated mean ability of the reference group is -0.26 logit (standard error= 0.03) and the estimated mean ability of the focal group is 0.48 logit (standard error= 0.05) higher than that of the reference group. The variance of the content dimension is 0.45 and the variance of the argumentation dimension is 0.20. Now that the influence of English proficiency is modeled, the correlation between the two dimensions decreases, as expected, to 0.56. Also, the person separation reliability is 0.71 for the content dimension, and 0.59 for the argumentation dimension.

Table 4.4 presents the information of the items identified as displaying non-ignorable DIF effects in the multidimensional DIF model. Please see Table 3 in the Appendix E for a complete list of items and corresponding DIF results. Comparing Table 4.3 and Table 4.4, we can first observe that the number of items identified as displaying non-ignorable DIF in the multidimensional DIF model is much smaller than that in the unidimensional DIF model. The result of the previous section has shown that content and argumentation items assessed two distinct constructs, and provided support for the use of a multidimensional model. Past research has reported that, if the multidimensionality of a test was ignored and a DIF analysis was performed under the assumption of unidimensionality, items might be identified as displaying DIF falsely (Ackerman, 1992;Huang, 2010). The results reported here suggest the importance of accounting for multidimensionality, if applicable, in the DIF detection procedure.

Table 4.4. Information of the DIF Items Identified in the Multidimensional DIF Model: Each Argumentation Bundle Was Scored as One Single Partial Credit Item (12 Content Items and Four Argumentation Items)

Item	**Type**	**Interaction (γ_{ig})***	**Standard Error**	**Effect Size**
2a	Content	-0.40	0.11	Large
291	Content	0.27	0.10	Intermediate
205	Content	0.26	0.10	Intermediate

* The group-by-item interaction term reported in this table is for the reference group.

As shown in Table 4.4, only three content items were found to display non-ignorable DIF in the multidimensional DIF model. None of the four argumentation bundled items, each of which is a partial credit item generated from a bundle, were found to display non-ignorable DIF effects. An examination was conducted for each of the three DIF items. Please see Appendix E for these items.

First, item 2a has already been identified as a tricky item in the internal research meetings with content specialists. Specifically, whether a broken airplane could still be regarded as an “airplane” could be confusing to students. Second, item 291 was found to be an easy item to both reference and focal groups: 91.81% of the students in the focal group and 73.10% of the students in the reference group got this item correct. A further analysis showed that about 12% of the students in the reference group chose “decreases” and about 11% of the students in the reference group chose “increases.” The item does not say explicitly that the chocolate bar melts “in a closed system,” and that should be revised in a future pilot testing. Last, item 205 was found to be an easy item to the focal group (84.80% correct) and more challenging to the reference group (58.48% correct). A further analysis showed that the most popular distractor for students in the reference group is option B (20.47%): “Air that you are breathing out from your lungs.” Compared to the correct answer, option A (“Water vapor in the form of water droplets from your breath that has cooled.”), option B is more straightforward with respect to both the vocabulary and sentence structure, and may thereby be more appealing to the students in the reference group.

4.7 Summary

This chapter investigated the function of content and argumentation items in a science test with a multidimensional IRT modeling framework, the MRCML framework, with two types of analyses: (1) dimensionality analysis, which examined whether the argumentation items assessed a distinct and separate construct from the content items as desired or not, and (2) DIF analysis, which investigated what content and argumentation items function differently from the other items and favor students with advanced English literacy level.

As to the dimensionality analysis, the multidimensional between-item model was found to fit better than the unidimensional model and the multidimensional within-item model. This result is expected. As Osborne and colleagues (Osborne, Henderson, MacPherson, & Szu, 2012; p.5) described, they “*sought to emphasize students’ ability to engage in reasoning in a scientific context*” and “*seek to minimize the domain-specific knowledge by providing much of the knowledge either in the opening scenario or the stem of question*.” Also student responses to argumentation items were evaluated with respect to the hypothetical argumentation learning progression only. The item design and scoring scheme in this study were made to foster the emphasis of the current argumentation assessment on argumentation construct “mainly,” if not “only.” The finding that the multidimensional between-item model fits best supports the rationale of the item design that Osborne and his colleagues described.

Granted, students’ reasoning ability is dependent on and confounded by their content knowledge. This close relationship between content knowledge and argumentation proficiency is revealed through the high correlation coefficient in the current results. The limitation of the current design is that, if a student failed on an argumentation item, the current scoring scheme does not collect evidence on whether the failure is due to the student’s insufficient content knowledge or insufficient argumentation proficiency. To obtain this information, the development of blended assessments that explicitly evaluate student performance with respect to BOTH content and argumentation constructs are needed. An example of such efforts is found in research by Gotwals and colleagues (Gotwals et al., 2012), which was based in the *BioKIDS: Kids’ Inquiry of Diverse Species* project. The goal of the project is to develop a learning-progression-based curricular and assessment system to understand how students develop knowledge that fuses content and practices. Their learning progression includes a content progression and a practice progression. Items were designed to allow students to demonstrate their proficiency in both content and practice progressions, and student responses were coded for both levels of content knowledge and evidence-based explanations. For examples of these items, see Gotwals, Songer, & Bullard (2012).

As suggested in the DIF analysis, students identified as having advanced English proficiency in the CST were also found to have a higher proficiency level of the Changes of State content knowledge and argumentation ability than their counterparts. These results matched with expectation, since literacy has been regarded as an essential element of learning and doing science in the literature of science education (e.g., Anderson, 1999;Norris & Phillips, 2003). Controlling for the effects of literacy on the proficiency of content knowledge and argumentation ability, three content items were found to display large or intermediate DIF effects, while none of the four argumentation bundles were found to display non-ignorable DIF effects. Note that individual argumentation items might display non-ignorable DIF effects that canceled within the bundle. As mentioned above, the focus here is on the aggregate amount of DIF a bundle displays, since it is the construction unit of the argumentation assessment. Not finding DIF provides evidence that the investigated argumentation bundles did not unfairly favor students with advanced English proficiency level as identified by their CST performance.

Chapter 5: Conclusions and Discussions

With the increasing focus on incorporating argumentation in the teaching and learning of science in recent years, how to assess students' scientific argumentation ability becomes an emerging challenge. There exists very little work on development and validation of new kinds of assessments for scientific argumentation. In response to this challenge, a progression-based assessment was developed to assess scientific argumentation ability in middle school. This dissertation examined validity evidence for the scientific argumentation assessment with three separate yet related studies. The first study applied factor analysis and Rasch modeling to obtain evidence of internal structure validity for the investigated argumentation items. The second study investigated the use of latent class analysis to obtain diagnostic information about students' performance level with respect to the hypothetical argumentation learning progression. The third study applied a multidimensional item response modeling framework to investigate the differences, or similarities, in the function of the studied argumentation items and traditional content items. Collectively, these three studies contribute to the understanding of the performance of the assessment of interest, and inform future research on assessments of argumentation within the domain of science. The findings of each study and implications for future research were discussed as follows.

5.1 Results from Factor Analysis and Rasch Analysis

The exploratory factor analysis and Rasch analysis both provided evidence on the internal structure validity for the investigated argumentation assessment. First, the exploratory factor analysis suggested that, statistically there is no violation of unidimensionality and, thus, the investigated argumentation items measured one salient latent construct as intended. Second, the Wright map obtained from the Rasch partial credit model served as a graphical representation of the locations of item thresholds. Over half of the investigated argumentation items were found to have locations of thresholds that conform to the expected pattern: (a) item thresholds of a lower progression level have lower difficulty levels than item thresholds of a higher progression level, and (b) item thresholds of the same progression level are located at similar positions on the Wright map. The ordinal structure of these item thresholds shown on the Wright map provided support for the hypothetical structure of the three hypothesized learning progression levels.

In addition to collecting validity evidence on the internal structure, the validation process also provides several implications for assessment development. First, the cognitive demand of an item may depend on how the item is written. In the case of this study, it was found that the amount of information a student needs to process to solve an argumentation item influences the difficulty of the item. Between two items designed to measure the same learning progression level, the item that requires students to process more pieces of information may impose an increasing cognitive demand and thus appears to be more difficult than the other item. Second, ambiguity of the language can also contribute to an unwanted increase in item difficulty level. Third, in the current study student responses that are regarded as being partially correct are considered to demonstrate partial performance at the targeted progression level. The distinction between performance "at" and performance "partially at" a progression level is potentially vague, and calls for reflections on two questions: (a) what level of sophistication student responses should demonstrate to be considered as fully demonstrating performance at a particular learning progression level, and (b) how item design can be improved to prompt students to demonstrate their understanding and proficiency level to the fullest.

5.2 Results from Latent Class Analysis

Results from latent class analysis provided both substantive implication for the future work on the studied learning progression and raised methodological lessons with respect to the latent class analysis of learning-progression-based assessments.

5.2.1 Substantive Implication for the Future Development of the Learning Progression

The hypothetical learning progression consists of three performance levels. An examination of the latent class analysis results suggests that none of the latent classes appears to fully attain the highest level of the hypothesized learning progression. Among other reasons, lack of appropriate instruction may contribute to this outcome. Past research on learning progressions has clearly pointed out that, students' progression along the defined levels of a learning progression is crucially dependent on the instructional activities designed to further that progression (Corcoran, Mosher, & Rogat, 2009; Wilson, 2012). The studied argumentation learning progression, and the corresponding learning-progression-based items, were developed and piloted to obtain information of students' argumentation proficiency levels at the stage of middle school. It was hypothesized that students would learn about composing or critiquing an argument from other sources (J. Osborne, personal communication, January, 2012). Given that no instructional activities were provided, however, students were neither introduced to the critical elements of arguments nor educated about the level of sophistication expected in performance on an argumentation task systematically. This lack of instructional activities may also contribute to the inconsistency in performance across tasks. The instructional aspect of the learning progression (Wilson, 2012) should thus be addressed in future work to guide students' progression.

5.2.2 Methodological Implications for the Analysis of Learning-Progression-Based Assessments

As West and colleagues pointed out (West et al., 2012, p.257), "*A central challenge in using learning progressions (LPs) in practice is modeling the relationships that link student performance on assessment tasks to students' levels on the LP*." Below is a discussion of the issues encountered in establishing the link between a latent class and a hypothetical progression level in this study, the implications of which may be generalized to other quantitative analyses of learning-progression-based assessments.

Issue 1. How to handle responses that do not represent a specific progression level

In learning-progression-based assessments, deciding how to interpret and appropriately code students' partial understandings or skills is critical. One of the main purposes of basing assessment design in learning progressions is to understand where students stand in their progression to the expert state of knowledge or skills, instead of simply determining whether students succeed or fail on a task. For this purpose, in the past research on learning progression, coding schemes were developed based on the levels of the hypothetical learning progression, such that students' responses to or performance on a task could be directly mapped to progression levels (Alonzo & Steedle, 2009; Gotwals, Songer, & Bullard, 2012).

Nevertheless, mapping students' partial understandings to a particular progression level could be a challenge. In this study, student responses demonstrating partial understandings were found to indicate incomplete or imperfect performance at a hypothetical progression level; however, these could not be mapped to a specific progression level. This is related to the "*messy middle*," which has been reported in other research on learning progressions too (e.g., Gotwals & Songer, 2010; Gotwals, Songer, & Bullard, 2012). Students with a middling range of knowledge might not have all the pieces of knowledge or skills to succeed on a task, and provide vague or discrepant responses that did not fit neatly into progression-level-based coding schemes.

In traditional educational testing, it is a common practice to give partial credits to student responses that are partially correct, and this exactly what is intended to be the focus of "middle-range" levels. The prerequisite of applying the aforementioned latent class models to validating a learning progression is that, each response category should correspond to a specific progression level. In the framework of latent class analysis, the interpretation of latent classes is based on the characteristics of their response probability patterns. Lack of a direct correspondence between a response category and a particular progression level jeopardizes the subsequent interpretation of latent classes with respect to learning progression levels (i.e., what progression level the performance of a particular latent class resembles). In this study, we evaluated the average ability estimates of students across response categories to determine whether a partial-credit response category should be combined with the full-credit or the incorrect-response category for each item. Nevertheless, the decision regarding how to deal with partial-credit response categories that do not represent performance at a progression level involves subjective judgment and requires an established criterion. Thus, the best strategy is to always design partial credit coding so that they correspond to a level of the learning progression. This may be realized by either (a) expanding the number of levels (and enhancing the criteria for) the levels and dimensions of the learning progression, or (b) changing the ways that questions are asked and responses coded, or (c) both.

Issue 2. Heterogeneity of students within each latent class

An important assumption of latent class analysis is the homogeneity of students within each latent class. In the literature on the application of latent class analysis to investigate a hypothetical learning progression, the logic of mapping a latent class to a particular progression level is built upon an examination of the response probability patterns. If a latent class is not homogeneous, as assumed by the latent class approach, the response probabilities of a latent class are obtained from a group of students with a range of different proficiency levels. In such a case, the basic assumptions of the latent class approach are not justified. In particular, using response probabilities to align the latent class with a specific progression level and classifying all of the students within the latent class into that progression level would be hard to justify.

In this study, Rasch analysis was used to examine how well latent classes separate persons with respect to a continuous proficiency scale. Examining the distribution of the ability estimates obtained from the Rasch model within each latent class, it was found that the variation in student abilities within each of the obtained latent classes was large. Also, the ability distributions of the students across the three latent classes were not well separated. The progression level each student demonstrated according to the Wright map was further examined within each latent class, which also provided graphical evidence of student heterogeneity within each latent class. These findings suggest the importance of checking the homogeneity assumption when using latent class analysis. In future research, models that allow heterogeneity within each class, such as Rasch mixture models, should be considered as well as traditional latent classes.

Issue 3. Alignment between a latent class and a progression level

Among the five components of learning progressions specified in the literature (Corcoran et al., 2009), one component is to specify operational definitions of what knowledge or skills would look like at each learning progression level. In the current learning progression research, an operational definition of a progression level was mainly provided as what knowledge or skills students were expected to demonstrate at that level. However, students' performance on items targeting the same level may vary. Past learning progression research (e.g., Gotwals & Songer, 2010; Gotwals, Songer, & Bullard, 2012) has reported that, students might be able to demonstrate a certain level of knowledge or skill on one task but not on another. The variation in student performance calls for reflection on what response behaviors students should demonstrate to be classified into a learning progression level.

In the framework of latent class analysis, if a latent class resembles a group of students at a particular hypothetical learning progression level, students in the latent class are expected to give responses indicative of that progression level with high probabilities consistently in the whole test. There are no established rules of thumb in the literature, however, with respect to what those conditional probabilities of giving responses representative of a progression level should be for a latent class to be regarded as resembling students at the particular progression level. In this study, I adopted a set of rules (M. Wilson, personal communication, June, 2013) to interpret the performance of a latent class with respect to learning progression levels using conditional response probabilities. Specifically, a latent class was regarded as (a) having mastered a learning progression level if the average correct-response probability of items targeting the level is larger than 0.75, (b) actively learning knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is between 0.50 and 0.75, (c) beginning learning knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is between 0.25 and 0.50, or (d) struggling with knowledge or skills defined at a progression level if the average correct-response probability of items targeting that level is below 0.25. While these rules may seem arbitrary, this set of rules can serve as a start point for future researchers to consider what "being at a learning progression level" means (e.g., does it mean that students have mastered the level?) and how to interpret the performance of a latent class with respect to a progression level.

5.3 Results from Dimensionality Analysis and DIF Analysis

The third study investigated the function of content and argumentation items in a science test with a multidimensional IRT modeling framework, the MRCML framework, with two types of analyses: (1) dimensionality analysis, which examined whether the argumentation items assessed a distinct and separate construct from the content items as desired or not, and (2) DIF analysis, which investigated what content and argumentation items function differently from the other items and favor students with advanced English literacy level.

As to the dimensionality analysis, the multidimensional between-item model was found to fit better than the unidimensional model and the multidimensional within-item model. This result is expected. As Osborne and colleagues (Osborne, Henderson, MacPherson, & Szu, 2012; p.5) described, they "*sought to emphasize students' ability to engage in reasoning in a scientific context*" and "*seek to minimize the domain-specific knowledge by providing much of the knowledge either in the opening scenario or the stem of question*." Also student responses to argumentation items were evaluated with respect to the hypothetical argumentation learning progression only. The item design and scoring scheme in this study were made to foster the emphasis of the current argumentation assessment on the argumentation construct "mainly," if not "only." The finding that the multidimensional between-item model fits best supports the rationale of the item design that Osborne and his colleagues described and provides evidence that the studied argumentation items assessed a distinct and separate construct from the content items as desired.

Granted, students' reasoning ability is dependent on and confounded by their content knowledge. This close relationship between content knowledge and argumentation proficiency is revealed through the high correlation coefficient in the current results. The limitation of the current design is that, if a student failed on an argumentation item, the current scoring scheme does not collect evidence on whether the failure is due to the student's insufficient content knowledge or insufficient argumentation proficiency. To obtain this information, the development of blended assessments that explicitly evaluate student performance with respect to BOTH content and argumentation constructs are needed. An example of such efforts is found in research by Gotwals and colleagues (Gotwals et al., 2012), which was based in the *BioKIDS: Kids' Inquiry of Diverse Species* project. The goal of the project is to develop a learning-progression-based curricular and assessment system to understand how students develop knowledge that fuses content and practices. Their learning progression includes a content progression and a practice progression. Items were designed to allow students to demonstrate their proficiency in both content and practice progressions, and student responses were coded for both levels of content knowledge and evidence-based explanations. For examples of these items, see Gotwals, Songer, & Bullard (2012).

As suggested in the DIF analysis, students identified as having advanced English proficiency in the CST were also found to have a higher proficiency level of the Changes of State content knowledge and argumentation ability than their counterparts. These results matched with expectation, since literacy has been regarded as an essential element of learning and doing science in the literature of science education (e.g., Anderson, 1999; Norris & Phillips, 2003). Controlling for the effects of literacy on the proficiency of content knowledge and argumentation ability, three content items were found to display large or intermediate DIF effects, while none of the four argumentation bundles were found to display non-ignorable DIF effects. Note that individual argumentation items might display non-ignorable DIF effects that canceled within the bundle. Nevertheless, the focus here is on the aggregate amount of DIF a bundle displays, since it is the construction unit of the argumentation assessment. Not finding DIF provides evidence that the investigated argumentation bundles did not unfairly favor students with advanced English proficiency level as identified by their CST performance.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91.

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

Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000 technical report*. Paris: OECD Publishing.

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions*. Netherlands: Sense Publishers.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389–421.

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

Anderson, C. W. (1999). Inscriptions and science learning. *Journal of Research in Science Teaching*, *36*(9), 973–974.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*(1), 137–172.

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review*, *23*(2), 221–234.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, *27*(4), 509–540.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence based approach to reform.* (CPRE Research Report No. #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology*, *2*(43), 171–192.

Diakow, R., Irribarra, D. T., & Wilson, M. (2011). Analyzing the complex structure of a learning progression: Structured construct models. Presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*(4), 465–484.

References

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, N. J.: Prentice Hall.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education*, *92*(3), 404–423.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*, *94*(2), 259–281.

Gotwals, A. W., Songer, N. B., & Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 183–210). Netherlands: Sense Publishers.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods & Research*, *16*(3), 379–405.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Huang, X. (2010). *Differential item functioning: The consequence of language, curriculum, or culture?* (Unpublished doctoral dissertation). University of California, Berkeley, CA.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10*(2), 193–205.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Kennedy, C. A., & Wilson, M. (2007). *Using progress variables to interpret student achievement and progress* (BEAR Report Series No. 2006-12-01). University of California, Berkeley. Retrieved from http://bearcenter.berkeley.edu/publications/UsingProgVars_Kennedy_Wilson.pdf

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

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88–91.

References

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics*, *2*(2), 99–120.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McLachlan, G., & Peel, D. A. (2000). *Finite mixture models*. New York, NY: Wiley.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, *22*(3), 267–305.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.

Osborne, J. F., Henderson, B., MacPherson, A., & Szu, E. (2012). Assessing scientific argumentation by middle school pupils and testing a learning progression for argumentation. Presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Paek, I. (2002). *Investigations of differential item functioning: Comparisons among approaches, and extension to a multidimensional context* (Unpublished doctoral dissertation). University of California, Berkeley, CA.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher*, *18*(1), 16–25.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in Earth Science. *Journal of Research in Science Teaching*, *49*(6), 713–743.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*(3), 425–435.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*(3), 349–359.

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement*, *72*(1), 5–36.

Schwartz, R., & Ayers, E. (2011). *Delta dimensional alignment: Comparing performances across dimensions of the learning progression for assessing data modeling and statistical reasoning* (Unpublished manuscript). University of California, Berkeley, CA.

Shapiro, A. (1982). Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, *47*(2), 187–199.

References

Shapiro, A., & Ten Berge, J. M. F. (2002). Statistical inference of minimum rank factor analysis. *Psychometrika*, *67*(1), 79–94.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, Florida: CRC Press.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, *46*(6), 699–715.

Ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika*, *56*(2), 309–315.

Tiffin-Richards, S. P., Pant, H. A., & Köller, O. (2013). Setting standards for English foreign language assessment: Methodology, validation, and a degree of arbitrariness. *Educational Measurement: Issues and Practice*, *32*(2), 15–25.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2005a). *Latent Gold 4.0 user's guide*. Belmont, Massachusetts: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2005b). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, Massachusetts: Statistical Innovations Inc.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement*, *24*(3), 185–201.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*(3), 197–219.

Walton, D. N. (1990). What is reasoning? What is an argument? *The Journal of Philosophy*, *87*(8), 399–419.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*, 221–261.

Wang, W.-C., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. *Objective Measurement: Theory into Practice*, *4*, 139–155.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., DiCerbo, K. E., … Behrens, J. T. (2012). A Bayesian network approach to modeling learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 257–292). Netherlands: Sense Publishers.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, *12*(4), 353–364.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, *16*(4), 309–325.

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

References

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730.

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice: Hypothesized links between dimensions of the outcome progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions.* (pp. 317–343). Netherlands: Sense Publishers.

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

Wright, B. D., & Masters, G. N. (1981). *Rating scale analysis*. Chicago: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Camberwell, Victoria: Australia Council for Educational Research Press.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442.

Appendix A: The Hypothetical Scientific Argumentation Learning Progression

Level	Constructing Arguments	Critiquing Arguments	Description
1a	Making a claim		This is the essential sine qua non of any reasoning process that must be advancing an argument for something. As such it cannot really be tested. [Cognitive Demand: This requires only one feature or element to be identified]
1b		Identifying a Claim	Engaging in argument requires the ability to identify the elements of an argument. The most basic of these is what claim is being advanced. This may be in the form of a hypothesis. Without this, it is impossible to engage in argument. Undertaking this process requires a knowledge of what the features of a claim are, i.e., a tentative statement about the material or social world [Cognitive Demand: This requires only one feature or element to be identified]
1c		Identifying data or a reason supporting a claim	This level requires students to analyze text or spoken discourse being used to support an argument and identify the reasons that are being used to justify the significance of such data. It requires the ability to identify the claim first and then the reasons that are being used to support the claim. [Cognitive Demand: This requires only one feature or element to be identified]
2a	Constructing an argument with data or reasons.		This ability builds on the previous level in that it requires the ability to make a claim, select reasons or data that might support that argument, and then construct a synthesis between the claim and the warrant. The focus of interest here is whether students can justify the claims that they advance. This requires students next to construct a warrant or

			justification for how that piece of data supports a claim. This can either take the form of: a) Giving a relational rule, e.g., metal will sink as heavy things sink and light things float. This is essentially a form of deductive argument making use of a general rule or an established rule. Object A is heavy; all heavy objects sink; therefore object A will sink. b) Making a generalization: This is an argument which draws on N data samples to make a generalization, e.g., object A, object B and object C are all made of glass. All three sank. Hence, all glass objects sink. Using an analogy or model to make a prediction. For instance, using the behavior of water waves to construct an argument for what will happen to sound waves when they hit a solid object. Models and analogies are used as the basis of 'If…then…therefore…' reasoning, e.g., if gases consist of tiny particles in constant collision, then when the space they occupy is halved, they will hit the walls twice as often and the pressure will double. Or if the atom is like a mini solar system (analogy) then when it is bombarded with high velocity helium nuclei, most of these should go straight through. [Cognitive Demand: This requires only two features or elements to be identified]
2b	Constructing a Rebuttal	(Identifying weak arguments)	This is the ability to construct a rebuttal because the deductions may be flawed; there may be exceptions or qualifications to the argument or the evidence contradicts a given hypothesis. Rebuttals can only be constructed if students can identify claims and their supporting justification. For instance, that heavy things sink and light things sink is not supported by the fact that iron boats float. The construction of a rebuttal is reliant on the ability to identify a claim with data or observations and come build an alternative argument. [Cognitive Demand: This requires two features or

			elements of an argument to be identified]
3a	Making a Comparative Argument		Here the process required is one of evaluation, which requires two arguments to be juxtaposed and compared. This is the basis of inferential (abductive) arguments, which require a process of evaluating which is the most likely or best explanation. This form of argument relies on being able to identify why other arguments are potentially flawed. In this form of argument ideas are evaluated not in isolation but by comparing them to other ideas. This is the form of argument used by historians and requires relational based reasoning where one idea is compared with another. This makes it harder. In essence this type of reasoning requires the comparison of two arguments and an evaluation of their competing merits. It is possible to test elements of this facility separately by asking students to evaluate two competing explanations and see if they can provide reasons for why one is better than another. [Cognitive Demand: This requires three or more features or elements to be identified]
3b		Identifying relative significance of several pieces of evidence.	Not all evidence is equally important. This is the ability to identify that certain pieces of evidence are more important than others. For instance, what possible reason could there be for wearing sunglasses if vision was active? Sunglasses would serve no purpose. To undertake this requires the relation between each piece of evidence and the claim to be identified and then an evaluative judgment made of the quality of the evidence and the relative importance of one piece of evidence with another. [Cognitive Demand: This requires three or more features or elements to be identified]
3c	Constructing a counter claim with justification		Of itself, this is a harder task, as it requires the individual to construct an alternative explanatory hypothesis, compare it with the existing argument and evaluate why it is superior. The latter

		requires a reason for why it offers greater explanatory coherence and to refute alternative theories or arguments. [Cognitive Demand: This requires three or more features or elements to be identified]

Appendix B: Argumentation Items

(A1) Desalinating Water

California is increasingly short of fresh water. Lack of fresh water limits the amount of food that can be grown. Clearly, there is a need to better manage this increasingly valuable resource. 97% of the world's water is in the oceans. The trouble is that it is salty and cannot be drunk. The process of removing salt from water is called *desalination*.

One method of desalination is to use energy to heat seawater so that it evaporates leaving the unwanted salt behind. The water vapor is then passed over a cold object where it turns back to pure water. A great deal of energy is lost during this process.

Another method involves pushing the seawater through a sheet with tiny holes in it called a membrane. Only the water particles pass through leaving all the salt on the other side. This uses half the energy but the membranes get clogged and have to be cleaned regularly. All the sludge produced then has to be taken away.

3a. What is the main claim that is being made in this piece?

3b. What are two pieces of data that are used to support the argument for desalination?

3c. What is the main argument against the standard method of desalinating water?

3d. What is the main argument against using membranes?

3e. Using only the information given above, which method do you think is the best one? Explain why you think this is the best method.

Appendix B

(A2) What happens to sugar in water?

Three students are discussing what happens to sugar grains when the sugar is added to a cup of hot water.

Paul says: *The sugar breaks up and disappears for good.*

Mary says: *The sugar breaks up but is still there and has not changed.*

Laura says: *The sugar has changed to make a new substance.*

They have several pieces of evidence for their argument.

1. The water tastes sweet.
2. After stirring the sugar can no longer be seen.
3. They leave the cup for several days. All the liquid evaporates leaving a sticky, solid substance at the bottom of the cup. This tastes sweet but is one large lump rather than separate grains.
4. They weigh the cup of water and the sugar separately. The weight of the cup of water and the sugar added together is the same as the weight of the cup of water with the sugar added.
5. They have been told that you cannot destroy or make new matter.

5a. Which evidence do you think best supports Paul's argument?

(a) Number ____

(b) Explain why you think this evidence supports Paul's argument.

Appendix B

5b. Which evidence challenges Paul's argument?

(a) Number____

(b) Explain why you think this evidence challenges Paul's argument.

5c. Which piece of evidence best supports Mary's argument?

(a) Number ____

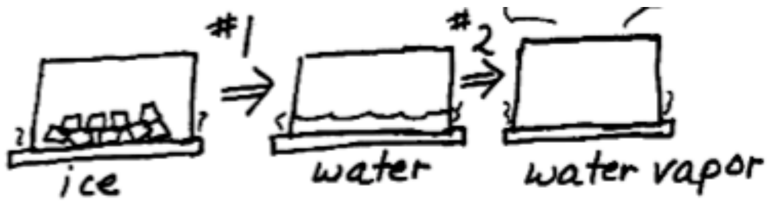
(b) Explain why you think this evidence supports Mary's argument.

5d. Laura claims that the evidence available is not enough to decide whether she or Mary is right. Do you agree? Why?

Appendix B

<u>(A3) Heating Ice to Water Vapor</u>

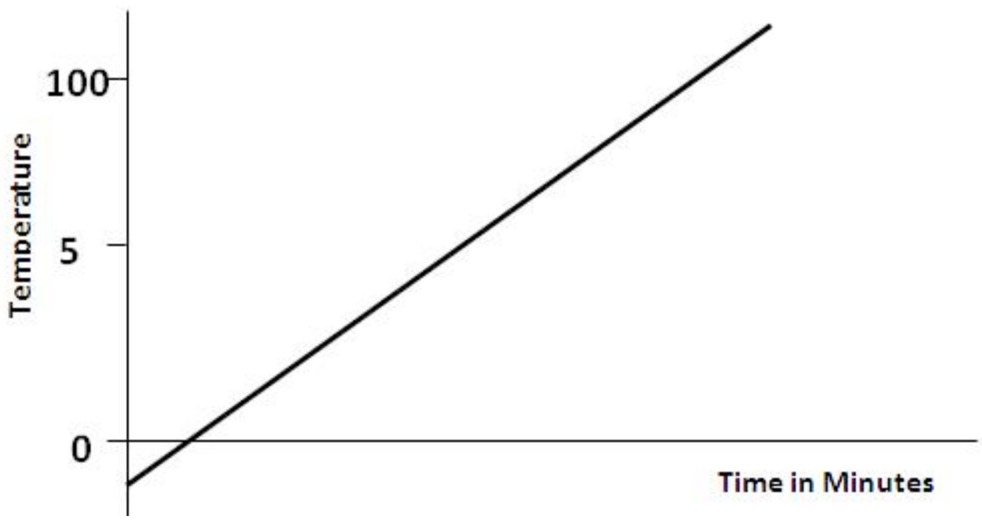
Evan and Anna are wondering what happens when ice in the container below is heated and boils.



0°C 50°C 100°C

Evan thinks that the temperature will increase like graph A. While Anna thinks it will be more like graph B. This is what they each say about their graphs:

Graph A: Evan's Prediction



Evan: The temperature in the beaker will increase steadily. Heat is supplied by the electric plate at a constant rate. This makes the water particles move faster. My measurements show that ice melts at 0°C and boils at 100°C. When they are moving fast enough they will break free and become vapor. This happens when the water boils. The temperature will then carry on increasing steadily.

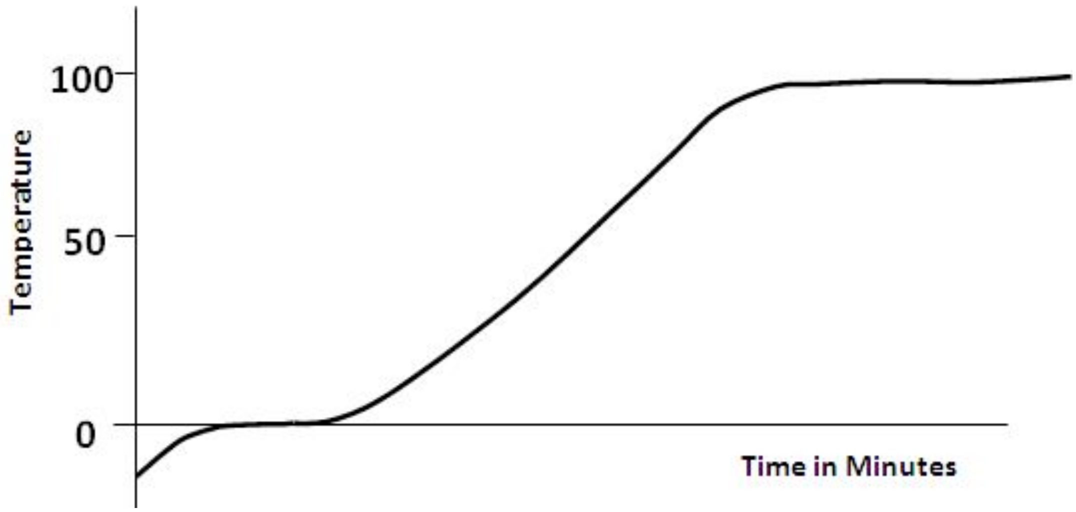
14a. For Evan's argument

a. Underline the claim that he is making.

b. Double underline the data he is using to support their argument.

c. Give one reason that he uses to justify his argument.

Graph B: Anna's Prediction



Anna: The temperature in the beaker will be steady at two points – first as the ice turns to water, which we measured to be at 0°C. It will also be steady at 100°C when it boils. She says that ice does not melt instantly and boiling water takes time to disappear. As the ice is turning to water, energy is used to break the bonds holding the water particles together. When this is happening the temperature of water does not increase. The same thing happens when it turns to vapor. Energy is needed for the water particles to escape from the liquid.

14b. For Anna's argument

a. Underline the claim that she is making.

b. Double underline the data she is using to support their argument

c. Give one reason that she uses to justify her argument.

14c. Evan then says that Anna must be wrong. After all, if the heater is on and warming the water, then how can the temperature not be increasing? Anna thinks this is a fair point but that it is wrong. What do you think she says?

Appendix B

(A4) Something in the Air?

Ted and Alexis are wondering where the water on the outside of the glass of water with ice comes from.

They each have a different idea:

Ted: The water came over the top of the glass.
Alexis: The water came from the air.

7a. How would you argue that **Ted** is wrong?

____

____

____

7b. How would you argue that **Alexis** is wrong?

____

____

____

Appendix B

7c(a). Ted thinks of a way to test who is right. He wipes the water off the glass and measures the height of the water in the glass. He returns an hour later and notes that drops of water have returned to the outside of the glass while the height of the water inside the glass has not changed. Whose claim does this support - Ted or Alexis?

7c(b). Explain why you think it supports Ted or Alexis.

7d. Ted now argues that the water came through the glass. Alexis is not convinced. What do you think she says?

Appendix C: Outcome Space

A1-3a			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer with appropriate reasoning. *Student report of the main claim includes something referring to a water shortage AND something referring to how the water shortage is to be addressed (i.e., desalination).*	"California has a water shortage and the oceans can help solve this problem if we can filter the salt out of the ocean water."	"California's water problem and how seawater can be used to make freshwater." 40001: "That California needs more water and a better way to make pure water without using as much energy."
1	Reasoning is largely correct, i.e., correct answer with inadequate reasoning or good reasoning that is used to justify an incorrect answer. *Student refers to EITHER something about a water shortage OR something about how to address a water shortage, BUT NOT BOTH.*	"California has a water shortage." OR "We can get water from the oceans by removing the salt from it."	"The main claim that is being made is the methods to use to drink fresh water out of sea water." 40401: "California is lacking fresh water because most of it is salty."
0	Answer fails to address the question. *Student refers to NEITHER something about a water shortage NOR something about how to address a water shortage.*	"The main claim being made in this piece is that [any isolated fact not linked to the main idea, e.g., 97% of the world's water is in the oceans]."	"97% of the world's water is in the oceans."
9	No Response		

Appendix C

A1-3b			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *Two correct pieces of data are identified.*	Two pieces of data that are used to support the argument for desalination are any two of the following: ● "California is increasingly short of fresh water" ● "97% of the world's water is in the oceans" ● water in the ocean is too salty and cannot be drunk ● "living standards are lowered" ● "lack of fresh water limits how much agriculture and industry can be supported"	"Lack of fresh water limits how much agriculture and industry can be supported and lowers living standards. Water in the ocean is too salty and cannot be drunk." 40033: "Lack of fresh water decreases/limits the amount of food that can be grown. 97% of the world's water is in the oceans that are salty and non-drinkable water."
1	Partially correct answer. *Only one correct piece of data is identified.*	Student identifies <u>only one</u> of the following: ● "California is increasingly short of fresh water" ● "97% of the world's water is in the oceans" ● water in the ocean is too salty and cannot be drunk ● "living standards are lowered" ● "lack of fresh water limits how much agriculture and industry can be supported"	"To use energy to heat seawater so that it evaporates leaving the unwanted salt behind and involves pushing the seawater through a sheet with tiny holes in it called a membrane." 40030: "97% of the Earth is water. Not all is drinkable. We need more of this precious resource." 40002: "Two pieces of data are the two different methods of desalination and the fact that most of the water on Earth is from the ocean."
0	Answer fails to address the question.	Student identifies <u>none</u> of the following: ● "California is	"Salty water can't be drunk. Fresh water are evaporating faster."

	The student fails to identify any of the correct pieces of data.	increasingly short of fresh water” • “97% of the world’s water is in the oceans” • “living standards are lowered” • “lack of fresh water limits how much agriculture and industry can be supported” [any combination of statements that are in fact not data that support desalination, e.g., “clearly, there is a need to better manage this valuable resource”].	40001: “That desalination uses a great deal of energy and that using membranes could take a lot of work.” 40419: “Heating salt and water separates it.” 40006: “Two methods of desalinating water with evaporation and a membrane.”
9	No Response		

Appendix C

A1-3c			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *The student answer mentions energy and compares the two desalination methods in terms of energy.*	"The main argument against the standard method of desalinating water is that it uses twice as much energy as the membrane method."	"That desalinating water will use a lot more energy than the membrane method."
1	Partially correct answer. *The student mentions energy, but uses the concept incorrectly, OR, uses the concept without comparing the two methods.*	"A great deal of energy is lost through desalination."	"This uses half the energy but the membranes get clogged and have to be cleaned regularly." 40414: "It takes too much energy."
0	Answer fails to address the question. *Failure to identify energy.*	"The main argument against the standard method of desalinating water is that [any isolated fact not linked to the main idea, e.g., California is increasingly short of fresh water" or "97% of the world's water is in the oceans]."	"97% of the world's water in in the oceans." 40412: "Using membranes to pull salt out of water is more efficient in desalinating water." (this is well-written, but they in no place specify *energy* – had they said "energy efficient" this would get a score of 2). 40415: "Because it has its disadvantages."
9	No response		

Appendix C

A1-3d			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *The student mentions that the membranes get clogged AND therefore require cleaning.*	"The main argument against using membranes is that [any argument against the membrane method, e.g., the membranes get clogged AND therefore they require cleaning]."	"The membranes get clogged and it gets cleaned regularly." 40004: "It must be cleaned regularly because it gets clogged with sludge that must be taken away." 40003: "The main argument against using membranes is that they get clogged and have to be cleaned regularly. All the sludge produced then has to be taken away."
1	Partially correct answer. *The student mentions that the membranes get clogged OR that it requires cleaning, BUT NOT BOTH.*	"The main argument against using membranes is that [argument against the standard method, e.g., the membranes get clogged OR it requires cleaning]."	"Membranes get clogged & using half the energy." 40414: "Cause it will clogged up" 40036: "When you use membranes, they must be cleaned regularly, and the sludge must be removed."
0	Answer fails to address the question. *The student mentions NEITHER that the membranes get clogged NOR that they require cleaning.*	"The main argument against the standard method of desalinating water is that [any isolated fact not linked to the main idea, e.g., California is increasingly short of fresh water or 97% of the world's water is in the oceans].	"Membranes are what pushes the seawater through a sheet with tiny holes."
9	No Response		

Appendix C

A1-3e		
Score	Descriptions	Empirical Examples
2	Claim + reason provided in the question prompt. *A good answer would state which they have chosen and provide a relevant reason – that is provide claim plus relevant reason*	41004: "I think the best method is to heat seawater so that it evaporates leaving the unwanted salt behind, because even if this method seems to use more energy, it doesn't matter, because just because method 2 uses less energy, you have to do more work to clean and take away." 41001: "I think that the second method would be the best way, it takes less energy but it has its cons too." 41003: "I think the membrane is the best method because it uses half the energy and you can probably turn the sludge into some kind of alternative fuel."
1	Claim + reason not provided in the question prompt *Here they provide claim but the reason is not in the information provided in the question prompt*	40412 'Using membranes is more efficient the distilling water' (the kind of efficiency is not specified. Energy efficiency? Time efficiency? 41009: "I think the first method is better when can obtain energy from solar panels." 41016: "I think using the one that takes energy is better because you can have artificial energy or sun energy." 40403: "Evaporation, I don't like to clean things [40403]
0	Claim only. *The student makes a claim without any justification or very irrelevant justification*	41005: "The best method." 41007: "To leave it there." 41006: "Both."
9	No Response	

Appendix C

A2-5a			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer with appropriate reasoning. *The student identifies that #2 supports Paul, and justification is provided for why in fact this piece of evidence supports his claim.*	"Number 2 best supports Paul, as he claims the sugar disappears for good, and indeed the sugar can no longer be seen."	40426: [PICKED 2] "Paul thinks that sugar disappears for good so it could no longer be seen." 40025: [PICKED 2] "The sugar is gone and cannot be seen, which appears to have made the sugar be gone forever." "2. Because it says when you stir it, it will be gone. Paul said that when you put in the sugar it disappears." "2. The sugar has disappeared, giving Paul the impression that it is gone for good."
1	Reasoning is partially complete, i.e., correct answer with inadequate reasoning or good reasoning that is used to justify an incorrect answer. *The student identifies that #2 supports Paul, but no justification is provided for why in fact this piece of evidence supports his claim.* *alternatively,* *The student fails to identify that #2 supports Paul, but at least reasoning is provided as to how the student thinks this piece of evidence supports his claim.*	"Number 2 best supports Paul, as [any reasoning that does not justify why Number 2 supports Paul]." *alternatively,* "Number [any piece of evidence other than 2] best supports Paul, as [any reasoning that tries to justify why the evidence they selected supports Paul]."	40410: [PICKED 2] "Because it disappears." 40034: [PICKED 2]: "After stirring the sugar with the water, a new substance is made." (here they select the right piece of evidence, but provide reasoning that actually supports Laura) "2. The sugar dissolves in the cup."
0	Answer fails to address the question. *The student not only fails to identify a piece of evidence that*	"Number [any piece of evidence other than 2] best supports Paul, as [any reasoning that does not justify why the	40402: "I don't know." 40428: [PICKED 1] "When the sugar breaks up it is still sweet."

	supports Paul, but no justification is provided for why in fact they believe the piece of they chose supports his claim.	student believes the evidence they selected supports Paul]."	
9	No Response		

Appendix C

A2-5b			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer with appropriate reasoning. *A correct piece of counter-evidence is identified and reasoning is provided as to how it counters Paul's claim.*	"Number [anything but 2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any piece of evidence besides #2] suggests that something remains."	40002: [PICKED 5] "Paul said the sugar was gone for good but due to the law of conservation, the sugar is still there just that it has dissolved." (note how they tied this evidence back to Paul's claim) "5. You cannot destroy matter, but Paul says the sugar disappeared."
1	Reasoning is largely correct, i.e., correct answer with inadequate reasoning or good reasoning that is used to justify an incorrect answer. *A correct piece of counter-evidence is identified, but reasoning is not provided as to how it counters Paul's claim.* *alternatively,* *A correct piece of counter-evidence is not identified, but at least reasoning is provided as to how the student thinks it counters Paul's claim.*	"Number [anything but 2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that does not justify why the selected evidence does not support Paul]." *alternatively,* "Number [2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that tries to justify why the selected evidence does not support Paul]."	40430: [PICKED 1] "Followed by what Paul said, the sugar breaks up and mixed with the water." 40404: [PICKED 5] "Because you cannot destroy or make new matter." (they did not tie this back to Paul's claim) "4. Because number 4 said that the sugar is still there."
0	Answer fails to address the question. *The student not only fails to identify a piece of evidence that challenges Paul's argument, but no justification is provided for why in fact they believe the piece of they chose counters Paul.*	"Number [2] challenges Paul's argument, as he claims the sugar disappears for good, but in fact [any reasoning that does not justify why the selected evidence does not support Paul]."	40428: [PICKED 2] "Because Paul says the sugar breaks apart and disappears, not stirring the water." "2. Tells more details, and makes his statement sort of clear."
9	No Response		

Appendix C

A2-5c **Combination of A2-5ca and A2-5cb as shown on the following two pages**		
Score	Descriptions	Empirical Examples
2	Correct answer with appropriate reasoning. *The student identifies a piece of evidence that supports the idea that sugar is left behind (#1 or #3), and justifies the evidence they selected.*	[Scored 2s on both A2-5ca and A2-5cb] 40401" [PICKED 1]: "The water is unnaturally sweet, so the sugar must still be in there."
1	Reasoning is largely correct, i.e., correct answer with inadequate reasoning or good reasoning that is used to justify an incorrect answer. *A correct piece of evidence is identified, but reasoning is not provided as to how it best supports Mary's claim.* *<u>alternatively</u>,* *A correct piece of evidence is not identified, but at least reasoning is provided as to how the student thinks it best supports Mary's claim.*	[Partially correct on A2-5ca and A2-5cb] 40404: [PICKED 5] "Because it breaks up, but it is still there." "4. The measurement shows that the weight of the mug and the sugar are the same. Mary says that the sugar changes in appearance, but it is still there." "5. You can't make a new substance out of sugar. It remained to be sugar."
0	Answer fails to address the question. *Neither a correct piece of evidence is identified, nor appropriate reasoning is provided.*	[Got no credit on both A2-5ca and A2-5cb]
9	No Response	

Appendix C

A2-5c-a			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *#1 or #3. The student identifies a piece of evidence that supports the idea that sugar is left behind.*	"Number [1 or 3] best supports Mary's argument."	(1) or (3)
1	Partially correct answer. *#4 or #5. The student identifies a piece of evidence that, while perhaps supporting the notion that something is left behind, it is not clear if that "something" is still sugar.*	"Number [4 or 5] best supports Mary."	(4) or (5)
0	Answer fails to address the question. *#2. The student identifies the piece of evidence that most weakly supports the idea that sugar is left behind (it actually best supports Paul's argument).*	"Number [2] best supports Mary."	[No empirical examples found]
9	No Response		

Appendix C

A2-5c-b			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *The student justifies why the evidence they selected in A2-5c-a suggests that sugar hasn't changed.*	[PICKS 1] "The sweetness indicates that the substance left behind is still sugar, therefore the substance hasn't changed."	40401" [PICKED 1]: "The water is unnaturally sweet, so the sugar must still be in there." 40432: [PICKED 4]: "The individual mass of the sugar and the water equal their combined mass, meaning the sugar never changed."
1	Partially correct answer. *The student identifies the relevant part of the evidence but fails to show how it in fact justifies that the sugar has not changed.*	[PICKS 1] "The substance left behind is sweet."	40404: [PICKED 5] "Because it breaks up, but it is still there." "4. The measurement shows that the weight of the mug and the sugar are the same. Mary says that the sugar changes in appearance, but it is still there." "5. You can't make a new substance out of sugar. It remained to be sugar."
0	Answer fails to address the question. *The student identifies a feature of the evidence that is irrelevant to the question of whether or not the substance has changed.*	"All the liquid evaporated from the mug."	40418: [PICKED 4] "The sugar was still sugar." 40413: {PICKED 4] "It says the same thing."
9	No Response		

Appendix C

A2-5d			
Score	Descriptions	Mock Student Responses	Empirical Examples
3	Recognize uncertainty AND provide reasoning. *The student recognizes that the evidence is not conclusive and makes a case for why this uncertainty does not allow them to side with either Laura or Mary*	"Laura is correct that the evidence is not conclusive because while [anything but 2] supports the notion that something is left behind, it is not clear based on this evidence whether or not what is left over is a new substance."	"[Yes – implied]. Because there are many things that could had happened to the sugar." 40426: "Yes, because the evidence given doesn't support either of their hypotheses." 40358: "I believe that there not enough evidence to prove whether Mary or Laura is right. Because they do say it's still there but not [whether] the substance has changed or not." 40393: "The substance that is left after the water evaporates may be sugar, or an entirely different substance do to a chemical reaction."
2	Recognize uncertainty but DO NOT provide reasoning. (this gets a higher score than students that provide reasoning but DO NOT recognize uncertainty, because on our argumentation progress map, recognizing uncertainty in evidence is considered to be more sophisticated than providing reasoning to back a claim) *The student recognizes that the evidence is not conclusive, but does not make a case for why this uncertainty does not allow them to side with either Laura or Mary.*	"I think Laura is right that you can't decide between her argument and Mary's argument."	40370: "Yes, because it could change every time." 40356: "Yes, because it wasn't tested."

Appendix C

1	Provide reasoning but DO NOT recognize uncertainty. *Here students, failing to recognize the uncertainty in the evidence, conclusively side with either Laura or Mary, but at least they provide a plausible reason why they chose to side with Laura or Mary.*	[The student sides with either Laura or Mary and provides reasoning as to why].	40403: "I don't because that have been told when you mix sugar and water it doesn't make a new substance, therefore Mary is correct." "Yes, the evidence is conclusive because it is true that the sugar disappeared."
0	DO NOT provide reasoning but DO NOT recognize uncertainty. *The student not only fails to see the uncertainty in the evidence, but they are unable to correctly reason for the claim (i.e., side with either Laura or Mary) in which they feel the evidence is indeed sufficient.*	[The student sides with either Laura or Mary but fails to provide reasoning as to why].	40392: "No because." 40352: "I think Mary's argument is right because she says the sugar breaks up but its still there but it hasn't changed." (note that this is just a regurgitation of Mary's hypothesis, not an attempt to justify it) "Well, I'm not really sure, but everyone has their own opinion." 40419: "No, because I know for a fact water breaks up and dissolves."
9	No Response		

Appendix C

A3-14a-a			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *Student underlines, <u>within a word or two</u>, "The temperature in the beaker will increase at a steady rate" AND/OR "The temperature will then carry on increasing steadily."*		
1	Partially correct answer. *Student underlines <u>some part</u> of "The temperature in the beaker will increase at a steady rate" AND/OR "The temperature will then carry on increasing steadily."*		
0	Answer fails to address the question. *Student underlines <u>no part</u> of "The temperature in the beaker will increase at a steady rate" AND/OR "The temperature will then carry on increasing steadily."*		
9	No Response		

Appendix C

A3-14a-b			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *Student double underlines, within a word or two, either "My measurements show that ice melts at 0°C and boils at 100°C" or "Heat is supplied by the electric beaker at a steady rate."*		
1	Partially correct answer. *Student double underlines some part of either "My measurements show that ice melts at 0°C and boils at 100°C" or "Heat is supplied by the electric beaker at a steady rate."*		
0	Answer fails to address the question. *Student double underlines no part of either "My measurements show that ice melts at 0°C and boils at 100°C" or "Heat is supplied by the electric beaker at a steady rate."*		
9	No Response		

Appendix C

A3-14a-c			
Score	Descriptions	Mock Student Responses	Empirical Examples
3	*The student writes a response that indicates any of the points in the adjacent box are a reason:*	• "Heat is supplied by the electric plate at a steady rate."	40539: "Heat is supplied by the electric plate at a constant rate."
2	*The student writes a response that indicates any of the points in the adjacent box are a reason:*	• "This makes the water particles move faster." • "When they are moving fast enough they will break free and become vapor. This happens when the water boils."	"He is saying the molecules are moving more quickly, making the water evaporate." 40506: "Water vapor is evaporated water and moves quick." (this answer seems to capture both of the bullet points) 40522: "When they are moving fast enough, they will break free and become vapor." 40520: "One reason he uses to justify his argument is that the heat will, make the particles move faster and become free to make vapor."
1	*The student provides reasoning that <u>partially</u> justifies their choices (any reasoning that is <u>partially</u> consistent with the mock student responses for Level 3 or Level 2 reasoning listed above).* *They at least gave reasoning a shot.*	"A reason Evan uses to justify his argument is [any reasoning that is PARTIALLY CONSISTENT with the mock student responses for Level 3 or Level 2 reasoning listed above, e.g., because of his data]."	"He uses the steady rate and ^{o}C." 40540: "When something turns into gas it rises." 40519: "The temperature has increased because it got hotter as it boiled."

Appendix C

0	*The student fails to provide reasoning that justifies Evan's argument (i.e., none of their reasons are in line any of the mock student responses for Level 3 or Level 2 reasoning listed above).* *No reasoning is apparent.*	"The graph says so." (note this is a zero because the graph is Evan's prediction, not a reason used to support his prediction)	40502: "Because his argument is true." 40518: "He's using the graph to show us the data." "Evan thinks he's right because of his data."
9	No Response		

Appendix C

A3-14b-a			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *Student underlines, within a word or two, "The temperature in the beaker will be steady at two points—first as ice turns to water and when it boils."* *The student can underline the whole two sentences: "The temperature in the beaker will be steady at two points—first as the ice turns to water, which we measured to be at 0°C. It will also be steady at 100°C when it boils."*		
1	Partially correct answer. *Student underlines some part of "The temperature in the beaker will be steady at two points—first as the ice turns to water, which we measured to be at 0°C. It will also be steady at 100°C when it boils."*		
0	Answer fails to address the question. *Student underlines no part of "The temperature in the beaker will be steady at two points—first as ice turns to water and when it boils."*		
9	No Response		

Appendix C

A3-14b-b			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *Student double underlines, within a word or two, "Which we measured to be at 0°C. It will also be steady at 100°C when it boils."*		
1	Partially correct answer. *Student double underlines some part of "Which we measured to be at 0°C. It will also be steady at 100°C when it boils."*		
0	Answer fails to address the question. *Student double underlines no part of "Which we measured to be at 0°C. It will also be steady at 100°C when it boils."*		
9	No Response		

Appendix C

A3-14b-c			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *The student writes a response that indicates any of the points in the adjacent box are a reason:* *The student says something about energy and or bond breakage at a phase transition (i.e., from ice to water or from water to water vapor)*	• "As the ice is turning to water energy is used to break the bonds holding the water particle together so their temperature does not increase." • "The same thing happens when it turns to vapor. Energy is needed for the water particles to escape from the liquid."	40528: "She says that during the ice to water process, the energy used should not increase the temperature." 40529: "As ice is turning to water, energy is used to break the bonds holding the water particles together. When this happens the temperature of water doesn't increase." 40539: "Energy is needed for the water particles to escape from the liquid." 41003: Since ice doesn't melt instantly the temperature will not increase at a steady pace.
1	Partially correct answer. *Student at least talks about one or both of the phase transitions, but does not mention energy and/or bond breakage*	"A reason Anna uses to justify her argument is [any reasoning that is PARTIALLY CONSISTENT with the mock student responses above, e.g., because of her data]."	40507: "She says ice doesn't melt instantly." 40501: "She made it for 3 point, ice, melt point, and boling point. When it go up to 100^0C, it stay. 41015: "Ice turns into water at 0 degrees Celsius" (this is a WEAK Level 1 response)

Appendix C

0	Answer fails to address the question. *Student response has nothing to do with the phase transition.*	"A reason Anna uses to justify her argument is [any reasoning that is NOT CONSISTENT with the mock student responses for Level 4 reasoning listed above]."	40502: "Is that he is true." 40513: "She is giving us information on the data that she has." 40519: "The water is moving faster but then slows down." 41012: "Ice will not melt quickly in boiling water." (here the student conflates the two phase transitions).
9	No Response		

Appendix C

A3-14c			
Score	Descriptions	Mock Student Responses	Empirical Examples
2	Correct answer. *Student provides a MICROSCOPIC response from which it can reasonably be inferred that the student understands that phase changes require energy.* *More specifically, they need to know WHY energy goes into phase changes (i.e., bond breaking).*	"Anna might say that the same amount of heat being transferred to the water is being used right away to break apart the water molecule bonds."	"Water particles in ice are locked in one position, so it does need energy to break the bonds into water and the temperature will stay the same." 40572: "She says the temperature doesn't increase, because energy is needed to break bonds. That is when it stays the same." 40043: "Anna might say that since energy is needed for the bonds to break, energy is exerted from the ice, thus balancing the heat the heater provides."
1	Partially correct answer. *Student provides an entirely MACROSCOPIC response from which it can reasonably be inferred that the student understands that phase changes require energy.* *More specifically, they need to know that energy goes into phase changes, not necessarily WHY – knowing WHY energy goes into phase changes (i.e., bond breaking) is a Level 2 response.*	"Anna might say that at the times when ice changes to water and when water changes to vapor, the temperature does not change."	"I think she says it takes time to melt ice, it just doesn't heat up in a second." 40013: "I think Anna says that even though the heater is on, it takes time to have the temperature increase because water keeps on evaporating." 40563: "When water turns to vapor, the temperature does not increase." 40014: "The heat will not increase because the water

Appendix C

			is using the energy to evaporate which cools the water while it is being heated." (they understanding the constant heat from the heater is being diverted into the evaporation phase change, but they do not explain why on a microscopic level) 40036: "It takes energy to melt ice and evaporate water, so it would take time to do that." 40560: "The heat from the water is being used to melt the ice and evaporate water instead of increasing the temperature."
0	Answer fails to address the question. *The student fails to provide reasoning to address Evan's question (e.g., "if the heater is on and warming the water, then how can the temperature not be increasing?").*	"Anna might say that [regurgitation of any fact used in the passage that does not pass as original reasoning]."	"I think she will say that the heat will not increase until the heat goes over the whole house." "That he proved his point. But she tried." 40018: "I think she says that because the ice is cold, it cause the water to not heat up, and it takes a while for the heat to overcome the coldness of the ice. 40576: "Because raising the temperature takes time." 40559: "She says that the ice cools down the burner."

			40562: "I think Anna is the one that is right. Her evidence is more supported." 40003: "I think she says that the water of bonds will be breaking." ("water of bonds" does not make sense…note if they said "water bonds will be breaking" and that this required some of the energy supplied by the heater, this would be Level 2) 40002: "She says that even if the plate is increasing at a constant rate the heat also at the same temperature." 40017: I think she says that might be at first heat up slowly and then faster because it starts out at a different temperature
9	No Response		

Appendix C

A4-7a			
Levels	Description	Mock Student Responses	Examples of Student Responses
2	Student provides a reasonable case for why Ted was wrong.	"Water cannot climb up on its own." "There is nothing to lift the water up."	40604: "If it came from over the glass, there would be an amount of water loss." 40608: "Nobody touched the glass cup in order to make it shake." (indeed, shaking would be one way that water cold come over the top – the student makes a speculation)
1	Student provides partial/incomplete reasoning for why Ted is wrong. Essentially, the students give reasoning a shot.	"Water does not escape from containers."	40601: "How did it go over the glass?" 40616: "I'll ask how did the water from inside of the cup come over the top of the glass." 40617: "Because you don't always fill your cup all thew way to the top, you could have just a little bit and it still does that." 40606: "Because it come from inside the cold glass water."
0	No reasoning is provided.	"Because Ted is wrong."	
9	No response		

Appendix C

A4-7b			
Levels	Description	Mock Student Responses	Examples of Student Responses
2	Student provides a reasonable case for why Alexis was wrong.	“Air is a gas and does not contain water.” “If water was in the air we would get wet.”	40602: “Because the air is not cold enough to make it form.” 40609: “You can’t because the ice is cold and it turns the moist air into water.”
1	Student provides partial/incomplete reasoning for why Alexis is wrong.	“Water comes from the ground and not the air.’” ‘”Air is clear.”	40619: “Alexis answer is wrong because water can’t come from air.” 40610: “That the air can’t just become water and stick on the glass.”
0	No reasoning is provided.		
9	No response		

Appendix C

A4-7c			
Levels	Description	Mock Student Responses	Examples of Student Responses
3	Student chooses Alexis and provides a reasonable explanation	This supports Alexis as, if the water had come over the top, the level of water would have gone down. It did not The water level did not change, yet there are water droplets on the glass again, hence the water must not have come from inside the glass, which refutes Ted's claim.	41003: I think that Alexis is right because if the water level has not changed how does it come from the top of the glass. 41001: The water from the air returned, and was not water that spilled over. Something is attracting that water to the cup.
2	They provide partial/incomplete reasoning for why they think Alexis's claim is supported. Essentially, they correctly side with Alexis and give reasoning a shot.	The water level did not go down, which supports Alexis.	41011: Alexis said the water air comes from the air. 41018: because the air must have been hot.
1	Student provides a reasonable case for why they think Ted's claim is supported	The water level was not checked carefully enough, meaning Ted could still be right.	41019: Ted because air is part of H20 so water falls into the cup but we can't see it.
0	Student sides with Alexis but no reasoning is provided. Student provides partial/incomplete reasoning for why they think Ted's claim is supported. Essentially, the students give reasoning a shot.	It supports Alexis. I think it supports Ted because Ted is smarter than Alexis.	41010: Because she is right.
9	No response		

Appendix C

A4-7d			
Levels	Description	Mock Student Responses	Examples of Student Responses
2	Alexis points to the water level not changing.	Alexis says that this is wrong because if the water came through the glass, the water level would decrease over time. Ted is incorrect because how could some of the water escape through the glass and yet the water level inside the glass not change?	40580: “Alexis might say that the water cannot pass through the glass and the water inside has not moved.” 40581: “If it came through the glass, the height of the water would be lower which it isn’t.”
1	Student provides partial/incomplete reasoning for why Alexis is correct in not being convinced. OR why they think Ted’s new claim (the water came through the glass) is incorrect. In either case, the student gives reasoning a shot.	Ted is incorrect because the water can’t come through the glass. Alexis shouldn’t be convinced because Ted is wrong.	41022: Water cannot go through a sealed environment. 41026: No water can come through the glass when the glass is blocking the water. 40563: “Water can’t got through the glass unless there are cracks.” 40564: “There are no holes in the glass.” 40572: “She says there are no holes in it and if so the amount of water inside would still be less.” 40582: “How can a liquid pass through a solid.” 40574: “A liquid could not pass thorugh a solid because the amounts are tightly pressed together.”

Appendix C

0	No reasoning	I don’t know but Alexis is right. Everyone is entitled to their own opinion.	41020: I don’t agree with you. 41008: How does it come through the glass?
9	No response		

Appendix D: Average Ability of Students within Each Score Category

Item A1_3a					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	75	0	-0.33	0	-0.61
1	191	1	0.03	1	0.17
2	81	2	0.21	1	0.25

Item A1_3b					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	245	0	-0.13	0	-0.19
1	31	1	0.02	0	-0.06
2	71	2	0.39	1	0.79

Item A1_3c					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	97	0	-0.40	0	-0.76
1	243	1	0.14	1	0.34
2	7	2	0.15	1	-0.36

Item A1_3d					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	65	0	-0.72	0	-1.09
1	95	1	0.01	1	0.19
2	187	2	0.23	1	0.31

Item A1_3e					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	75	0	-0.57	0	-0.92
1	108	1	0.12	1	0.23
2	164	2	0.16	1	0.31

Appendix D

Item A2_5a					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	54	0	-0.80	0	-1.23
1	129	1	-0.01	1	0.11
2	164	2	0.25	1	0.36

Item A2_5b					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	58	0	-0.68	0	-1.14
1	176	1	0.01	1	0.16
2	113	2	0.32	1	0.39

Item A2_5c					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	26	0	-0.73	0	-1.23
1	271	1	0.02	1	0.08
2	50	2	0.20	1	0.32

Item A2_5d					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	157	0	-0.27	0	-0.37
1	82	1	0.30	1	0.78
2	80	2	0.05	0	-0.16
3	28	3	0.38	1	0.47

Item A3_14aa					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	140	0	-0.42	0	-0.72
1	22	1	0.10	1	0.25
2	185	2	0.29	1	0.55

Appendix D

Item A3_14ab					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	121	0	-0.42	0	-0.72
1	96	1	0.19	1	0.40
2	130	2	0.23	1	0.42

Item A3_14ac					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	161	0	-0.35	0	-0.58
1	47	1	0.07	0	0.02
2	68	2	0.27	1	0.67
3	71	3	0.46	1	0.76

Item A3_14ba					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	136	0	-0.29	0	-0.57
1	211	1	0.17	1	0.40

Item A3_14bb					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	250	0	-0.08	0	-0.13
1	94	1	0.16	1	0.41
2	3	2	0.24	1	0.38

Item A3_14bc					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	172	0	-0.29	0	-0.48
1	111	1	0.19	1	0.33
2	64	2	0.40	1	0.81

Item A3_14c					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	309	0	-0.05	0	-0.09
1	18	1	0.25	1	0.72
2	20	2	0.46	1	1.09

Appendix D

Item A4_7a					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	69	0	-0.40	0	-0.70
1	88	1	-0.08	1	-0.07
2	190	2	0.17	1	0.32

Item A4_7b					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	62	0	-0.38	0	-0.75
1	134	1	-0.01	1	0.07
2	151	2	0.15	1	0.29

Item A4_7c					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	147	0	-0.41	0	-0.61
1	5	1	-0.27	0	-0.19
2	42	2	0.07	1	0.36
3	153	3	0.37	1	0.54

Item A4_7d					
Code	# of Students	Score: PCM	Average Ability: PCM	Score: OPM	Average Ability: OPM
0	119	0	-0.44	0	-0.71
1	163	1	0.13	1	0.27
2	65	2	0.43	1	0.71

Appendix E: Multidimensionality Analysis

Table 1. What Argumentation Items Measure Changes of State Content Construct (B. Henderson, personal communication, March, 2013)

Item #	Bundle	Item Name	Measure Changes of State Content Construct Or Not (Y=Yes; Blank=No)
1	A1.	A1-3a	
2	Desalinating Water	A1-3b	
3		A1-3c	
4		A1-3d	
5		A1-3e	
6	A2.	A2-5a*	Y
7	Sugar in the Water	A2-5b*	Y
8		A2-5c*	Y
9		A2-5d	Y
10	A3.	A3-14aa	Y
11	Ice to Water Vapor	A3-14ab	Y
12		A3-14ac	Y
13		A3-14ba	Y
14		A3-14bb	Y
15		A3-14bc	Y
16		A3-14c	Y
17	A4.	A4-7a	Y
18	Something in the Air	A4-7b	Y
19		A4-7c*	Y
20		A4-7d	Y

Note. Items with * are those with two parts: the first part asks students to select one of the choices given in the item prompt and the second part asks students to provide an explanation. One overall score was given and used in the subsequent analysis.

Appendix E

Table 2. Result of the Unidimensional DIF Model: Each Argumentation Bundle Was Scored as One Single Partial Credit Item (12 Content Items and Four Argumentation Items)

Item	Type	DIF Effect*	Standard Error	Effect Size
291	Content	0.38	0.03	Large
219b	Content	-0.09	0.03	Ignorable
222	Content	-0.13	0.03	Ignorable
6	Content	0.15	0.03	Ignorable
117	Content	0.13	0.03	Ignorable
205	Content	0.30	0.03	Intermediate
225b	Content	-0.25	0.03	Intermediate
113	Content	0.03	0.03	Ignorable
2c	Content	0.56	0.03	Large
2a	Content	-0.30	0.03	Intermediate
286_7	Content	0.02	0.03	Ignorable
5b	Content	0.06	0.03	Ignorable
A1	Argumentation	-0.18	0.03	Ignorable
A2	Argumentation	-0.25	0.03	Intermediate
A3	Argumentation	-0.14	0.02	Ignorable
A4	Argumentation	-0.29	0.12	Intermediate

* The group-by-item interaction term reported in this table is for the reference group.

Appendix E

Table 3. Result of the Multidimensional DIF Model: Each Argumentation Bundle Was Scored as One Single Partial Credit Item (12 Content Items and Four Argumentation Items)

Item	**Type**	**DIF Effect***	**Standard Error**	**Effect Size**
291	Content	0.27	0.10	Intermediate
219b	Content	-0.05	0.07	Ignorable
222	Content	-0.16	0.09	Ignorable
6	Content	0.12	0.09	Ignorable
117	Content	0.11	0.09	Ignorable
205	Content	0.26	0.10	Intermediate
225b	Content	-0.19	0.05	Ignorable
113	Content	0.02	0.09	Ignorable
2c	Content	-0.06	0.13	Ignorable
2a	Content	-0.40	0.11	Large
286_7	Content	0.08	0.07	Ignorable
5b	Content	-0.01	0.30	Ignorable
A1	Argumentation	0.01	0.03	Ignorable
A2	Argumentation	-0.04	0.03	Ignorable
A3	Argumentation	0.08	0.02	Ignorable
A4	Argumentation	-0.05	0.04	Ignorable

* The group-by-item interaction term reported in this table is for the reference group.

Appendix E

QUESTION #2A [Broken Airplane]

2A.1 Here is a wooden airplane. If I break it into many pieces, is it still an airplane? Is it still wood? (L1)

A. It is still an airplane and it is still wood.
B. It is still an airplane but it is not wood.
C. It is no longer an airplane but it is still wood.
D. It is no longer an airplane and it is not wood.

QUESTION #291 [Chocolate Bar]

(291) When a chocolate bar melts, the amount of chocolate ____________. L2(b)

A. increases
B. decreases
C. stays the same
D. There is no way to predict.

QUESTION #205 [Breath Mirror]

(205) If someone breathes on a mirror, part of the mirror clouds up. What are you actually seeing when you see the mirror cloud up? L2(b)

A. Water vapor in the form of water droplets from your breath that has cooled.
B. Air that you are breathing out from your lungs.
C. Liquid from your tongue.
D. A combination of molecules from the surrounding air that attach to the mirror.