ABSTRACT

Title of Dissertation: MULTIDIMENSIONALITY IN THE NAEP SCIENCE ASSESSMENT: SUBSTANTIVE PERSPECTIVES, PSYCHOMETRIC MODELS, AND TASK DESIGN

Hua Wei, Doctor of Philosophy, 2008

Dissertation Directed by: Professor Robert J. Mislevy

Department of Measurement, Statistics, and Evaluation

Educational assessments are characterized by the interplay among substantive theories, task design, and measurement models. Substantive theories define the nature of inferences to be made about students and types of observations that lend support to the targeted inferences. Task design represents the schemes for the design of tasks and extraction of evidence from student behaviors in the task situations. Measurement models are the tools by which observations of students' performances are synthesized to derive the targeted inferences.

This dissertation elaborates on the interplay by specifying the entities that are involved and how they work in concert to produce an effective assessment and sound inferences. Developments in several areas are contributing to interest in more complex educational assessments: Advances in cognitive psychology spark interest in more complex inferences about students' knowledge, advances in technology make it possible to collect richer performance data, and advances in statistical methods make

fitting more complex models feasible. The question becomes how to construct and analyze assessments to take advantage of this potential. In particular, a framework is required for understanding how to think about selecting and reasoning through the multivariate measurement models that are now available.

Illustrations of the idea are made through explicating and analyzing the 1996 National Assessment of Educational Progress (NAEP) Science Assessment. Three measurement models, each of which reflects a particular perspective for thinking about the structure of the assessment, are used to model the item responses. Each model sheds light on a particular aspect of student proficiencies, addresses certain inferences for a particular purpose, and delivers a significant story about the examinees and their learning of science. Each model highlights certain patterns at the expense of hiding other potentially interesting patterns that reside in the data. Model comparison is conducted in terms of conceptual significance and degree of fit. The two criteria are used in complement to check the coherence of the data with the substantive theories underlying the use of the models.

MULTIDIMENSIONALITY IN THE NAEP SCIENCE ASSESSMENT: SUBSTANTIVE PERSPECTIVES, PSYCHOMETRIC MODELS, AND TASK DESIGN

By

Hua Wei

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Professor Robert J. Mislevy, Chair
Professor Paul J. Hanges
Professor Gregory R. Hancock
Assistant professor Jeffrey R. Harring
Adjunct assistant professor Amy B. Hendrickson

Acknowledgements

First of all, I would like to thank my advisor, Dr. Mislevy, for his vision, patience, and tremendous support throughout the stages of this dissertation. Without his encouragement and mentorship, all the work would not have been completed. I would also like to thank Dr. Hancock, Dr. Harring, Dr. Hendrickson, and Dr. Hanges for their guidance and insightful suggestions. Discussions with them helped me to focus and clarify my thinking. My gratitude also goes to my fellow students in EDMS, whose friendship and support made my studies much more pleasant and productive than they would otherwise have been. Finally, I want to thank my parents and my husband for their great support along the way.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Advances in cognitive and measurement sciences have inspired development of assessment practices that can address more ambitious questions. Evolving conceptions about how students acquire, organize, and use knowledge offer the potential for richer and more coherent assessments that can better assist learning and teaching. However, this potential can be realized only if the general scientific principles for assessment design are explicated and implemented in assessment applications. Especially, the interplay among substantive, statistical, and operational aspects of an assessment is the foundation for developing an effective assessment that suits the purpose for which it is designed and achieves its potential for informing instruction and learning (National Research Council, 2001; Mislevy, Steinberg, & Almond, 2003).

Essentially, every assessment is characterized by the interplay among substantive theories, patterns in the data, and measurement models (National Research Council, 2001; Wilson, 2005). Substantive theories define the nature of inferences to be made about students, how observations of student performance should be collected and in what task situations, and what aspects of student performance are relevant evidence that lends support to the targeted inferences. Patterns in the data are salient aspects of students' performances that bear evidence about their unobserved proficiencies. They are the target of analysis. Measurement models are the tools by which patterns in the data are analyzed to derive the targeted inferences.

Substantive theories, patterns in the data, and measurement models should be

coordinated to produce an effective and coherent assessment and generate sound inferences. With regard to a measurement model, on the one hand, it should be formulated in a way that is consistent with test developers' substantive theories and appropriate in grain-size for the purpose of the assessment. It is not a haphazard collection of variables randomly appearing in the form of a mathematical function. Rather, the inclusion of variables and distributions in a measurement model and the level of detail at which they are defined are determined by the important substantive relationships in the assessed domain and the theoretical constructs and observations that are involved (Mislevy, Wilson, Ercikan, & Chudowsky, 2002). Although it may be unrealistic to model every subtlety and complexity of the substantive relationships, a simplified version should be reflected in a measurement model. The conceptual significance of a measurement model is determined by the validity of its underlying substantive theories and the strength of linkage between the model and the theories.

On the other hand, a measurement model should be structured to capture the significant patterns in the data (Ibid). Features of task situations and students' responses to the tasks should be closely monitored in the model. Any salient aspect of data that exists in the real-world setting but is left unmodeled will contaminate the validity of the model and distort inferences made through the model. The extent to which a measurement model represents and explains patterns in the data can be signified by statistical data-model fit indices. Model criticism tools, such as tests of person fit or item fit, are available to detect suspected departures of a measurement model from particular aspects of observed data (Wilson, 2005).

In this dissertation, the three pillars of a coherent assessment and their interplay were discussed to illuminate the different approaches to multidimensionality. Multidimensionality is a recurring issue in educational assessments. Essentially, it results from the interaction between features of examinees and features of tasks in the test settings. Three types of multidimensionality were analyzed in the presentation to illustrate ideas. Two of these types are distinguished by Adams, Wilson, and Wang (1997). A test measuring several parallel unidimensional subscales with no common item across subscales is *multidimensional between items*. In contrast, a test measuring several latent dimensions with some items related to more than one dimension is *multidimensional within items*. A third type of multidimensionality is conceptualized by mixture models (Rost, 1990; Mislevy & Verhelst, 1990; Yamamoto & Everson, 1995). Distinctions among the three types of multidimensional test structures reflect different perspectives for thinking about knowledge and learning in the domain of interest, different rationales by which test developers want to characterize students' knowledge and proficiency, and different design choices they make to implement their rationales.

The assessment being analyzed in this dissertation is the 1996 National Assessment of Educational Progress (NAEP) science assessment. In many large-scale assessments like NAEP, unidimensionality is a basic assumption. Item responses are usually analyzed by unidimensional item response theory (IRT) models. The narrative theme supported by unidimensional IRT models is that all persons and items are placed along a single continuum of latent trait and persons' positions on the

continuum indicate their overall propensities to answering items correctly and items' positions indicate their overall probabilities of being answered correctly. Stories can be told in the following form: some students are more likely to give correct responses to all items than other students, and some items are more difficult than others for all students.

However, the assumption of unidimensionality does not hold for the NAEP science assessment, which is designed to cover three content areas: physical science, life science, and earth science. The measurement theme of the NAEP science assessment appears to go beyond a single unidimensional IRT model. Therefore, multivariate modeling techniques merit consideration for assessments like NAEP which are targeted at multiple aspects of students' knowledge, skills, and abilities.

As described in the NAEP 1996 Technical Report (Allen et.al., 1999), each of the test items is classified into one of three fields of science and the three fields of science constitute the scales for score reporting. Creating a scale for each of the three fields of science is consistent with the conception that each discipline of science has its own special ways of knowing and that the patterns of development of competence are unique to some extent within each subject domain (National Research Council, 1996). Having three separate subscales, each of which is associated with a different cluster of items, complies with the definition of between-item multidimensionality. The procedure used in NAEP to model between-item multidimensionality follows two steps. In the first step, a unidimensional IRT model is fit to each subscale to estimate item parameters. In the second step, the parameters of the underlying multivariate

latent space are estimated by having the item parameters fixed at their estimated values. In this study, a multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997) rather than three unidimensional models was fit to the item responses to estimate item parameters as well as parameters of the latent abilities. The primary advantage of fitting a multidimensional IRT model such as MRCMLM is that it provides better estimates of item parameters and it yields consistent estimates of the correlations among the latent abilities (ibid).

Although various fields of science may differ in terms of theories, themes, and factual information, the essentials of learning natural sciences are common across scientific disciplines. This viewpoint is evident in the design of the NAEP science assessment. As specified in the NAEP Science Framework (National Assessment Governing Board, 2000), the NAEP science assessment is created to measure three latent dimensions that cross the three content areas. Inspired by a consideration of the design feature, an exploratory item factor analytic model was applied to the NAEP science data in this study to investigate the number of dimensions that actually underlie students' performances, how each item is differentially associated with each dimension, and what each dimension represents.

In a large-scale science assessment such as NAEP, the examinees come from distinct subpopulations with different background characteristics. Thus, one cannot assume a priori that item difficulties are equal across examinees. In fact, there may be factors that make some items easier for one examinee but harder for another. These factors are the cause of multidimensionality. In recent years, discrete mixtures

of IRT models (for example, Rost, 1990; Mislevy & Verhelst, 1990) are increasingly being used to deal with multidimensionality. These models look at multidimensionality from a different perspective in that they fit models of lower dimensionality—indeed often unidimensionality—but allow the item parameters to vary across subgroups, each of which has a distinct latent ability distribution. These models have proved useful for studying tests from the perspective of more subtle shifts in difficulty due to developmental change, strategy use, and curricular emphasis. In this dissertation, an exploratory analysis using a mixture Rasch model (MRM; Rost, 1990) was performed to detect potential latent dimensions. A mixture Rasch model integrates the Rasch and latent class models, with item parameters estimated for each latent class and ability distributions obtained within latent classes. The major advantage of fitting an MRM to the NAEP science data is that it identifies latent subpopulations and the distinct characteristics of the subpopulations can be used to explain the existence of multidimensionality in the data at large. Importantly, the substantive story that accords with this model may differ in instructionally or pedagogically meaningful ways from those associated with standard within- or between-item multidimensional models.

In this dissertation, data from the NAEP science assessment were analyzed by three types of measurement models. Each measurement model sheds light on a particular aspect of student proficiencies, addresses certain inferences for a particular purpose, and delivers a significant story about the examinees and their learning of science. Each model highlights certain patterns at the expense of hiding other

potentially interesting patterns that reside in the data. Model fit indices, including the Akaike's (1973) information criterion (AIC), the Consistent Akaike's information criterion (CAIC; Bozdogan, 1987), and the Bayesian information criterion (BIC; Schwarz, 1978), were used to compare the posited models and, in turn, evaluate the coherence of the data with the substantive theories underlying the use of the models.

Analyzing the same assessment data through three different types of models is meant to illustrate the interplay among substantive theories, psychometric models, and patterns in the data. Specifically, I am interested in knowing how different theories about science learning motivate the use of different statistical models, what specifics of the assessment data are highlighted by the models, how the models compare with one another in terms of substantive meaningfulness and statistical fit, and how the three aspects of assessment work together to bring about targeted inferences.

In addition to answering the specific research questions, this dissertation serves to address the following meta-questions that are considered to be of primary importance to assessment applications:

1. In light of the interplay among the three aspects of an assessment, how does a measurement model integrate with the other two aspects of an assessment? Especially, how does it connect with the substantive theories of the subject domain being assessed, and how does it represent and model the assessment data?
2. How do alternative measurement models express different conceptions about knowing and learning in the subject domain of interest? How do

they highlight different patterns in the data and model them in different ways? What are the considerations in choosing an appropriate measurement model?

3. How should the models be discussed and compared in terms of the way they model students' performance data, and the types of inferences they support regarding students' knowledge and learning?
4. How does the interplay among substantive theories, measurement models, and patterns in the data inspire efforts in task design? Specifically, in the cycle of assessment development, when does the phase of modeling fitting, model interpretation, and model evaluation take place and how does it inform practices in other phases?

The major contribution of the dissertation is that it elaborates on the interplay among substantive theories, patterns in the data, and measurement models in an assessment, and illustrates this idea in the analysis of a complex assessment. The often-mentioned issue of multidimensionality provides the backdrop for the discussion. Data obtained from the NAEP science assessment, which is designed to be multidimensional, are analyzed by three different multidimensional models, each of which accords with a different conception about knowing and doing in science and addresses inferences targeted at a different aspect of student proficiencies. Three different rationales for the existence of multidimensionality are clearly articulated and supported with data analyses and model evaluation. Based on the results of the analysis, three different stories are told about how different patterns of achievement

vary across subject areas and subpopulations of examinees and why. Integrating ideas from assessment arguments, substantive perspectives on knowing and learning in science, and statistical modeling, this dissertation represents an effort to orchestrate model fitting, model interpretation, and model evaluation within the conceptual framework of an assessment argument.

Chapter 2: Literature Review

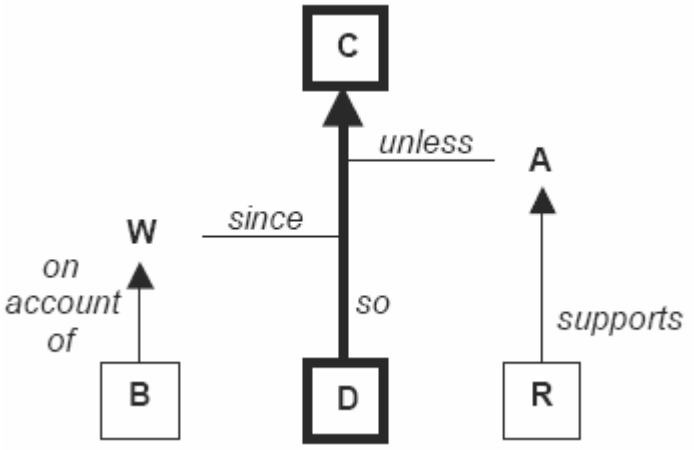
The theme of this dissertation is the interplay among substantive arguments, statistical modeling, and data patterns in an assessment. This chapter starts by reviewing the structure of an assessment argument, highlighting the role of probability-based measurement models. The theoretical background of the interplay among the substantive, statistical, and operational aspects of an assessment, also known as the assessment triangle (National Research Council, 2001) is provided. What follows is a review of the measurement models that are used for data analysis, along with the fit indices by which the models are evaluated. The last section of this chapter is a discussion of the perspective on knowing and learning in science as reflected in the NAEP science assessment and the rationale for the design of the assessment.

2.1 Assessment argument

Every assessment is a special case of evidentiary argument (Mislevy, 1994). The line of reasoning starts from observations of students' performances in a handful of task situations to inferences about their knowledge or proficiency in more broadly construed domains. Although the specifics may vary from one assessment to another, the structure that organizes the specifics into a coherent argument is common across all assessments. The structure of educational assessments can be understood in terms of concepts and representational forms introduced by Toulmin (1958). In Toulmin's terms, an argument is reasoning from particular *data* to particular *claims*. Data are things that we observe and claims are propositions that we want to support

with data. The inference from particular data to a particular claim is justified by a *warrant*, which is in turn supported by *backing*. The backing for a warrant is grounded upon substantive theories and accumulated experience. In any particular case, the inference from data to a claim is qualified by *alternative explanations*, which are supported by *rebuttal evidence*. The structure of a simple argument is outlined in Figure 1.

Figure 1: Toulmin's structure for arguments.

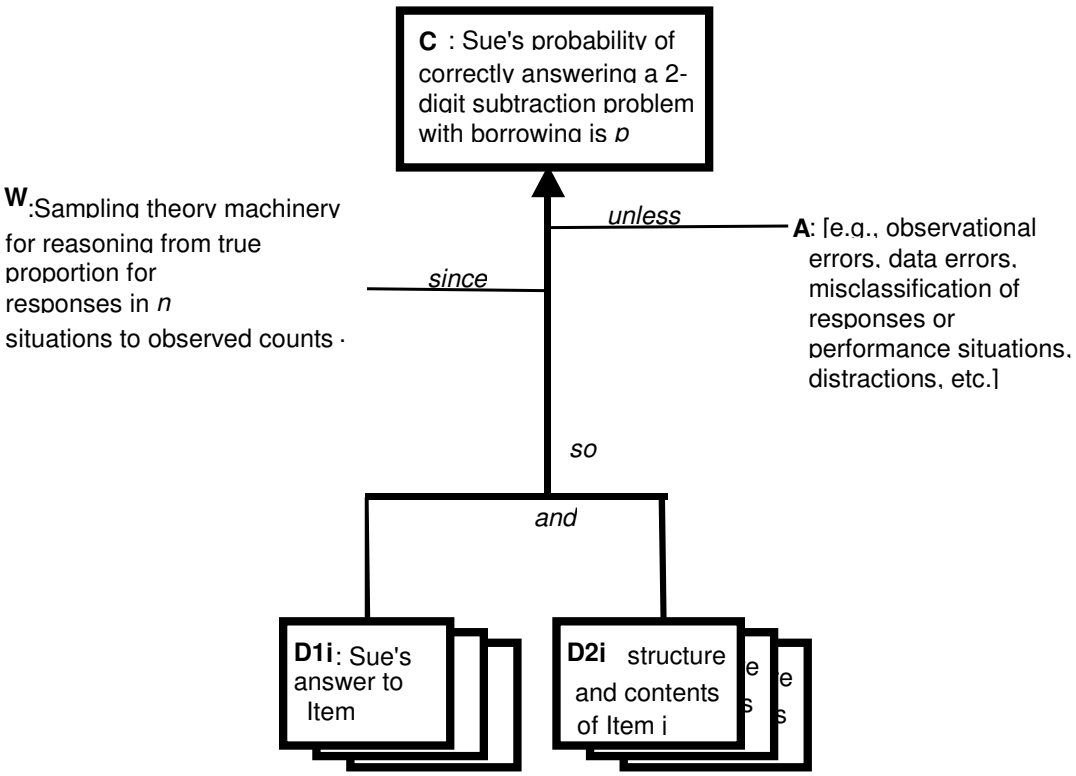


Reasoning flows from *data* (D) to *claim* (C) by justification of a *warrant* (W), which in turn is supported by *backing* (B). The inference may need to be qualified by *alternative explanations* (A), which may have *rebuttal evidence* (R) to support them.

Educational assessments are more complex than Figure 1. An assessment often consists of many claims and data elements, involves multiple chains of reasoning, and contains interweaving dependencies among claims and pieces of data (Mislevy, 2003). Figure 2 displays the structure of an assessment argument that leads from observing Sue's responses to multiple two-digit subtraction items with borrowing to inferences about her ability of solving similar problems. Observations of Sue's performances in

multiple targeted tasks are held as evidence to make an inference about her likely performances in the domain of such tasks. The warrant that justifies the connection between the data and the claim encompasses definitions of the targeted task situations, response classifications, and, most importantly, a probability-based inference model. Reasoning through the model establishes the relationship between observed proportions of correct responses and true proportions in terms of probability.

Figure 2: Elaborated Toulmin's diagram



Toulmin's diagram delineates the basic structure of an assessment, but the substance of every element in the assessment structure and the rationale that orchestrates them as a coherent argument are defined by substantive theories and psychological perspectives on knowledge and learning (Mislevy, 2003). Substantive theories deal with a specific domain of inquiry and are expressed in narrative forms such as categories and properties. They provide the substance for an assessment

argument. Psychological perspectives are of particular significance because they determine the nature of claims that we want to say about students, types of data that we want to obtain to support the claims, and the rationale that justifies the connection between data and claims. Different perspectives emphasize different aspects of knowing and learning and have different implications for what should be highlighted in an assessment and how it should be implemented. Interested readers may refer to Greeno, Collins, and Resnick (1996) and National Research Council (2001) for a thorough discussion of the implications of the different psychological perspectives for assessment practices.

Assessments motivated by different psychological perspectives may appear different on the surface, but a closer look at the arguments underlying the assessments reveals a deeper level of invariability. Every assessment has the same narrative structure, which is fleshed out by substantive theories and psychological perspectives. The narrative structure emphasizes the flow of reasoning from data to claims through the justification of a warrant and qualification of alternative explanations. Domain knowledge and perspectives about learning make explicit what types of competencies are to be determined about students, what to look for in what they say, do, or make, and how it constitutes evidence about what they know and can do. Filling in the general narrative structure with domain-specific substance enables an assessment to tell stories with regard to how students interact with the test tasks, what aspects of test performance bear evidence about student proficiencies, and how students at different levels of proficiency exhibit different patterns of behavior.

A characteristic of educational assessments that makes them different from regular evidentiary arguments is the use of measurement models as one aspect of a warrant. The framework of an assessment argument consists of a narrative structure overlaid by a probability-based measurement model (Mislevy & Huang, 2006). The narrative structure makes explicit the elements of an assessment and their relationships, and connects them into a coherent argument. Measurement models come into the picture at another level. The substantive claims about students are expressed in terms of student model variables and probability distributions, values for which are inferred through a probability-based measurement model from values of observable variables, which are extracted from students' performances in the test, such as item responses. The measurement model quantifies the relationships between the two sets of variables in the form of conditional probability distribution of observable variables given latent student model variables and thus supports probability-based reasoning from observations of student performance to targeted inferences. The role that a measurement model plays in an assessment will be further discussed in the subsequent sections.

2.2 Assessment triangle

Every assessment, regardless of its purpose or the context in which it is used, is based on a triad of interconnected elements, namely, cognition, observation, and interpretation. This framework, referred to as the assessment triangle (National Research Council, 2001), embodies the principle of evidentiary reasoning and can be used to analyze existing assessments and design new ones.

The cognition vertex of the triangle refers to the theory or set of beliefs about how people learn, what they know, and what should be assessed in a subject domain. The theory is often derived from educational research and cognitive studies of how people acquire, represent, and use knowledge and develop expertise in a particular domain. The word "cognition" does not imply that the theory must come from the cognitive perspective. Rather, it should be consistent with a psychological perspective, as appropriate for the purpose of the assessment, and targeted at a level of detail sufficient to get the job of assessment done (Ibid). An effective assessment often starts from a clearly conceptualized cognitive model centered around a well-defined theoretical construct, which is considered most important to assess based on substantive theories and through the lens of a particular psychological theory.

Observation refers to the collection of tasks and observations used to elicit demonstrations of important knowledge and skills from students. This aspect of assessment is essentially a set of schemas for the design of tasks that elicit illuminating responses from students. Tasks should be carefully designed to provide evidence that is linked to the cognitive model and lends support to the theoretical construct measured in the assessment (Ibid).

The interpretation corner of the triangle refers to all the methods and tools used to reason from fallible observations to inferences about students. Observations of students' performances in a set of tasks are synthesized into inferences about students' knowledge, skills, or other attributes through some interpretational framework, which consists mainly of scoring rubrics and a probability-based measurement model. The

scoring rubrics are the rules for extracting salient aspects from students' performances and expressing them as values of observable variables, which are used to update beliefs about students' knowledge, skill, and abilities through the machinery of a measurement model.

Each of the three elements of the triangle should be connected to the other two in a meaningful way in order to produce an effective assessment and generate sound inferences (Ibid). The nature of cognition, the kinds of observation, and the details of interpretation may differ in their particulars, but the challenge of assessment design is the same for every task at hand. Essentially, assessment design is an iterative process (Wilson, 2005). An entire cycle of the process includes formulating a cognitive model about the type of knowledge and skills to be measured in an assessment, creating tasks that will address the targeted knowledge and skills, trying out the tasks on samples of students and observing their performances, analyzing the performance data through the use of a statistical model, and interpreting the results within the framework of the substantive model. It is almost certainly necessary to repeat the cycle one or more times whenever mismatch or inconsistency is identified. If that happens, the cognitive model may need to be refined, data recollected, and measurement design re-contemplated. Each iteration represents an effort of strengthening the linkage among the elements of an assessment and enhancing the cohesion and effectiveness of the underlying argument.

2.3 Role of measurement models

As mentioned above, educational assessments differ from regular evidentiary

arguments in that they use measurement models as one aspect of a warrant. Measurement models are employed to quantify the relationships between students' proficiencies that we want to know about and aspects of their performances that bear evidence of the proficiencies. Student characteristics such as their knowledge, skill, and proficiencies are indexed by parameters of the student model, and important aspects of student behavior that evidence the measured proficiencies are captured by observable variables. Observable variables and student model variables are linked by probability-based functions, and conditional independence is usually assumed for observable variables given student model variables. Through the machinery of probability-based reasoning, especially through the application of Bayes theorem (for example, Mislevy, 1994), observations in the assessment setting are rendered into beliefs or conjectures about students' states with respect to the proficiencies of interest.

A measurement model, however, is not about variables or distributions per se. What underlies a measurement model is a substantive model that specifies the measurement theme and connects the variables and distributions to real-world phenomena. Decisions about a measurement model, such as the number and nature of variables to be included, how they are connected to each other, and what form the model takes, are made in accordance with the substantive model. This substantively-determined measurement model combines the collection of evidence from observations into support for summary conjectures through structures of mathematical probability, and delivers the story of the assessment argument in a more

succinct way. In particular, the probability-based reasoning through the structure of the measurement model permits one to synthesize the information from observations into beliefs about students' knowledge, skills, and abilities.

Measurement models are the lenses through which we view patterns in the data. They are not intended to explain every single detail of data. Rather, they are designed to capture the most important patterns in the data (Mislevy, Wilson, Ercikan, & Chudowsky, 2002). Variables and distributions are the integral components that build up the lenses, through which stories can be told with regard to how students interact with the test tasks, what aspects of test performance bear evidence about student proficiencies, and how students at different levels of proficiency exhibit different patterns of performance (Mislevy & Huang, 2006).

A measurement model, therefore, should be evaluated by two criteria. First, it should be meaningful in the sense that it formalizes the relationships posited in the substantive model. A measurement model disconnected from its substantive context is meaningless and will result in meaningless or even misleading conclusions. The meaningfulness of a measurement model can be evaluated by examining its degree of match with the substantive model. Moreover, a measurement model, however closely it represents a substantive model, should be able to describe data adequately. The goodness of fit of a measurement model is evaluated in terms of the extent to which observed data deviate from predictions of the model. Severe departures alert us to the possibility of model misspecification or failure of the built-in assumptions, such as conditional independence or unidimensionality. Moreover, the ways in

which the observed data differ from the model predictions give us clues about possible causes of misfit. In case of model-data misfit, the substantive model that precedes the measurement model also needs to be reexamined to validate the conceptual underpinnings of the measurement model.

The two criteria by which a measurement model is judged should be used to complement each other to inform model modification exercises. Model construction and modification efforts should be oriented toward stressing the link between a model and its substantive context and improving the fit between the model and the data. A model that is substantively sound may not be able to account for the characteristic features of the data, simply because what the theories predict is not observed in the data. In this sense, checking the degree of model-data fit is also checking the relevance and coherence of the underlying substantive theories. As Embretson (1998) explicitly stated, "The cognitive models are evaluated by the overall fit of a mathematical model" (p. 383). In the other direction, a model that fits the data at hand but lacks a solid theoretical basis needs to be tied back to relevant substantive theories to validate the results obtained from data analysis.

To put it in a broader context, making sense of data collected in an assessment situation through the use of a measurement model is an instance of model-based reasoning (Stewart & Hafner, 1994). Establishing correspondences between elements of data and entities in the structure of a measurement model is the starting point for making explanations and predictions with regard to students' performance data. After the model is formulated, reasoning is carried out through the model, i.e.

the assessment data are analyzed in terms of the variables, distributions, and the quantitative relations specified in the model, and explanations and predictions are made about students, test items, and how their interactions give rise to the observed data. The explanatory and predictive power of the model is evaluated by checking the degree of fit between the data and the model. Any anomalous data that are inconsistent with model predictions suggest directions for model-revising efforts, which are targeted at reconceiving and restructuring the data with a better-fitting model. Given the iterative nature of model-based reasoning, the application of a measurement model to assessment data may involve multiple cycles of model formation, model use, model evaluation, and model revision until the revised model is sufficient.

2.4 Review of measurement models

In this section, three types of measurement models are reviewed and compared in terms of assumptions, properties, and estimation methods. Each type of model represents a distinct approach to accounting for and modeling multidimensionality, and produces inferences targeted at different aspects of students' proficiency in science.

2.4.1 The multidimensional between-item model

In the last two decades, substantial amounts of work has been done on the development and application of multidimensional item response theory (MIRT) models (for example, Reckase, 1997; McDonald, 1999). A multidimensional Rasch-type model, called the multidimensional random coefficients multinomial logit

model (MRCMLM; Adams, Wilson, & Wang, 1997), is particularly useful in practical testing situations due to its flexibility and generalizability.

The MRCMLM is a multidimensional extension of the unidimensional random coefficients multinomial logit model (RCMLM; Adams & Wilson, 1996). It assumes that a set of D traits underlie the persons' responses. The probability of a response in category k of item j is modeled as

$$P(X_{jk}=1;\mathbf{A},\mathbf{B},\boldsymbol{\xi}\,|\,\boldsymbol{\theta})=\frac{\exp(\mathbf{b}_{jk}\boldsymbol{\theta}+\mathbf{a}'_{jk}\boldsymbol{\xi})}{\sum_{k=1}^{K_i}\exp(\mathbf{b}_{jk}\boldsymbol{\theta}+\mathbf{a}'_{jk}\boldsymbol{\xi})} \qquad (1)$$

where $\mathbf{b}_{jk}$ is the vector of scoring functions of response k to item j,

$\boldsymbol{\xi}$ is the item parameter vector,

$\mathbf{a}'_{jk}$ is the design vector, i.e. a linear combination of $\boldsymbol{\xi}$ for response category k of item j.

Suppose an item with four response categories (0, 1, 2, 3) is designed and scored based on the following matrices: (a response in category 0 is denoted by a vector of 0s)

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

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

The response categories are modeled as

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

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

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

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

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

The MRCMLM can be used to model different types of multidimensional tests. A subclass of the model is used for between-item multidimensional tests in which there are several parallel unidimensional subscales and each item measures only one subscale. Another subclass of the model is used for within-item multidimensional tests which are designed to measure several latent dimensions and some or all items are related to more than one dimension. The two types of multidimensionality can be modeled by having appropriate design and scoring matrices in the MRCMLM.

For a between-item multidimensional test in which the items are all dichotomous, the item response function of the MRCMLM is simplified as:

$$P(X_{ij}=1\,|\,\boldsymbol{\theta}_i)=\frac{\exp(\mathbf{r}_j'\boldsymbol{\theta}_i+\beta_j)}{1+\exp(\mathbf{r}_j'\boldsymbol{\theta}_i+\beta_j)} \tag{2}$$

where β_j is the difficulty of item j, $\boldsymbol{\theta}_i$ is the multivariate vector of person i, indicating person i's positions on the multiple latent continuous scales, and

$\mathbf{r}_j=(r_{j1},r_{j2},r_{j3},.....,r_{jm},.....,r_{jM})'$ where

$$r_{jm}=\begin{cases}1, & \text{if item } j \text{ measures to the } m\text{th dimension}\\ 0, & \text{otherwise.}\end{cases},$$

and M is the total number of dimensions assessed in the test.

The MRCMLM can be estimated by marginal maximum likelihood estimation (MML; Bock & Aitkin, 1981) via the application of the EM algorithm (Dempster, Laird, & Rubin, 1977). Consistent estimates of structural item parameters are

obtained. Person parameters can then be estimated by fixing the item parameters at their estimated values. Estimation of the parameters in the MRCMLM is implemented in the ConQuest program (Wu, Adams, & Wilson, 1998).

2.4.2 Item factor analytic model

Item factor analysis is a technique used to investigate the dimensionality of test items. It provides evidence with respect to whether a set of items are indeed measuring a single latent ability or several kinds of abilities, and estimates the strength of relationships between items and latent abilities. Unlike classical factor analysis of continuous measured variables, factor analysis of items cannot be implemented on the observed Pearson product moment correlations of item responses that are either dichotomously scored or scored as discrete values within certain bounds, due to a number of problems (Mislevy, 1986). General practice is to assume that a vector of latent response variables underlie the observed item response variables and to perform common factor analysis on the estimated correlations of the latent response variables.

Suppose a test is composed of J dichotomously scored items, it is assumed that the observed variable X_j is governed by the continuous latent response variable Y_j and the threshold γ_j as

$$X_j = \begin{cases} 1, \text{ if } \mathrm{Y_j} \geq \gamma_j \\ 0, \text{otherwise.} \end{cases}$$

The unobserved continuous variable Y_j is then modeled as a linear function of M ($M < J$) latent factors $\boldsymbol{\theta}$ plus its own unique factor v_j as

$$Y_j = \lambda_{1j}\theta_1 + \lambda_{2j}\theta_2 + \ldots + \lambda_{mj}\theta_m + \ldots + \lambda_{Mj}\theta_M + v_j. \quad (3)$$

The latent factors $\boldsymbol{\theta}$ are assumed to be distributed as MVN (0, $\boldsymbol{\Phi}$), and the unique factors or residuals v_j are distributed as MVN (0, $\boldsymbol{\Psi}$), where $\boldsymbol{\Psi}$ is a diagonal matrix with positive diagonal values. Assuming that the latent factors and residuals are not correlated, the variables Y_j are distributed as MVN (0, $\boldsymbol{\Sigma}$). To eliminate the problem of indeterminacy introduced by the unobserved nature of both Y_j and $\boldsymbol{\theta}$, specifications are made that $\boldsymbol{\Phi} = \mathbf{I}$ where $\mathbf{I}$ is an identity matrix of order *M* and $\Sigma_{jj} = 1$ for each *j*. It follows that $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ and $\boldsymbol{\Psi} = \mathbf{I}$ –diag ($\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$), where $\boldsymbol{\Lambda}$ is the matrix of factor loadings.

There are a number of methods of estimating the parameters in an item factor model. A traditional approach is to use information in the two-by-two contingency tables of joint frequencies to obtain tetrachoric correlations, and then obtain unweighted least squares, weighted least squares, or maximum likelihood solutions of the parameters. Christoffersson (1975) and Muthén (1978) proposed estimation approaches that uses three-, and four-way margins of the raw data table, that is, joint frequencies for items taken three and four at a time, to obtain estimates of parameters. A full information approach (Bock, Gibbons, & Muraki, 1988), as its name suggests, uses all the available information in the data matrix, estimates parameters through the marginal maximum likelihood (MML) method. These methods have been implemented in computer programs such as LISREL (Jöreskog & Sörbom, 1996), Mplus (Muthén & Muthén, 1998), and TESTFACT (Wilson, Wood, & Gibbons, 1991). In this dissertation, the full information approach was used via the implementation of TESTFACT.

Researchers (e.g., McDonald, 1999) recognize that item factor analytic models and multidimensional IRT models have many similarities in methodology. Moreover, the formal equivalence of the two types of models has been established in the literature (e.g., Bock & Aitkin, 1981; Takane & De Leeuw, 1987). Following the notation used in the preceding discussion, the probability of a correct response to item *j* as a function of $\boldsymbol{\theta}$ is given by a normal ogive model

$$P(x_j = 1 \mid \boldsymbol{\theta}) = \Phi\left[\frac{\lambda_{1j}\theta_1 + \lambda_{2j}\theta_2 + ...\lambda_{kj}\theta_k + ... + \lambda_{Mj}\theta_M + v_j}{\sigma_j}\right], \quad (4)$$

where Φ is the notation for a normal ogive model. An alternative expression is

$$P(x_j = 1 \mid \boldsymbol{\theta}) = \Phi\left[a_{1j}\theta_1 + a_{2j}\theta_2 + ... + a_{kj}\theta_k + ... + a_{Mj}\theta_M + d_j\right], \quad (5)$$

where $a_{kj} = \frac{\lambda_{kj}}{\sigma_j}$ and $d_j = \frac{v_j}{\sigma_j}$. Given the well-known relationship between the logistic distribution function and the cumulative standard normal distribution function, the normal ogive model can be approximated by the logistic model

$$P(x_j = 1 \mid \boldsymbol{\theta}) = \frac{\exp\left(\sum_{k=1}^{m} 1.7a_{kj}\theta_k + d_j\right)}{1 + \exp\left(\sum_{k=1}^{m} 1.7a_{kj}\theta_k + d_j\right)}, \quad (6)$$

which is the form of the multidimensional linear logistic item characteristic function first presented by McKinley and Reckase (1982). Therefore, fitting a common factor model to a pool of item responses is equivalent to fitting a compensatory MIRT model, and the item parameters estimated in the factor model can be translated into their MIRT analogs.

2.4.3 Mixture of item response theory models

Traditional item response theory (IRT) describes the performance of all

examinees by using a single model. It provides estimates of students' overall propensities toward correct responses, but fails to give a detailed account about the processes or strategies by which students give correct responses to the items. Latent class analysis (Lazarsfeld & Henry, 1968; Dayton, 1998), on the other hand, does not assign proficiency estimates to individual examinees, but categorizes examinees into discrete latent classes, each of which is identified with a unique response pattern. The response patterns signal examinees' cognitive structures of understanding or developmental stages with regard to certain proficiency. Capabilities and limitations of the two types of models motivate the development of other modeling techniques. Mixtures of IRT models result from the integration of IRT models and latent class models. Within a mixed model framework, the quantitative differences between persons are accounted for by the latent continuous variable in the IRT model within each component of the mixture, and the qualitative differences between persons are explained by the latent categorical variable, whose categories correspond to the components of the mixture (Rost, 1990). Mixture models can be continuous, too, but in this study, the focus is on discrete mixture models in which there are a finite number of components.

One of the earliest attempts to combine IRT and latent class models was made by Yamamoto (1989). He introduced a hybrid model which assumes that the population is a mixture of a group of examinees who respond to items in accordance with an IRT model and a group of examinees whose response patterns cannot be explained by the IRT model but are associated with the latent classes they belong to. In the IRT group,

the probability of an examinee giving a correct response to an item is a function of his (or her) latent ability and the difficulty of that item. Responses of examinees in the latent class group follow certain patterns which reflect their class membership. Each class is characterized by a specific response pattern (called an idealized response pattern), which often results from a unique understanding or misunderstanding of the content being measured. An extended hybrid model (Yamamoto & Everson, 1995) was applied to detecting test speededness and strategy switching from systematic responding to random guessing.

Rost (1990) proposed a mixed Rasch model in which it is assumed that the population of examinees is composed of two or more latent groups. The responses of all the examinees within each latent group are modeled by a standard Rasch model. The item difficulty parameter for each item is assumed to vary across groups. Therefore, the probability of a correct response by examinee *i* from class *g* to item *j* is given by:

$$\mathrm{P}(X_{gij}=1 \mid \theta_{gi}, \beta_{gj}) = \frac{\exp(\theta_{gi}-\beta_{gj})}{1+\exp(\theta_{gi}-\beta_{gj})}. \tag{7}$$

The probability of a correct response by an examinee chosen at random from the population is given by:

$$\mathrm{P}(X_{ij}=1 \mid \boldsymbol{\theta}_i, \beta_{gj}, \pi) = \sum_{g=1}^{G} \pi_g \frac{\exp(\theta_{gi}-\beta_{gj})}{1+\exp(\theta_{gi}-\beta_{gj})} \tag{8}$$

where *G* denotes the number of latent groups, and π_g is the proportion of group *g* in the population, or the mixing parameter as Rost called it. Constraints are set on π_g that $0 < \pi_g < 1$ and $\sum_{g=1}^{G} \pi_g = 1$.

Conditional maximum likelihood estimates of item difficulties within each class and proportions of latent classes can be obtained through the use of an EM algorithm, as described by Rost (1990). Person proficiency for each person within each class can be estimated by using the empirical Bayes inference method. Rost's mixture model does not make any assumption about item parameters in the latent classes or class sizes, which makes it particularly useful in the context of exploratory analysis, where no strong theory exists *a priori* about the nature of the differences between the IRT models representing the distinct latent groups.

Mixture modeling can also provide a framework for testing theories about cognitive processes in a specified content domain. Mislevy and Verhelst (1990) used a mixed linear logistic test model (LLTM; Fischer, 1973) to model strategy use in solving spatial tasks. The model assumes that there are two latent classes of examinees, each attempting the items with a distinct strategy. The probability of item response is given by the Rasch model. Item difficulty is further modeled as a linear function of more basic parameters that reflect the effects of salient characteristics of the item as relevant under the strategy being used.

Unlike Rost's model, Mislevy and Verhelst's model requires substantial knowledge about the content domain prior to data modeling, such as the finite number of strategies being used, the salient item features relevant to each strategy, and the extent to which each characteristic is manifest in each item. Therefore, it is more often used in the context of confirmatory analysis, such as testing hypotheses.

Mixture models retain the advantages of both IRT and latent class models. The

quantitative characterization enabled by the IRT component makes mixture modeling widely applicable in educational testing practices, and the qualitative differentiation effected by the latent class component makes it particularly useful in addressing thorny issues that cannot be solved within the traditional IRT framework. As one of its applications, mixture models are used in detecting multidimensionality and, more importantly, understanding what causes it. Allowing item parameters to vary across latent classes sets the stage for modeling the different mode of interaction between items and examinees within each latent class. Different patterns of item parameters across latent classes are evidence that items that were designed to measure a single trait actually elicit different abilities from different types of examinees. That is exactly what results in multidimensionality. Mixture models have also been used in detecting differential item functioning (Cohen & Bolt, 2005), strategy use (Mislevy & Verhelst, 1990), and test speededness (Yamamoto & Everson, 1995; Bolt, Cohen, & Wollack, 2002).

2.5 Fit indices

In this dissertation, three types of measurement models are compared in terms of goodness of fit as well as conceptual significance. In this section, a number of fit indices that are used for model comparison are reviewed, and research findings regarding their performance are summarized.

The likelihood ratio chi-square difference statistic (G^2_{diff}) is used in many applications where the relative fit of a set of nested models is compared. It is usually presumed to be asymptotically distributed as a chi-square distribution, with its degrees

of freedom equal to the difference of the degrees of freedom between the two models. However, this presumption may not be valid in some cases. For small sample sizes, the chi-square ratio difference statistic is not closely approximated by a chi-square distribution. Besides, the use of this statistic in some applications, such as comparisons of latent class models or mixture models, may result in violations of regularity conditions (Bishop, Fienberg, & Holland, 1975), which are required for the asymptotic results to hold. Most importantly, it is only appropriate for comparisons among nested models. The three types of models being studied in this dissertation are non-nested models, and, therefore, cannot be compared with this test.

Akaike' Information Criterion (AIC) (Akaike, 1973) has been applied in a variety of model comparison and selection problems. When there are several contending models and the parameters within the models are estimated by the method of maximum likelihood, AIC is computed for each model by using the following formula:

$$AIC = -2\log(L) + 2p\,, \tag{9}$$

where L denotes the likelihood of the sample based on the maximum likelihood estimates of the model parameters, and p refers to the number of nonredundant parameters estimated in the model.

AIC is criticized in the literature (for example, Bhansali & Downham, 1977) on the grounds that it is not asymptotically consistent since sample size is not directly involved in its calculation. Without violating Akaike's main principles, Bozdogan (1987) made two analytical extensions to AIC, and one of them results in a selection

criterion called CAIC, which is denoted as follows:

$$CAIC = -2\log(L) + 2p(\log(N) + 1), \tag{10}$$

where N denotes the sample size. CAIC is asymptotically consistent, and, as compared with AIC, penalizes complex models more severely.

The Bayesian Information Criterion (BIC) (Schwarz, 1978) is of the same general form as AIC. It is defined as follows:

$$BIC = -2\log(L) + p\log(N). \tag{11}$$

It differs from AIC only in the second term which depends on the sample size. Obviously, as N increases, BIC favors simpler models more strongly than AIC.

AIC, CAIC, and BIC are generally used to compare non-nested models. They differ in terms of their penalties for overparameterization. Generally speaking, for realistic sample sizes, BIC and CAIC tend to select simpler models than those chosen by AIC. This belief is strengthened by the results in Lin and Dayton (1997).

As mentioned above, AIC, CAIC, and BIC are calculated using the maximum likelihood estimates of the model parameters. When the model parameters are estimated via methods other than maximum likelihood estimation, modified versions of these criteria are considered appropriate. Congdon (2003) suggested calculating AIC and BIC using the posterior means of the parameters in Bayesian modeling when the parameters are estimated via Markov Chain Monte Carlo (MCMC) sampling methods. The AIC and BIC described by Congdon were studied along with other model selection indices for mixture IRT models in Li, Cohen, Kim, and Cho (2006), and the results suggested that BIC performed the best in terms of correctness and

consistency.

2.6 NAEP science assessment

The design of science assessments is a fairly broad area. In this dissertation, I am focused on the rationale for assessing science achievement in NAEP and the design choices that the assessment developers have made to implement the rationale.

The NAEP science assessment is a nationally representative and continuing assessment of what America's students know and can do in the subject area of science. Its primary goals are to detect and report the status of students' science achievement and track changes over time. It provides comprehensive, dependable national achievement data that help educators, legislators, and others reflect on the current practices in science education and make appropriate adjustments to increase the science literacy of students in the United States (National Center for Education Statistics, 1999).

Given the purpose of the NAEP science assessment, it should be envisioned in the large context of the national science education system. The National Science Education Standards (National Research Council, 1996) represent "a broad consensus about the elements of science education needed to permit all students to achieve excellence" (Ibid). Specifically, the standards for science content, which prescribe what students should know, understand, and be able to do in natural sciences, are of particular guidance to the design of NAEP science assessment.

The categories of the content standards include the subject matter of science associated with the divisions of the domain of science, as well as unifying principles,

concepts, and processes that transcend disciplinary boundaries. Science subject matter focuses on "the science facts, concepts, principles, theories, and models that are important for all students to know, understand, and use" (Ibid), while the unifying concepts and processes standard "describes some of the integrative schemes that can bring together students' many experiences in science education across grades K-12" (Ibid). Obviously, the standards not only emphasize the need to examine the extent and organization of students' knowledge, but also stress the need to "probe for students' understanding, reasoning, and the utilization of knowledge" (Ibid).

The development of the NAEP Science Assessment Framework (National Assessment Governing Board, 2000) was guided by the basic principles and perspectives of the National Science Education Standards. The Framework was structured as a matrix, having fields of science and knowing and doing science as its two major dimensions. The fields of science are, namely, earth, physical, and life sciences. Knowing and doing science includes conceptual understanding, scientific investigation, and practical reasoning. Each assessment task can be classified into one subcategory in each of the two dimensions. Besides, two other categories, namely nature of science and themes are specified in the framework that pervade science education but only pertain to a limited number of items in the assessment. Items that belong to these two categories are developed to measure knowledge of content within a specific field of science and an area of knowing and doing science, in addition to addressing knowledge of either of the two categories. Nature of science includes "the history of science and technology, the habits of mind that characterize

these fields, and methods of inquiry and problem solving" (Ibid). Themes represent big ideas or key organizing concepts that enable students to better understand natural phenomena, such as systems, models, and patterns of change.

The two major dimensions of the Framework reflect the idea that learning science involves learning the organized factual knowledge that is unique to each domain, as well as the essentials of learning natural sciences, such as the general process of scientific investigation, way of reasoning, and reliance on technology. The development of science literacy can be characterized as occurring along two dimensions: the content and cognitive dimensions. While students learn science facts, concepts, and theories in their everyday study, they also gain an acquaintance with conceptual and procedural schemes that enable them to understand the natural world better. Therefore, assessing science achievement involves assessing the science content, which is domain-specific, and, more importantly, assessing general abilities of understanding, doing, and using science (National Assessment Governing Board, 2000) that are believed to cross the various content areas.

Chapter 3: Cases of Convergence in Model and Narrative Relationships

In this chapter, the interplay between measurement models and narrative stories is further discussed with an emphasis on the homology between the two. The interrelationships among the types of models being studied in this dissertation are explored. Specifically, I am focusing on the conditions under which the models converge or contrast and the similarities or differences between the narrative stories associated with the models under those conditions.

3.1 Measurement models and narrative stories as two representations

Measurement models and narrative stories are two kinds of representations (Mislevy, 2006; Greeno, 1983) of the domains that assessment projects tap into. They characterize the real-world situation with objects, relationships, and properties that are not necessarily explicit in the situation. They provide a framework in which reasoning of the situation can proceed: the process of mapping between the problem situation and representations leads to understanding, explanations, and predictions of the situation. The two representations are the same in the sense that they involve the same set of conceptual entities (Greeno, 1983) when representing a problem situation. They differ in terms of the perspective from which the problem situation is characterized and the form in which the entities and their relations are represented.

Understanding the homology between narrative stories and measurement models facilitates understanding of the problem domain that they both represent. A measurement model is a mathematical abstraction of the key aspects, patterns, and

relationships that exist in the assessment situation. With each measurement model is associated a narrative space. Connecting the formal entities, including variables and distributions, in a measurement model with people, events, and contexts in the real world spells out all the narrative stories that can be told about the assessment itself and those who are assessed. The variables and their relationships defined in a model can be quite flexible, and, accordingly, the stories inferred from different models are of different versions.

In this dissertation, three types of measurement models are fit to the same assessment data. A look on the surface tells that these models have different structures and involve different variables. However, under certain conditions they become equivalent. The formal relations between these models are reflected in the similarities and differences between the narrative stories derived from the models. Typically, the stories told from the models are of different characteristics, but under particular conditions, they become the same. In relation to the two criteria of evaluating measurement models that were discussed in the previous chapter, i.e., conceptual meaningfulness and statistical fit, the conditions under which the models diverge can be gauged by tests of fit and it is with significant statistical results that the meaningfulness of different versions of narrative stories can be justified.

The questions addressed in this chapter then, are these:

- Under what conditions on model parameters do the models under consideration become mathematically equivalent?
- Are there corresponding equivalences of narratives in those

circumstances?

3.2 Unidimensional models versus multidimensional models

3.2.1 Unidimensionality and essential unidimensionality

Unidimensionality is an assumption required of many item response theory (IRT) models, such as the one-, two-, and three-parameter logistic models. Two definitions of unidimensionality, namely strict unidimensionality and essential unidimensionality, are distinguished in the literature (Stout, 1990; Junker, 1993). Strict unidimensionality means that the probabilities of correct responses to test items are strictly a function of only one latent variable, in addition to variables that represent item characteristics. The general form of a strictly unidimensional IRT model is expressed as

$$P\left(X_j = x_j\right) = \int P\left(X_j = x_j \mid \theta\right) f(\theta) d\theta . \tag{12}$$

Only a single factor, the unidimensional latent trait θ, fully accounts for an individual's performance in a test. Any IRT model that posits strict unidimensionality satisfies the assumption of local independence, which is written as:

$$P\left(X_1 = x_1, ..., X_j = x_j, ..., X_J = x_J \mid \theta\right) = \prod_{j=1}^{J} P\left(X_j = 1 \mid \theta\right)^{x_j} \left[1 - P\left(X_j = 1 \mid \theta\right)\right]^{1-x_j} . \tag{13}$$

It is shown that local independence implies

$$Cov\left(X_j, X_k \mid \theta\right) = 0 \tag{14}$$

for all pairs of $j, k \in \{1, ..., J\}$ and for all levels of θ (Sijtsma & Molenaar, 2002). The zero pair-wise conditional covariance is a result of local independence and is called weak local independence. In general, local independence holds approximately when weak local independence holds (for example, McDonald & Mok, 1995). In

practice, strict unidimensionality is considered too restrictive to be applicable in real-world test situations, where minor dimensions other than the trait being measured also affect examinees' responses.

Essential unidimensionality, a less stringent definition of unidimensionality, recognizes that every test is inherently multidimensional and that item responses are affected by a dominant latent trait and some non-significant latent factors. It assumes a vector of latent traits $\boldsymbol{\theta} = \{\theta, \theta_1, \theta_2, ...\theta_M\}$ relevant to the test items, where θ represents the dominant latent trait of interest and the rest of the vectors denote minor latent traits that are associated with the test items. An essentially unidimensional IRT model is represented as

$$P\left(X_j = x_j\right) = \int_{-\infty}^{\infty} P\left(X_j = x_j \mid \boldsymbol{\theta} = \theta\right) f(\theta) d\theta . \tag{15}$$

IRT models that postulate essential unidimensionality satisfy the assumption of essential independence. Essential independence is represented as

$$\frac{\sum_{1 \le j < k \le J} \left| Cov\left(X_j, X_k \mid \boldsymbol{\theta} = \theta\right) \right|}{\binom{J}{2}} \to 0 \quad \text{as } N \to \infty \quad \text{(Stout, 1990).} \tag{16}$$

This indicates that after conditioning on the dominant latent trait θ, the residual covariances between items are very small on average.

Essential independence is a weaker assumption of local independence. It focuses on the individual examinee differences that are essential or dominant in influencing test performance rather than all the individual differences that influence test performance. As Stout stated, essential independence holds if any of the following three conditions is satisfied: (1) only a few items depend on the trait(s)

other than the dominant trait; (2) each latent trait other than the dominant trait influences at most a small number of items, and these incidental traits are orthogonal to each other, conditioning on the dominant trait; (3) the magnitude of the dependence of the items on the trait(s) other than the dominant trait is small, even though most of the items may depend on them.

The definition of strict unidimensioanlity is impractical for psychological or educational testing and essential unidimensionality presents an efficient and appropriate approximation to it. Attending to the only dominant latent trait and ignoring other inessential or minor traits is not detrimental for any practical purposes and has no adverse impact on the inferences that we want to make about the examinees.

3.2.2 Essential unidimensionality and multidimensionality

Multidimensional IRT models are extensions of unidimensional IRT models. Instead of maintaining the assumption of (strict or essential) unidimensionality, multidimensional models assume that there is more than one latent dimension that significantly influences examinees' responses to test items. The general form of a multidimensional IRT model is written as

$$P\left(X_j = x_j\right) = \int P\left(X_j = x_j \mid \boldsymbol{\theta}\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{17}$$

Unlike unidimensional models in which the latent trait that underlies the item response is represented as a scalar, multidimensional models represent the latent trait as a vector. Local independence of a multidimensional model implies that

$$Cov\left(X_j, X_k \mid \boldsymbol{\theta}\right) = 0 \tag{18}$$

for all pairs of $j,k \in \{1,...,J\}$ and for all values of $\boldsymbol{\theta}$.

Comparison of the formulas for essential unidimensionality and multidimensionality leads to a conclusion that an essentially unidimensional model is technically a multidimensional model but approaches a unidimensional model in the limit. In other words, it is multidimensional in nature, but predictions for item responses made on the basis of the unidimensional approximation approach the correct multidimensional predictions. Mathematically, this is written as

$$P\left(X_j = x_j\right) = \int P\left(X_j = x_j \mid \theta\right) f(\theta) d\theta \approx \int P\left(X_j = x_j \mid \boldsymbol{\theta}\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (19)$$

Insights about the distinction between essential unidimensionality and multidimensionality can also be gained through comparing their geometric representations. In an essentially unidimensional test, items are represented as vectors in a multidimensional space. They all point to approximately the same direction and thus form a relatively homogeneous cluster. The orientation of the cluster in the latent space reflects the dominant latent trait measured by the test. In a multidimensional test, all the item vectors are plotted in a multidimensional space. The direction of each vector indicates the composite of latent traits it purports to measure. A pattern that suggests multidimensionality is that the item vectors form several distinct item clusters, implying that the test as a whole is multidimensional, but each item cluster can be treated as a unidimensional subtest. Another scenario is that the item vectors cannot be separated into distinct item clusters, implying that the test is multidimensional and that the items measure different combinations of latent traits. These two representations coincide with the two types of multidimensionality,

namely between- and within-item multidimensionality, which will be compared in the following section.

The dimensionality of a test can be evaluated by a variety of fit indices and significance tests (Tate, 2003). As implied by Tate, almost all of the methods for assessing test dimensionality are based on the concept of essential unidimensionality. No matter what assessment method, parametric or nonparametric, is used to examine dimensionality, the basic rationale is as follows: the assumption of local independence is assessed under a hypothesized model, unidimensional or multidimensional, and measures are obtained that indicate the amount of item dependence. If item dependence is stronger than what would be expected by chance, the assumption of local independence is suspected. If strong dependencies exist among items, the assumed model dimensionality would be rejected. It should be noted that in this approach, local independence is defined as pairwise or weak local independence and item dependencies are calculated on the basis of the conditional covariances of item responses for all item pairs at all levels of the latent traits.

Many achievement tests are designed to be unidimensional for scoring and ranking purposes. However, as Ackerman (1994) argued, "unidimensionality should never be assumed but should always be verified" (p. 257). When the conditions of unidimensionality are satisfied for a given test and the specified population, persons' differences in test performance are attributed to their differences with regard to the latent trait measured by the test. Stories about persons' propensities toward correct responses and items' difficulty, discriminating power, and other characteristics can be

told. The assessment of unidimensionality is related to the test in general, and there are item fit and person fit indices, which indicate the goodness of fit of individual items and persons. These fit indices can detect items or persons that exhibit unique response patterns deviant from those expected. Existence of such items or persons does not necessarily undermine the viability of unidimensionality, but should be carefully investigated and properly interpreted.

When the assumption of unidimensionality is not viable for a test, a multidimensional model should be considered. Multidimensional IRT models are applied for either exploratory or confirmatory uses. Exploratory MIRT models are used when the test developer has no strong theory about the structure of the test to be analyzed. As stated in the previous chapter, a MIRT model used in an exploratory mode is equivalent to a common factor analytic model. Mathematically, a common factor model is represented as Equation 2

$$Y_j = \lambda_{1j}\theta_1 + \lambda_{2j}\theta_2 + ... + \lambda_{mj}\theta_m + ... + \lambda_{Mj}\theta_M + v_j ,$$

where Y_j is the continuous latent response variable underlying the observable dichotomous variable X_j, λ_{mj} is the factor loading of Y_j on the latent factor θ_m, and v_j is the unique factor. It approaches the unidimensional model if either (i) the factor loadings approach one another (i.e., same loadings across items,

or $\lambda_{11} = \lambda_{12} = ...\lambda_{1j} = ... = \lambda_{1J}$, $\lambda_{21} = \lambda_{22} = ...\lambda_{2j} = ... = \lambda_{2J}$, …,

$\lambda_{m1} = \lambda_{m2} = ...\lambda_{mj} = ... = \lambda_{mJ}$,…, $\lambda_{M1} = \lambda_{M2} = ...\lambda_{Mj} = ... = \lambda_{MJ}$ where $j \in \{1,...,J\}$

and $m \in \{1,...,M\}$) or (ii) the factor correlations all approach 1. In the first case, multiple (*M*) latent traits are required but the exact same combination of them is

required for all (J) items. The same combination, the composite of latent traits, can be thought of as a single dimension. The second case means that multiple latent traits are involved, but people are lined up exactly the same way with regard to all latent traits. The common lineup may as well be thought of as the single dimension involved in this set of items, for this group of examinees.

When the test developer has substantial knowledge about the content domains or cognitive abilities assessed in the test and how they are needed in combination for a correct response to each item, a MIRT model can be used to verify (or reject) his (or her) knowledge. A family of MIRT models developed for confirmatory uses is the multidimensional random coefficients multinomial logistic model (MRCMLM), which was described in the previous chapter. The MRCMLM approaches the unidimensional random coefficients multinomial logistic model (RCMLM) when the underlying ability is unidimensional.

The goodness of fit of a hypothesized multidimensional model is evaluated by comparing it to that of a unidimensional model. If the improvement in fit is significant, the multidimensional model is retained. Stories that are consistent with expectations of the model can be told. For example, persons are compared with regard to their knowledge in the content areas or abilities covered in the test. Similarly, items are characterized by their required combinations of knowledge or abilities. If the multidimensional model does not fit significantly better than a unidimensional model, the more parsimonious unidimensional model is retained and explanations of persons and items are made in accordance with the unidimensional

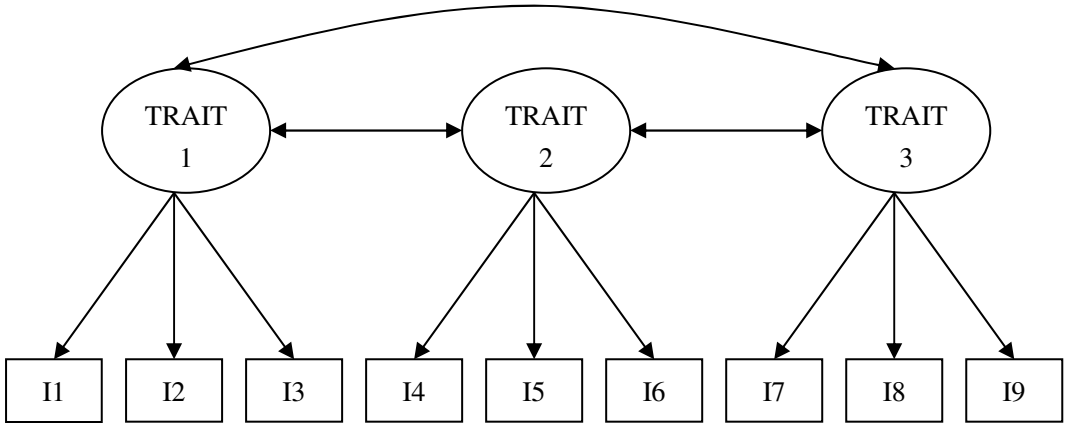
model.

3.3 Between-item versus within-item multidimensionality

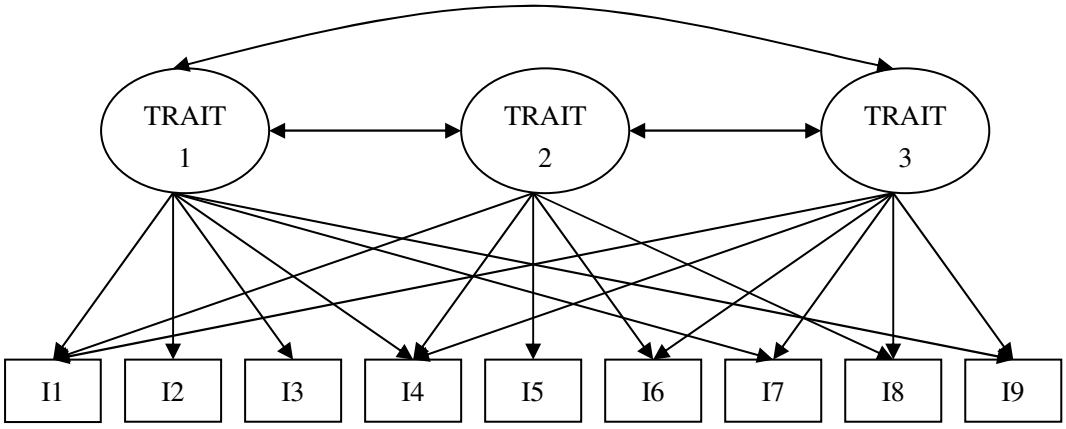
As discussed above, the distinction between between- and within-item multidimensionality can be best understood with their geometric representations. A between-item multidimensional test is represented as being made up of several dimensionally homogeneous item clusters, while a within-item multidimensional test can not be broken into a smaller number of distinct clusters. A mathematics achievement test that contains subtests of arithmetic, algebra, geometry, and measurement is an example of a between-item multidimensional test. Within-item multidimensionality is most easily recognized in a test composed of mathematics word problems, which require both computation and reading skills (Ansley & Forsyth, 1990). To use the concepts in factor analysis, a between-item multidimensional test can also be understood as exhibiting simple structure (Stout et al., 1996), while a within-item multidimensional test as displaying complex structure. Figure 3 illustrates the structural differences between the two kinds of multidimensionality.

Figure 3: Structures of between-item and within-item multidimensionality

Between-item multidimensionality



Within-item multidimensionality



Understanding the number and nature of abilities that determine the response to each item is important in modeling item responses. The structure of a between-item multidimensional test allows for separate measurements on each dimension. However, this approach is criticized because of its failure to use all the available data (Adams, Wilson, & Wang, 1997). In a between-item multidimensional test, the

subscales are distinct but correlated. A joint analysis, which takes advantage of the correlations among the latent dimensions, leads to improved estimates of item and person parameters. Within-item multidimensional tests are often analyzed with compensatory or noncompensatory models, depending on the nature of the interaction between the required latent dimensions. Compensatory models are used more often than noncompensatory models in practical testing situations, partly because they have well-developed estimation algorithms. The most commonly used compensatory models are the two-parameter linear logistic model (McKinley & Reckase, 1982) and its variations. Besides, the MRCMLM is developed to incorporate a wide class of multidimensional models, which can be used in contexts in which either a compensatory or noncompensatory model is deemed desirable.

3.4 Multidimensional models versus mixture models

Multidimensional models and mixtures models represent two different techniques of modeling item responses that cannot be adequately accounted for by a single unidimensional trait. Although different in structure, they are equivalent under certain conditions. Specifically, Rijmen and De Boeck (2005) studied two extensions of the Rasch model, the between-item multidimensional model (Adams, Wilson, & Wang, 1997) and the mixture Rasch model (Rost, 1990), and proved their equivalence in certain circumstances.

In the Rasch model, the marginal probability of a response pattern $\mathbf{y}$ is equal to

$$P(\mathbf{y} \mid \boldsymbol{\beta}) = \int_{\theta} \left\{ \prod_{j=1}^{J} \frac{\exp\left[y_j(\theta + \beta_j)\right]}{1 + \exp(\theta + \beta_j)} \right\} df(\theta), \tag{20}$$

where $f(\theta)$ is the distribution function of the latent variable θ in the population of

examinees. The first extension of the Rasch model is its multidimensional version. Suppose a test consists of *K* subgroups of items and each group can be modeled by the Rasch model. Let $\mathbf{r}_j = (r_{j1}, r_{j2}, ..., r_{jk}, ..., r_{jK})'$ where

$$r_{jk} = \begin{cases} 1, & \text{if item } j \text{ belongs to the } k\text{th group} \\ 0, & \text{otherwise.} \end{cases}.$$

Based on the between-item multidimensional model, the marginal probability of a response pattern $\mathbf{y}$ is equal to

$$P(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{R}) = \int_{\theta} \left\{ \prod_{j=1}^{J} \frac{\exp\left[y_j \left(\mathbf{r}_j' \boldsymbol{\theta} + \beta_j\right)\right]}{1 + \exp\left(\mathbf{r}_j' \boldsymbol{\theta} + \beta_j\right)} \right\} df(\boldsymbol{\theta}) \tag{21}$$

where $\boldsymbol{\theta}$ is a *K*-dimensional vector which represents the latent traits, $\boldsymbol{\beta}$ is the vector of item parameters, and $\mathbf{R}$ is a $J \times K$ matrix with $\mathbf{r}_j$ as is *j*th row.

The second extension of the Rasch model is its mixture version. Suppose the population of examinees consists of *G* latent classes. According to the mixture Rasch model, the marginal probability of a response pattern $\mathbf{y}$ is equal to

$$P(\mathbf{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_G) = \sum_{g=1}^{G} \pi_g \int_{\theta} \left\{ \prod_{j=1}^{J} \frac{\exp\left[y_j \left(\theta + \beta_{jg}\right)\right]}{1 + \exp\left(\theta + \beta_{jg}\right)} \right\} df_g(\theta) \tag{22}$$

where π_g is the proportion of the *g*th class, $\boldsymbol{\beta}_g$ is the vector of item difficulties for the *g*th class, and $f_g(\theta)$ is the distribution function of θ for the *g*th class.

In exploring the formal relation between the two models, Rijmen and De Boeck (2005) rewrote the between-item multidimensional model in Equation 21 as

$$P(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{R}) = \int_{\xi} \left\{ \prod_{j=1}^{J} \frac{\exp\left[y_j \left(\mathbf{r}_j' \mathbf{S} \boldsymbol{\xi} + \beta_j\right)\right]}{1 + \exp\left(\mathbf{r}_j' \mathbf{S} \boldsymbol{\xi} + \beta_j\right)} \right\} dq(\boldsymbol{\xi}) \tag{23}$$

where $\boldsymbol{\xi} = \boldsymbol{T}\boldsymbol{\theta}$, with $\boldsymbol{T}$ a nonsingular transformation matrix and $\mathbf{S} = \boldsymbol{T}^{-1}$, and $q(\boldsymbol{\xi})$ is the distribution function of $\boldsymbol{\xi}$. Equation 23 can be reformulated as

$$P(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{R}) = \int_{\xi_{(k)}} \int_{\xi_k} \left\{ \prod_{j=1}^{J} \frac{\exp\left[y_j \left(\mathbf{r}_j' \mathbf{S} \boldsymbol{\xi} + \beta_j \right) \right]}{1 + \exp\left(\mathbf{r}_j' \mathbf{S} \boldsymbol{\xi} + \beta_j \right)} \right\} dq_k \left(\xi_k \mid \boldsymbol{\xi}_{(k)} \right) dq_{(k)} \left(\boldsymbol{\xi}_{(k)} \right), \tag{24}$$

if we condition the *k*th latent variable ξ_k on the vector of *K*-1 other latent variables $\boldsymbol{\xi}_{(k)}$. If $\boldsymbol{T}$ is chosen such that the inner integral of the function corresponds to the function of the Rasch model, this reformulation of the between-item multidimensional model can be regarded as a continuous version of the mixture Rasch model, where each class has a different set of values for $\boldsymbol{\xi}_{(k)}$.

For a given class in which $\boldsymbol{\xi}_{(k)} = \mathbf{a}_{(k)}$, the probability of observing a response pattern $\mathbf{y}$ is equal to

$$P(\mathbf{y} \mid \boldsymbol{\xi}_{(k)} = \mathbf{a}_{(k)}, \boldsymbol{\beta}, \mathbf{R}) = \int_{\xi_k} \prod_{j=1}^{J} \frac{\exp\left[y_j \left(\sum_{k=1}^{K} r_{jl} \left(s_{lk} \xi_k + \mathbf{s}_l^{(k)} \mathbf{a}_{(k)} \right) + \beta_j \right) \right]}{1 + \exp\left[y_j \left(\sum_{k=1}^{K} r_{jl} \left(s_{lk} \xi_k + \mathbf{s}_l^{(k)} \mathbf{a}_{(k)} \right) + \beta_j \right) \right]} dq_k \left(\xi_k \mid \boldsymbol{\xi}_{(k)} = \mathbf{a}_{(k)} \right) \tag{25}$$

where $\mathbf{s}_l^{(k)}$ is the *l*th ($l = 1,2,...,k,....,K$) row of $\boldsymbol{S}$, without element s_{lk}. Considering the fact that each item belongs to one dimension in the between-item multidimensional model, Equation 25 can be simplified as

$$P(\mathbf{y} \mid \boldsymbol{\xi}_{(k)} = \mathbf{a}_{(k)}, \boldsymbol{\beta}, \mathbf{R}) = \int_{\xi_k} \prod_{j=1}^{J} \frac{\exp\left[y_j \left(s_{l_j k} \xi_k + \mathbf{s}_{l_j}^{(k)} a_{(k)} + \beta_j \right) \right]}{1 + \exp\left[y_j \left(s_{l_j k} \xi_k + \mathbf{s}_{l_j}^{(k)} a_{(k)} + \beta_j \right) \right]} dq_k \left(\xi_k \mid \boldsymbol{\xi}_{(k)} = \mathbf{a}_{(k)} \right) \tag{26}$$

where $s_{l_j k} = s_{1k}$ for items that belong to the first dimension, $s_{l_j k} = s_{2k}$ for items that belong to the second dimension and so on. The inner integral of Equation 26 will be equal to the formulation of the Rasch model as in Equation 19 if $s_{lk} = 1$ for all *l*, or, in other words, the *k*th column of **S** consists of ones only.

As seen in Equation 26, the item parameter for the *j*th item within a latent class is equal to the item parameter β_j of the between-item multidimensional model and "a

term that is common for all the items belonging to the same original dimension *l* and that depends on the value of $\mathbf{a}_{(k)}$ one is conditioning upon, $s_l^{(k)}\mathbf{a}_{(k)}$" (Rijmen & De Boeck, 2005, p. 486), plus a term that is common for all items due to the identification restriction that $E\left(\xi_k \mid \boldsymbol{\xi}_{(k)} = \mathbf{a}_{(k)}\right) = 0$. To put it simply, within a given class, the item parameters are equal to the item parameters of the multidimensional model plus a shift parameter that is specific for the dimension an item belongs to in the multidimensional model. The term $s_l^{(k)}\mathbf{a}_{(k)}$ can be seen as an interaction between the latent class a person belongs to (identified by the values of $\mathbf{a}_{(k)}$) and the group a particular item is associated with (i.e., dimension *l*). A mixture model with this property is similar to a Saltus model (Wilson, 1989; Mislevy & Wilson, 1996), except that it consists of a continuous mixture of classes while in a Saltus model the classes are often discontinuous.

The formal equivalence between the two types of models suggests that a between-item multidimensional model is approximated by a mixture Rasch model in which the item parameters for items associated with the same dimension are equal across classes up to a class-specific shift parameter, as the number of classes approaches infinity. In other words, the effect of being in a particular class upon responding to an item is common for all items of the same dimension. The interaction between person group and item class in the mixture model, as represented by the shift parameter, approximates the algorithm in the multidimensional model that items measure different dimensions and people have different distributions along those dimensions.

In addition to establishing the formal equivalence between the between-item multidimensional model and a continuous mixture Rasch model, Rijmen and De Boeck did a simulation study which suggests that the equivalency relationship holds approximately between a two-dimensional between-item model and a finite mixture Rasch model with only two latent classes.

In another line, Reise and Gomel (1995) compared the mixture Rasch model and the full information item factor analytic model and argued that these two models "represent two conceptually distinct ways of accounting for heterogeneity in an item response matrix" (p. 342). To put it more exactly, the two types of models can be distinguished as addressing "item heterogeneity" or "person heterogeneity". This study is a complement to the study by Rijmen and De Boeck in some sense because it is also an exploration of the relationship between a MIRT model and the mixture Rasch model. Besides, solutions from the exploratory analyses suggested that the personality assessment being analyzed can be well represented by a two-dimensional between-item model or a two-class mixture Rasch model. It should be recognized that the item factor analytic model resembles the between-item multidimensional model and its rotated solution may suggest between-item multidimensionality. It differs from the between-item multidimensional model in some ways such as, the item discrimination parameters are included in an item factor analytic model, the group membership of each item is estimated but not specified a priori, and a probit link is used instead of a logit link.

For a given test, the solutions given by the between-item multidimensional model

and the mixture Rasch model share some similarities due to the isomorphic relation between the two types of models. The distinctions between the dimensions in the between-item multidimensional model often correspond to the characteristic features of the latent classes in the mixture Rasch model. For example, the results from Reise & Gomel (1995) indicated that a 2-factor IRT model and a 2-class mixture Rasch model provided the best representations of the data. The two dimensions were called "agency" and "communication", and the two latent classes identified by the mixture Rasch model turned out to be the "agentic" and "communal" types of people.

The equivalence between the two models is discussed in the context that the true model underlying the test data is known to be the between-item multidimensional model. If the true model is unknown and the two models are fit to the test data, they do not always fit equally well. Generally speaking, the between-item multidimensional model is more parsimonious than the mixture Rasch model. However, as discussed in the previous chapter, choosing between models is not merely a statistical issue. There are circumstances in which one model is preferred over the other (Reise & Gomel, 1995). For example, if items can be reasonably divided into groups a priori, the between-item multidimensional model will be considered more plausible.

On the other hand, the two models emphasize different aspects of the variations observed in a response matrix, and tell different stories about persons and items. The between-item multidimensional model associates items with multiple latent

dimensions and expresses person differences along these dimensions. It emphasizes the differences between items with regard to the particular latent dimension each measures and explains response heterogeneity with multiple dimensions of individual differences. The mixture Rasch model assumes a single latent trait and identifies multiple subpopulations. It focuses on the distinctions among persons with regard to the class each belongs to. Response heterogeneity is accounted for by person differences, which are described both qualitatively, as class membership, and quantitatively, as values on the latent dimension specific to each class.

3.5 Unidimensional models versus mixture models

The general form of a mixture IRT model can be expressed as

$$P\left(X_j = 1 \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \pi\right) = \sum_{g=1}^{G} \pi_g P\left(X_j = 1 \mid \theta_g, \beta_g\right),$$

where π_g is the proportion for the *g*th class, and $P\left(X_j = 1 \mid \theta_g, \beta_g\right)$ is the response function for the *g*th class that accords with a certain unidimensional IRT model. A mixture model is collapsed to a unidimensional model under two circumstances. When one of the components of the mixing parameter π is essentially equal to 1 and the others are essentially equal to 0, the mixture model is essentially a unidimensional model. In that case, there is a mixture in principle, but virtually all examinees are in the same class. Another scenario would be that the item parameters in the different classes all approach one another. In that case, there are different classes of examinees, but all classes share the same response probabilities for the test items.

Unidimensional models and mixture models are comparable in the sense that they both assume unidimensionality. Unidimensional models assume that examinees all

come from the same population and that their differences are quantitative with regard to a single latent dimension. Items are also located on the same scale. A mixture model assumes that the population of examinees consists of multiple qualitatively different classes, and that the class-membership is latent. Within each latent class, unidimensionality still holds. However, the item parameters are class-specific, implying that the particular dimension the test items are measuring varies across classes.

The differences between a unidimensional model and a mixture model make them applicable in different contexts and for different purposes. In a unidimensional model, a single latent trait is assumed to underlie the response process of every item for every examinee, and a common set of item parameters are estimated for all examinees, which implies that all examinees are assumed to perceive and respond to the items in the same manner. This assumption does not hold at all times, especially when the examinees come from distinct subpopulations with different background characteristics, just like those in the NAEP science assessment. In modeling their item responses, the examinees cannot simply be characterized as possessing different quantities of a latent trait. They should also be distinguished in terms of the qualities of latent traits they may have demonstrated in the process of responding to the items. In case the assumption of unidimensionality fails, a mixture model, for example, the mixture Rasch model, can be used in an exploratory mode to investigate the heterogeneity among examinees. Of course, mixture models (for example, Mislevy & Verhelst, 1990) can also be used in confirmatory analysis to support hypotheses

about the causes for heterogeneity among persons in the item-solving process.

Unidimensional models and mixture models tell different stories about persons and items. In the unidimensional Rasch model, each item is equally easy or hard for all the examinees. Examinees are quantified with regard to the single latent construct measured in the test. In the mixture Rasch model, by contrast, one or more items are harder for one class but easier for another. For all examinees, the probability of belonging to each latent class is calculated. Examinees with response patterns of a particular type have a larger likelihood of belonging to one latent class than to another. A mixture model almost always fits better than a unidimensional model because it can account for departures from unidimensionality and detect more refined distinctions about examinees. However, it does not always make more sense than a unidimensional model. It tends to pick up characteristics that are only unique to the particular items that are included in the test and the particular examinees who took the test. Interpretation of the latent classes should be based on the generalizability of these characteristics.

As noted in each section of this chapter, under certain conditions, each two of the three types of models converge in both mathematical forms and narrative stories. In cases of convergence, tests of statistical fit or fit indices cannot tell the differences between the models, and the narrative stories associated with the models are the same.

Chapter 4: Methodology

4.1 Data

Data used in the study were item responses of the 8th graders from the national comparison sample in the NAEP 1996 science assessment. As described in the Data User Guide (National Center for Education Statistics, 1999), the NAEP 1996 science assessment was administered to samples of students in the participating states as well as to a national sample. NAEP employs a stratified cluster sampling scheme in selecting participants. The national comparison sample was a subsample from the full national sample created to allow for valid state and national comparisons. It was representative of all students in Grade 8 enrolled in public and nonpublic schools in the United States in the assessment year.

Each student was administered a booklet containing cognitive and background items. The pool of cognitive science items for Grade 8 was divided into fifteen mutually exclusive blocks. Each block contained both multiple-choice and constructed-response items. The number of items within a block ranged from 6 to 16. Four of the fifteen blocks were hands-on tasks in which students were given a set of equipment and asked to conduct an investigation and answer questions (mostly constructed-response) related to the investigation. Three of the remaining eleven regular paper-and-pencil blocks were theme blocks, which were designed to address the themes in science education, as described in the NAEP science assessment framework.

Based on a complex matrix sampling design (Allen et al., 1999), blocks of items

were assembled into booklets, which were then assigned to students. Each booklet contained three blocks of cognitive items. Each booklet contained a block of hands-on tasks, which were always presented in the last position of the booklet. Theme blocks were placed randomly in student booklets, but not in every booklet. Each booklet contained no more than one theme block. Each non-theme paper-and-pencil block appeared in the first or second position of a booklet the same number of times. The administration of booklets to students followed a spiraling design, which was a systematic way to ensure that each booklet appeared an appropriate number of times in the student sample.

In addition to three cognitive science blocks, each booklet also contained three segments of background items, i.e. a demographic questionnaire, a science background questionnaire, and a motivation questionnaire. Every student received the same background items. The student demographic questionnaire included questions about race, mother's and father's level of education, types of reading materials at home, and school attendance. The science background questionnaire included questions that addressed attitudes and perceptions about science, time spent studying science, and instructional experiences related to science in the classroom. The motivation questionnaire asked students questions such as how they thought about their performances in the test and how important it was for them to do well in the test.

In this dissertation, data analyses were performed with responses to cognitive science items. To circumvent the complexities of matrix sampling in the NAEP

assessment instrument, data analyses could be done on single blocks or combinations of blocks. Analyzing multiple blocks simultaneously would involve having more items but fewer examinees in the response data than analysis of a single block. In this study, data analyses were performed with three sets of response data. Analyses were first done on a single block of items, whose content and framework classifications were released to the public. Analyses could be desirably done on a booklet, which contained three blocks. However, since not a single booklet in the 1996 NAEP science assessment was released for public use, analyses were then done on two other data sets, each containing examinees' responses to two blocks of items.

To simplify data analysis, polytomously scored constructed-response items were dichotomized in a manner that the collapsing of categories resulted in the most balanced dichotomous response frequencies for each item. That is to say, categories were collapsed so that the frequency of correct responses would approximate that of the incorrect responses as much as possible. In this way, the loss of information due to dichotomization was minimized.

In the data files used in the study, missing responses were recoded as "omitted" or "not presented". Specifically, missing responses prior to the last observed response of a block were coded as omissions, while missing responses at the end of a block were coded as not reached. In the model analyses, omitted responses were treated as incorrect responses and not reached items were treated as if they had not been presented to the examinee. The same response data were analyzed by each model in the study.

4.2 Model analyses

In this study, item responses were analyzed by three different types of measurement models and the results were compared within and across types of models. The multidimensional between-item models were used in a confirmatory approach to examine two hypotheses that were implied by the design rationale of the NAEP science assessment. The first hypothesis, which was assumed in the NAEP science score reporting system, stated that the assessment was multidimensional because of the three content areas that were covered. The second hypothesis was that the assessment was multidimensional in terms of the three cognitive proficiencies that the items wee designed to measure. Following the confirmatory analyses, item responses were then subjected to exploratory analysis via two types of models. Exploratory item factor analysis was conducted to investigate the number of latent factors that underlay the item responses, how the items were associated with the factors, and how the factors were to be interpreted. The mixture Rasch models were used to identify multiple latent classes of examinees whose response patterns were qualitatively dissimilar. Quantitative differences among students were scaled within each class.

The three models reflect three different conceptualizations of multidimensionality. The multidimensional between-item model approaches multidimensionality at a global level by evaluating the structure of the test instrument. The factor model and mixture IRT model account for multidimensionality in terms of “item heterogeneity” and “person heterogeneity”, respectively (Reise & Gomel, 1995). The three models,

when used in complement with each other, provide a general as well as a nuanced picture about the test, the items, and the examinees.

4.2.1 The multidimensional between-item model

Application of the multidimensional between-item model is inspired by the design feature of the NAEP science assessment. Each block as a whole covers three fields of science (physical, earth, and life sciences) and three types of cognitive proficiencies (conceptual understanding, practical reasoning, and scientific investigation), but each item within a block is designed to measure knowledge of content of only one field of science and one type of cognitive proficiency. Thus, the structure of the assessment can be accounted for by the multidimensional between-item model and the dimensions can be defined in terms of content areas or science process skills.

As mentioned above, the multidimensional between-item model is a subclass of the multidimensional random coefficients multinomial logit model (MRCMLM). Estimation of the MRCMLM is based on the marginal maximum likelihood (MML) estimation procedure, which is implemented in ConQuest. When estimating the item parameters of the model, the vector-valued person parameter $\boldsymbol{\theta}$ is assumed to follow a multivariate normal distribution. Results provided by ConQuest include estimates of the item parameters, means, variances, covariances and correlations of the latent dimensions, and deviance of the model. ConQuest also provides expected a posterior (EAP) ability estimates and maximum likelihood ability estimates for the person parameters. The overall fit of the multidimensional model can be examined

by testing the difference in the model deviance between a unidimensional model and the multidimensional model, which approximates a chi-square distribution asymptotically with degrees of freedom equal to the number of additional parameters estimated in the multidimensional model. A significant test result indicates that the multidimensional model fits the item response data significantly better than the unidimensional model.

ConQuest also provides fit statistics for individual items, which are residual-based indices similar to the weighted and unweighted fit statistics that were developed by Wright and Masters (1982). Weighted fit statistics are preferred because they are less sensitive to unexpected responses made by persons for whom the item of interest is far too easy or far too difficult. Wu (1997) has shown that these statistics have approximate scaled chi-square distributions and can be transformed to approximate normal deviates (t-values). Following standard guidelines, an item is considered as a misfit item if the absolute value of its associated t-statistic is greater than 2.0. A t-value greater than 4.0 or less than -4.0 indicates serious misfit. However, according to Hambleton & Murray (1983), sample size can significantly impact the detection of misfit items. Based on their simulation study, the number of detected misfit items tends to increase with the increase of sample size, and it seems that sample sizes around 600 to 1000 may give accurate results. When sample size is over 1000, the fit statistics need to be interpreted with caution.

4.2.2 The exploratory item factor analytic model

As mentioned above, the exploratory full-information item factor analysis is

implemented in the TESTFACT program, which uses the marginal maximum likelihood estimation (MML) method on the full item response data matrix to obtain item parameters. TESTFACT first performs a principal factor analysis on the smoothed tetrachoric correlation matrix by using the minimum squared residuals (MINRES) method. Factors are extracted and factor loadings are obtained. In the initial solution generated by MINRES, the factors are orthogonal to each other, and can be subjected to varimax (factors being orthogonal) or promax (factors being oblique) rotation, as indicated in the command. In this study, all the rotations performed on the factor solutions were promax rotations, and the correlations between the factors were estimated. The factor loadings are then converted to intercepts and slopes, which serve as the starting values for the MML procedure. The full-information item factor analysis results in a chi-square statistic for the model fit and parameter estimates for both the factor analytic and the multidimensional item response theory formulations.

Determining the number of factors with the exploratory solution provided by TESTFACT involves examining the latent roots of the tetrachoric correlation matrix, the root mean square residual (RMSR) statistic for the matrix of residuals, chi-square difference statistics, and the number of substantial loadings for the factors (Stone & Yeh, 2006). As suggested by many researchers (for example, Gorsuch, 1983), examination of scree plots is useful for determining the number of factors. RMSR is a statistic that summarizes the differences between the observed correlations and the model-implied correlations, i.e., the matrix of residuals provided in the TESTFACT

output. A value of .05 can be used to indicate an acceptable factor solution (Muthén & Muthén, 2001). The chi-square difference statistics between nested factor models can be tested to determine the number of factors. If adding another factor does not bring about significant improvement in fit, the current factor model should be retained as the most appropriate model. Finally, the factor loadings should be examined to identify the cluster of items that have high loadings on each factor. The magnitude of factor loadings indicates the strength of the relationships between the items and the factors. Factors are interpreted based on the items that are strongly associated with them. Typically, a factor loading is considered substantial if it is greater than .3 (Gorsuch, 1983).

A recent article by Stoel, Galindo-Garre, Dolan, & van den Wittenboer (2006) points out that the boundary conditions of the parameters in the common factor model make the chi-square difference tests no longer appropriate for comparing nested factor models. Other fit statistics such as the information criterion indices may be used instead in comparing this type of models. In this dissertation, I used both the chi-square difference tests and the information criteria in comparing factor models, but the latter ones should be given more attention because they are considered more reliable.

4.2.3 MCMC estimation of the mixture models

Rost's (1990) mixture Rasch model (MRM) is used in the third set of analyses. Estimation of the model is carried out by using the Markov chain Monte Carlo (MCMC) estimation algorithm, which finds more applicability than the traditional

MML/EM estimation in estimating complex types of item response models due to its straightforwardness (Patz & Junker, 1999). By adopting a perspective of Bayesian inference, MCMC methods impose a prior distribution (often very weak) for each parameter in the model and estimate the full conditional posterior distribution of each parameter given the observed data and other parameters in the model. The basic idea of MCMC is to simulate a Markov chain whose stages represent a sample from the parameter's posterior distribution and the sample mean of the stages of the Markov chain is taken as the estimate of the parameter.

MCMC methods have been found to be particularly useful in estimating mixture distributions (Robert, 1996). In MRM, a class membership parameter is sampled for each examinee at each stage of the chain, along with a continuous latent ability parameter at each stage of the chain. Specifically, for each examinee, the class membership parameter is sampled from the conditional distribution of the examinee's membership in that class given the sampled item parameters and parameters for the mixing proportions. Similarly, the parameters for the mixing proportions are sampled from their posterior distributions conditional on the sampled class memberships, abilities for all examinees, and item parameters. The parameters for the mixing proportions are defined according to the frequencies with which the examinees are sampled into the classes over the stages of the chain. The frequency with which each examinee is sampled into each class determines the posterior probability of the examinee's membership in that class.

In this dissertation, the WinBUGS software (Spiegelhalter et al., 2003) is used to

implement the MCMC estimation of the mixture Rasch model. The ability parameters are assumed to be normally distributed within classes. The mean of the ability distribution is fixed to zero for each class, and the standard deviation is left to vary across classes. Item difficulty for each item within each class is assumed to follow a normal distribution with mean of zero and standard deviation of two.

Label-switching is a common problem in running discrete mixture models via MCMC estimation. The class labels permute during the simulation run, which makes the output difficult to interpret. Several methods have been proposed for handling the problem, including imposing constraints on the parameters (Richardson & Green, 1997), cluster-based relabeling of the simulated parameters (Stephens, 1997), and preassigning one or more observations to each component with certainty (Chung, Loken, & Schafer, 2004). In this dissertation, I attempt to solve the problem of label switching by using the last approach since it has been shown to be both simple and effective.

It should be noted that in this study guessing is not modeled in any of the three types of models. The influence of guessing can be incorporated in TESTFACT to estimate the parameters of the item factor model. I choose not to do that because I am not particularly interested in the influence of guessing and I want to compare the results from TESTFACT with results from the other two types of models, in which guessing can not be accommodated.

Another point that needs to be brought into attention is that the cluster sampling scheme of NAEP violates the random sampling assumption of Item Response Theory

(IRT)-based measurement models, and would have a non-negligible impact upon parameter estimation and interpretation of results. It tends to reduce the accuracy of parameter estimates and make significance tests more powerful than they should be. In this dissertation, no special treatment was done to account for the effect of cluster sampling because this study was not focused on the technical details of the models or statistical solutions for any assumption violation problems.

4.3 Computation of the fit indices

The three types of models are expected to lead to different results, which all make sense if understood within their own conceptual framework. The differences among the three sets of solutions and interpretations need to be evaluated on the basis of model fit. In this dissertation, information criteria, including AIC, BIC, and CAIC, are used for comparing the models. These fit statistics are especially useful in comparing models with a non-nested relationship, which is true of the three types of models compared in this study.

As discussed above, the information criteria are appropriate when maximum likelihood estimates (MLE) of model parameters are obtained. The first two types of models are estimated via marginal maximum likelihood (MML) estimation, and the model estimates can be used directly in the computation of the information criteria. The mixture Rasch models are estimated by using the MCMC algorithm and the posterior means of the model parameters approximate the MLEs if the sample size is sufficiently large. For example, Li, Cohen, Kim, and Cho (2006) used the BIC criterion with posterior means for mixture models similar to the ones proposed here as

obtained from the MCMC estimation.

Computing the information criteria for each model is essentially computing the log likelihood of the response data by using the estimated parameters of that model, since the other two components of information criteria, sample size and degrees of freedom, can be directly obtained from the model and the data. Specifically, the log likelihood of the response data under the assumption of each model can be computed in the same framework by holding the model parameters fixed at their estimated values. All the models are built on the same matrix of dichotomous item response data. The general form of the log likelihood is

$$\ln(L) = \sum_{i=1}^{N} \ln P(\underline{X_i}) = \sum_{i=1}^{N} \sum_{j=1}^{n} x_{ij} \ln(P_{ij}) + (1 - x_{ij}) \ln(1 - P_{ij})$$

where $\underline{X_i}$ is the response pattern of the *i*th person and $\underline{X_i} = \{x_{i1}, x_{i2}, ...x_{ij}, ...x_{in}\}$, and

P_{ij} = Prob ($X_{ij} = x_{ij}$ | item parameters)

= $\int$ Prob ($X_{ij} = x_{ij}$ | item parameters, person parameters) $\times$ P [person parameters]

∂ [person parameters]

The log likelihood is estimated for each model with that model's form for the probability of the item response x_{ij}. The form of the model and the nature of the item and person parameters are determined in each model. In each case, optimal estimates, either MLEs or Bayesian posterior means, are used for item parameters in the calculation, and person parameters are integrated out.

In the between-item multidimensional model, the item parameters $\boldsymbol{\xi}$, along with the covariances of the latent dimensions are estimated via MML estimation. To compute the log likelihood, I randomly sample 200 $\boldsymbol{\theta}$ vectors from the multivariate

normal distribution $N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, where

$\hat{\boldsymbol{\mu}} = [\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3]$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_{33} \end{bmatrix} .$$

$\hat{\sigma}_{11}$, $\hat{\sigma}_{22}$, $\hat{\sigma}_{33}$ are the estimated variances of the three latent dimensions, and $\hat{\sigma}_{21}$ (or $\hat{\sigma}_{12}$), $\hat{\sigma}_{31}$ (or $\hat{\sigma}_{13}$), $\hat{\sigma}_{32}$ (or $\hat{\sigma}_{23}$) are the estimated covariances among the three latent dimensions. The 200 $\boldsymbol{\theta}$ vectors are regarded as quadrature points and each quadrature is equally weighted. Given that all the items being analyzed are dichotomized and only the item difficulty is estimated for each item, the response probability for the *j*th item conditional on the *m*th quadrature $\boldsymbol{\theta}_m$, is given by

$$P(\mathbf{X}_j = 1; \mathbf{A}, \mathbf{B}, \hat{\boldsymbol{\xi}}_j \mid \boldsymbol{\theta}_\mathbf{m}) = \frac{\exp(\mathbf{b}_j \boldsymbol{\theta}_\mathbf{m} + \mathbf{a}_j' \hat{\boldsymbol{\xi}}_j)}{1 + \exp(\mathbf{b}_j \boldsymbol{\theta}_\mathbf{m} + \mathbf{a}_j' \hat{\boldsymbol{\xi}}_j)}$$

where $\hat{\boldsymbol{\xi}}_j$ is the estimated difficulty for the *j*th item, and **A** and **B** are the design and scoring matrices. The probability of obtaining the response pattern $\underline{X_i}$, conditional on the *m*th quadrature $\boldsymbol{\theta}_m$ is

$$P(\underline{X_i} \mid \boldsymbol{\theta}_m) = \prod_{j=1}^{n} P(x_{ij} = 1 \mid \boldsymbol{\theta}_m)^{x_{ij}} \left[1 - P(x_{ij} = 1 \mid \boldsymbol{\theta}_m)\right]^{1-x_{ij}} .$$

The unconditional probability of obtaining the response pattern $\underline{X_i}$ is approximated as

$$P(\underline{X_i}) = \sum_{m=1}^{200} P(\underline{X_i} \mid \boldsymbol{\theta}_m) \cdot A(\boldsymbol{\theta}_m) .$$

As mentioned above, each quadrature is equally weighted and $A(\boldsymbol{\theta}_m) = \frac{1}{200} = .005$. Thus, the unconditional probability of observing the response pattern $\underline{X_i}$ is

$P(\underline{X}_i) = \sum_{m=1}^{200} P(\underline{X}_i \mid \boldsymbol{\theta}_m) \cdot .005$. Therefore, the likelihood of obtaining the entire response matrix for all the persons is approximately

$L = \prod_{i=1}^{N} P(\underline{X}_i) = \prod_{i=1}^{N} \sum_{m=1}^{200} P(\underline{X}_i \mid \boldsymbol{\theta}_m) \cdot .005$ and the log likelihood is

$$\ln(L) = \sum_{i=1}^{N} \ln P(\underline{X}_i) = \sum_{i=1}^{N} \ln \sum_{m=1}^{200} P(\underline{X}_i \mid \boldsymbol{\theta}_m) \cdot .005$$

where N is the number of persons in the sample.

In the item factor analytic model, the item response function is given by $P(x_j = 1 \mid \boldsymbol{\theta}) = \Phi\left[a_{1j}\theta_1 + a_{2j}\theta_2 + ... + a_{kj}\theta_k + ... + a_{Mj}\theta_M + d_j\right]$ where Φ denotes the cumulative normal density function. The item parameters a's and d's, along with the correlations of the latent factors are estimated via MML estimation (Bock, Gibbons, & Muraki, 1988) in TESTFACT. Let's take a two-factor item model as an example. To compute the log likelihood, I randomly sample 200 $\boldsymbol{\theta}$ vectors from the multivariate normal distribution $N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, where

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1 & \hat{r}_{12} \\ \hat{r}_{21} & 1 \end{bmatrix}$$ and

$\hat{r}_{21}$ (or $\hat{r}_{12}$) is the estimated correlation among the two latent factors. The probability of obtaining the response pattern $\underline{X}_i$, conditional on the mth quadrature $\boldsymbol{\theta}_m$ is

$$P(\underline{X}_i \mid \boldsymbol{\theta}_m) = \prod_{j=1}^{n} P(x_{ij} = 1 \mid \boldsymbol{\theta}_m)^{x_{ij}} \left[1 - P(x_{ij} = 1 \mid \boldsymbol{\theta}_m)\right]^{1-x_{ij}}$$

$$= \prod_{j=1}^{n} \left[\Phi\left(a_{1j}\theta_{1m} + a_{2j}\theta_{2m} + d_j\right)\right]^{x_{ij}} \left[1 - \Phi\left(a_{1j}\theta_{1m} + a_{2j}\theta_{2m} + d_j\right)\right]^{1-x_{ij}}$$

where θ_{1m} and θ_{2m} are the components of $\boldsymbol{\theta}_m$.

Since the 200 $\boldsymbol{\theta}$ points are equally weighted, the unconditional probability of

obtaining the response pattern $\underline{X_i}$ is approximated as

$P(\underline{X_i}) = \sum_{m=1}^{200} P(\underline{X_i} \mid \boldsymbol{\theta}_m) \cdot .005$. Therefore, the likelihood of obtaining the entire response matrix for all the persons is approximately equal to

$L = \prod_{i=1}^{N} P(\underline{X_i}) = \prod_{i=1}^{N} \sum_{m=1}^{200} P(\underline{X_i} \mid \boldsymbol{\theta}_m) \cdot .005$ and the log likelihood is

$\ln(L) = \sum_{i=1}^{N} \ln P(\underline{X_i}) = \sum_{i=1}^{N} \ln \sum_{m=1}^{200} P(\underline{X_i} \mid \boldsymbol{\theta}_m) \cdot .005$ where N is sample size.

Calculating log likelihood is more complicated for the mixture Rasch model. In the mixture Rasch model, the response probability of the ith person to the jth item is given by

$$\mathrm{P}(X_{\mathrm{ij}} = 1 \mid \boldsymbol{\theta}_i, \beta_{gj}, \pi) = \sum_{g=1}^{G} \pi_g \frac{\exp(\theta_{gi} - \beta_{gj})}{1 + \exp(\theta_{gi} - \beta_{gj})}.$$

The item parameters β_{gj} and class proportions π_g are estimated via MCMC estimation in WinBUGS. In this study, the θ distribution for each latent class is assumed to be normal. The mean of the distribution is constrained to be equal to zero for each class, while the standard deviation of the distribution is freely estimated. Take the two-class mixture Rasch model as an example. To compute the log likelihood, I randomly select 100 θ's from the normal distribution $N(0, \hat{\sigma}_1)$, and $\hat{\sigma}_1$ is the estimated standard deviation of the first class. Another 100 θ's are randomly selected from the normal distribution $N(0, \hat{\sigma}_2)$, and $\hat{\sigma}_2$ is the estimated standard deviation of the second class. The 100 θ points from each distribution are regarded as quadrature points and each quadrature has a weight of .01. Let θ_{1m} denote the mth θ point selected from the distribution of the first class, and θ_{2m}

denote the *m*th θ point from the second class. Conditional on the *m*th quadrature point of the first class, θ_{1m}, the probability of obtaining the response pattern $\underline{X_i}$ is equal to

$$P(\underline{X_i} \mid \theta_{1m}) = \prod_{j=1}^{n} P(x_{ij} = 1)^{x_{ij}} \left[1 - P(x_{ij} = 1)\right]^{1-x_{ij}}$$

$$= \prod_{j=1}^{n} \left[\frac{\exp(\theta_{1m} - \beta_{1j})}{1 + \exp(\theta_{1m} - \beta_{1j})} \right]^{x_{ij}} \left[1 - \frac{\exp(\theta_{1m} - \beta_{1j})}{1 + \exp(\theta_{1m} - \beta_{1j})} \right]^{1-x_{ij}}.$$

Conditional on the knowledge that the *i*th person belongs to the first class, the probability of obtaining the response pattern $\underline{X_i}$ is approximated as

$$P(\underline{X_i} \mid \phi = 1) = \sum_{m=1}^{100} P(\underline{X_i} \mid \theta_{1m}) \cdot A(\theta_{1m}) = .01 \sum_{m=1}^{100} P(\underline{X_i} \mid \theta_{1m}).$$

If the *i*th person's class membership is unknown, the probability of obtaining the response pattern $\underline{X_i}$ is approximated as

$$P(\underline{X_i}) = \sum_{g=1}^{2} \hat{\pi}_g \cdot P(\underline{X_i} \mid \phi = g) = P(\underline{X_i} \mid \phi = 1) \cdot \hat{\pi}_1 + P(\underline{X_i} \mid \phi = 2) \cdot \hat{\pi}_2)$$

$$= \hat{\pi}_1 \cdot .01 \sum_{m=1}^{100} P(\underline{X_i} \mid \theta_{1m}) + \hat{\pi}_2 \cdot .01 \sum_{m=1}^{100} P(\underline{X_i} \mid \theta_{2m})$$

where $\hat{\pi}_1$ and $\hat{\pi}_2$ are the estimated class proportions obtained from the two-class mixture Rasch model solution. Therefore, the likelihood of obtaining the entire response matrix for all the persons is approximately

$L = \prod_{i=1}^{N} P(\underline{X_i}) = \prod_{i=1}^{N} \sum_{g=1}^{2} \hat{\pi}_g \cdot P(\underline{X_i} \mid \phi = g)$, and the log likelihood is

$\ln(L) = \sum_{i=1}^{N} \ln P(\underline{X_i}) = \sum_{i=1}^{N} \ln \sum_{g=1}^{2} \hat{\pi}_g \cdot P(\underline{X_i} \mid \phi = g)$ where *N* is the sample size.

Computing the information criteria involves identifying the number of parameters to be estimated and sample size for each model, in addition to calculating the log

likelihood. In this study, all the models are used to analyze the same response data and the sample size is the number of examinees from whom the responses were obtained. The number of parameters varies across the models. For the between-item multidimensional model, the parameters to be estimated include item difficulties, the means of the hypothesized dimensions, and the unique elements of the variance-covariance matrix. To solve the identification problem, the mean of the item difficulties for each dimension is constrained to be zero. This is achieved by fixing the item difficulty of the last item on each dimension to be equal to the negative sum of the difficulties of the other items on that dimension. Suppose there are *J* items in the response data and the proposed model contains *D* dimensions, the number of parameters estimated in the model is equal to $J + \frac{D \cdot (D+1)}{2}$. For an item factor analytic model, the parameters to be estimated include item thresholds and factor loadings for all the items, minus the number of constraints. For an *M*-factor model, the number of constraints is equal to $M \cdot (M-1)/2$, and the number of estimated parameters to is equal to $J \cdot M + J - M \cdot (M-1)/2$. In the context of MCMC, the number of parameters estimated in a mixture Rasch model is the sum of item difficulties across all classes, the number of means of theta for all classes, and the number of class proportion estimates. Suppose the proposed mixture Rasch model has G classes, the number of parameters is equal to $J \cdot G + G + (G-1)$.

Chapter 5: Results and Discussion

In this dissertation, data analysis was first done on a single block of items. Students who responded to the items were the subjects of analysis. The three types of models were fit to the response data and results of model analysis were summarized in terms of substantive meaningfulness and statistical fit. Information criteria, including AIC, BIC, and CAIC, were computed for each model. Based on these indices, goodness of fit was compared both within and across types of models. Narrative stories about science proficiency were compared across models for a small number of examinees, and commonalities and distinctions among those stories were discussed. Besides, as a follow-up to mixture model analysis, examinees' demographic and background variables were extracted and their associations with latent class membership were studied.

Replicative data analysis was done on two other data sets to examine the generalizability of the findings found in the block-level analysis. As mentioned above, data analysis could be performed on an entire booklet. However, only four item blocks of the 1996 NAEP science assessment for Grade 8 were released for secondary use and not any three of them were bundled together as a booklet. Therefore, I picked two combinations of item blocks, each of which was administered as part of a booklet, and studied examinees' responses to the block combinations. Similarly, the three types of models were fit to the response data and analysis results were again compared across models in terms of substantive significance and statistical fit.

The last section of this chapter is a synthesis of the analysis results found across the three data sets, with their commonalities and differences highlighted and discussed. Implications of the findings to assessment design are also discussed.

5.1 Analysis results of a single item block

Block S20 was selected as the target of analysis because it was one of the publicly released blocks and detailed information about the items in the block was available. It was made up of 8 multiple-choice and 8 constructed-response items. It contained 6 items on physical science, 6 on earth science, and 4 on life science. There were 1251 8th graders to whom Block S20 was administered. Responses to the multiple-choice and short constructed-response items were scored as right or wrong, and responses to the extended constructed-response items were originally scored as discrete integer values within certain bounds (from 0 to 3 or 4). As mentioned above, responses to polytomously scored items were dichotomized in this study in order to simplify data analysis.

A preliminary exploratory factor analysis was performed on the response matrix in TESTFACT, and the factor loadings of the one-, two-, and three-factor solutions are summarized in Tables 1-3. The results of factor analysis suggested that Item 10 behaved oddly. In the one-factor solution, its loading on the single factor was negligible (= .091), while the other items had moderate to high loadings on that factor. In the two-factor solution, it did not load substantially on either of the two factors. In the three-factor solution, it loaded highly (= .881) on the third factor, while all the other items had negligible loadings on that factor. In addition, the estimates of item

parameters provided by NAEP indicated that Item 10 was a very difficult item (b = 4.0574), but students of the lowest ability had a 20% chance of getting it right (c = .1986). This is saying that the factor on which this item loaded highly was not a psychometrically meaningful factor but related to whether a student was a lucky guesser on that item. Due to this reason, I dropped this item in further analyses. (See Appendix A for Item 10. I suspect that most students chose a wrong answer to this item because they did not know the word "mitochondrion", which was the key word in the question, and those who got it right simply made lucky guesses.)

Table 1: Factor loadings of the one-factor model for all the items in Block S20

Item	Factor 1
S20_1	.399
S20_2	.550
S20_3	.504
S20_4	.090
S20_5	.665
S20_6	.495
S20_7	.532
S20_8	.604
S20_9	.338
S20_10	**.091**
S20_11	.346
S20_12	.248
S20_13	.529
S20_14	.508
S20_15	.730
S20_16	.543

Note: The item in bold, Item 10, has a negligible loading on the single factor, on which almost all the other items have substantial loadings.

Table 2: Factor loadings of the two-factor model for all the items in Block S20

Item	Factor 1	Factor 2
S20_1	.434	-.019
S20_2	.149	.461
S20_3	.068	.497
S20_4	-.180	.278
S20_5	.019	.782
S20_6	.561	-.022
S20_7	.261	.334
S20_8	.508	.129
S20_9	.129	.224
S20_10	**.217**	**-.130**
S20_11	.146	.231
S20_12	.290	-.040
S20_13	.606	-.050
S20_14	.500	.040
S20_15	.493	.314
S20_16	.650	-.054

Note: The item in bold, Item 10, has negligible loadings on both factors.

Table 3: Factor loadings of the three-factor model for all the items in Block S20

Item	Factor 1	Factor 2	Factor 3
S20_1	.458	.050	-.066
S20_2	.265	-.029	.353
S20_3	.181	.002	.397
S20_4	-.091	.014	.204
S20_5	.057	.093	.826
S20_6	.594	.069	-.090
S20_7	.408	-.122	.188
S20_8	.612	-.025	-.003
S20_9	.202	.006	.150
S20_10	**-.015**	**.881**	**.110**
S20_11	.247	-.115	.123
S20_12	.248	.139	-.005
S20_13	.629	-.004	-.101
S20_14	.574	-.031	-.063
S20_15	.598	-.056	.185
S20_16	.658	.076	-.092

Note: The item in bold, Item 10, has a substantial loading on factor 2, on which all the other items have negligible loadings. However, it has negligible loadings on factor 1 and factor 3, each of which is strongly indicated by a number of items. Based on the factor loadings of Item 10 in the three factor model solutions, we can make a conclusion that Item 10 is not measuring a psychometrically interesting factor and should be excluded in further analyses.

Analysis results for Block S20 are organized into four sections: results of model analyses, comparison of model fit by information criteria, comparison of narrative stories for selected examinees, and background characteristics of latent classes.

Results of model analyses for each model are summarized and discussed in three

sub-sections: the model/narrative frame, results in term of parameters and fit, and results in terms of substance.

5.1.1 Results of model analyses

5.1.1.1 The between-item multidimensional model

- The model/narrative frame

The between-item multidimensional model represents test structure in terms of several parallel unidimensional subscales. Each subscale is associated with a distinct set of items and no item is common across subscales. The association between items and subscales is determined in the test design and item writing stage. Therefore, the between-item multidimensional model is often used in confirmatory analysis to evaluate the hypothesized test structure.

In the between-item multidimensional model, items are clustered along predefined lines and item parameters are estimated on a priori grounds. Specifically, each item is categorized as an indicator of one subscale, and the parameter(s) of the item is estimated on that subscale. Students are characterized by a set of proficiency scores along the dimensions that are intended to be covered in the test. Distinctions along the hypothesized dimensions are believed to lead to different response patterns among the examinees.

According to the design framework of the NAEP science assessment, each item block can be treated as a multidimensional testing instrument, whose structure is described by a three-dimensional between-item model. The three dimensions correspond to the three fields of science or the three science process skills. Figures 4

and 5 depict the between-item multidimensional structure of Block S20 in terms of content areas and process skills, respectively. Based on the test structure illustrated in Figure 4, a three-dimensional between-item model was fit to the response matrix to test the hypothesis that the item block is multidimensional due to item content. Similarly, a three-dimensional between-item model was fit to the response matrix to test the hypothesis that the item block is multidimensional due to science process skills, based on what is illustrated in Figure 5.

Figure 4: Between-item multidimensionality in terms of content areas for Block S20 (excluding Item 10)

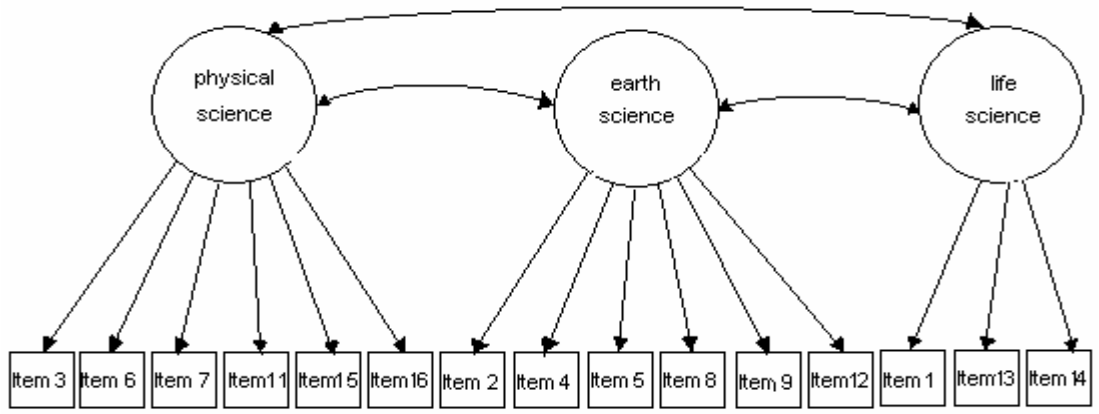
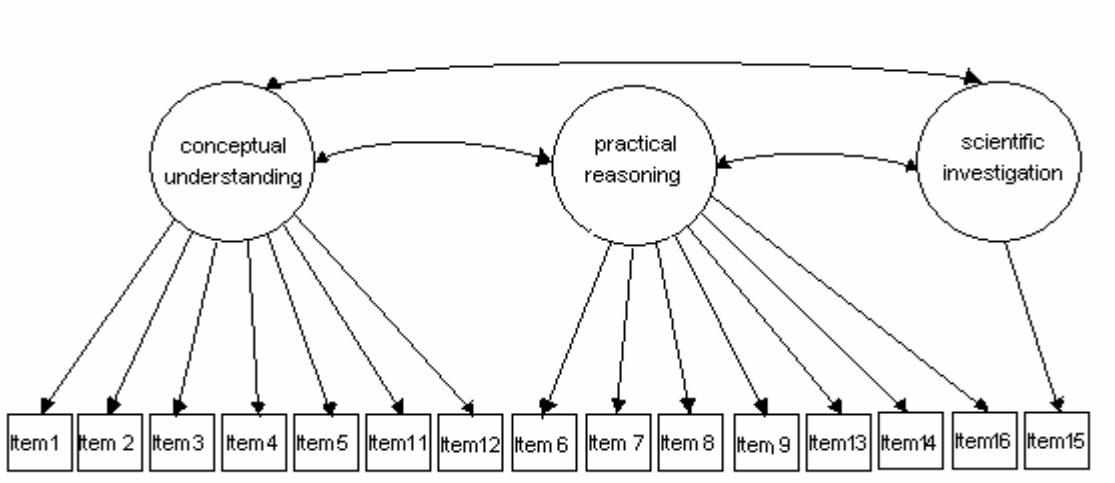


Figure 5: Between-item multidimensionality in terms of science process skills for Block S20 (excluding Item 10)



- Results in terms of parameters and fit

The parameters of the multidimensional between-item models were estimated in ConQuest. In estimating the item difficulties, constraints were applied so that the mean of the item difficulties was zero on each dimension. This was done by setting the difficulty of the last item in each dimension equal to the negative sum of the difficulties of the rest of the items on that dimension. The total number of parameters estimated in the model was equal to 21, which included twelve item difficulties, the means of the three latent distributions, and the six unique elements of the variance-covariance matrix. Below are the results from the two sets of confirmatory analysis.

Hypothesis 1: The item block is multidimensional due to item content.

Under this hypothesis, items are indicators of three content areas. Table 4 summarizes the estimates of item difficulties for the 15 items along with their standard errors and diagnostic statistics of fit. It should be noted that standard errors and fit statistics were not provided for constrained parameters. Based on the criterion discussed in Chapter 4, the analysis results suggest that all the items fit acceptably except for Items 2, 8, 11, and 15.

Table 4: Item difficulty estimates and fit statistics from ConQuest for Block S20 excluding Item 10 (The dimensions are defined in terms of content areas)

Item	Dimension	Estimate	Standard Error	Unweighted Fit MNSQ	Unweighted Fit T	Weighted Fit MNSQ	Weighted Fit T
S20_1	Life	-0.445	0.046	1.03	0.8	1.00	0.1
S20_2	Earth	-0.766	0.043	1.12	2.9	1.10	**3.1**
S20_3	Physics	-1.266	0.046	1.05	1.4	1.06	1.7
S20_4	Earth	2.033	0.053	1.08	1.9	1.07	1.9
S20_5	Earth	0.904	0.047	1.04	0.9	1.04	1.1
S20_6	Physics	0.715	0.049	0.97	-0.7	0.98	-0.6
S20_7	Physics	0.075	0.047	0.99	-0.1	1.01	0.4
S20_8	Earth	-1.594	0.045	1.09	2.2	1.08	**2.2**
S20_9	Earth	0.326	0.045	1.07	1.6	1.05	1.4
S20_11	Physics	0.128	0.047	1.13	3.1	1.12	**3.1**
S20_12	Earth	-0.902*					
S20_13	Life	-0.114	0.045	1.00	0.1	1.01	0.3
S20_14	Life	0.559*					
S20_15	Physics	0.303	0.048	0.86	-3.7	0.87	**-3.8**
S20_16	Physics	0.044*					

Notes: 1. * indicates that the item difficulty was constrained.
2. A weighted T statistic with an absolute value larger than 2 suggests moderate misfit. Items in bold are moderately misfit items.

The estimated mean of the latent distribution was -.828 for physical science, -.538 for earth science, and .661 for life science. This is not saying that this sample of students was more able in life science than in physical science or earth science. These were three separate scales and, as mentioned before, the origin of each of the three scales was set by making the mean of the item difficulties zero on each scale.

Therefore, the three scales did not have a common origin or unit. A general statement that can be made was that an average student did better in an average item in life science than in physical science or earth science. No constraints were placed on the variances. The estimated variance of the latent distribution was 1.144 for physical science, 0.759 for earth science, and 1.086 for life science. This implies that the variability of students' abilities as assessed by the test items was larger in physical science and life science than in earth science.

A unidimensional model was fit to the same response data and its goodness of fit was compared with that of the three-dimensional model. The deviance for the unidimensional model was 21486.669, and the number of estimated parameters was equal to 16. The deviance for the three-dimensional model was 21477.017, and the number of estimated parameters was equal to 21. The difference between the two deviances follows an asymptotic chi-square distribution with degrees of freedom equal to the difference between the numbers of estimated parameters for the two models. The deviance difference between the two models was equal to 9.652 with 5 degrees of freedom. A significance test of the difference statistic suggests that the multidimensional model does not fit significantly better than the one-dimensional model because the p-value ($p = .086$) of the statistic is larger than the nominal level of .05.

Hypothesis 2: The item block is multidimensional due to science process skills.

Under this hypothesis, items are indicators of three science process skills. Table 5 summarizes the estimates of item difficulties for the 15 items along with their

standard errors and diagnostic statistics of fit. Again, standard errors and fit statistics were not provided for constrained parameters. Based on the same criterion as was used before, all the items fit acceptably except for Items 2 and 13. Clearly, the number of misfit items under this hypothesis was smaller than the number of misfit items under the previous hypothesis.

Table 5: Item difficulty estimates and fit statistics from ConQuest for Block S20 excluding Item 10 (The dimensions are defined in terms of science process skills)

Item	Dimension	Estimate	Standard Error	Unweighted Fit		Weighted Fit	
				MNSQ	T	MNSQ	T
S20_1	Conceptual understanding	-1.440	0.045	1.02	0.6	1.01	0.2
S20_2	Conceptual understanding	-0.628	0.043	1.09	2.1	1.08	**2.6**
S20_3	Conceptual understanding	-0.811	0.043	1.06	1.5	1.06	1.7
S20_4	Conceptual understanding	2.137	0.052	1.10	2.3	1.07	1.9
S20_5	Conceptual understanding	1.020	0.047	1.04	1.0	1.03	0.9
S20_6	Practical reasoning	1.212	0.049	0.93	-1.7	0.96	-1.0
S20_7	Practical reasoning	0.578	0.047	0.94	-1.5	0.99	-0.1
S20_8	Practical reasoning	-1.424	0.048	0.95	-1.4	0.95	-1.6
S20_9	Practical reasoning	0.591	0.047	1.10	2.5	1.05	1.5
S20_11	Conceptual understanding	0.486	0.045	1.07	1.7	1.05	1.6
S20_12	Conceptual understanding	-0.762*					
S20_13	Practical reasoning	-1.089	0.046	0.93	-1.9	0.92	**-2.4**
S20_14	Practical reasoning	-0.416	0.045	0.98	-0.5	0.99	-0.3
S20_15	Scientific investigation	0.000*					
S20_16	Practical reasoning	0.548*					

Notes: 1. * indicates that the item difficulty was constrained.
2. A weighted T statistic with an absolute value larger than 2 suggests moderate misfit. Items in bold are moderately misfit items.

The estimated mean of the latent distribution was -.404 for conceptual understanding, -.318 for practical reasoning, and -1.470 for scientific investigation. Again, a general statement can not be made about students' proficiencies with regard to the three process skills because the latent dimensions were estimated in a way that they did not have a common origin. The estimated variance of the latent distribution was 0.673 for conceptual understanding, 1.076 for practical reasoning, and 3.769 for scientific investigation. The large variance associated with the dimension of scientific investigation was due to the fact that only one item was scaled on this dimension and the examinees' performances on that item varied a lot.

The three-dimensional between-item model was again compared with the unidimensional model in terms of goodness of fit. The deviance for the three-dimensional model was 21414.225 and the number of estimated parameters was equal to 21. The deviance difference between the three-dimensional model and the unidimensional model was equal to 72.444 with 5 degrees of freedom. A significance test of the chi-square difference statistic suggests that the multidimensional model fits significantly better than the unidimensional model because the p-value (p = .000) of the statistic is smaller than the nominal level of .05.

- Results in terms of substance

A comparison of the two sets of analysis suggests that multidimensionality in terms of cognitive processes makes more sense than multidimensionality in terms of content areas for the observed response data. If subscales are to be used for score reporting on the basis of this item block, they should be defined in terms of cognitive

processes rather than content areas.

The estimated correlations among the three process skills were .902, .915, and .942. The large values of the correlations imply that reporting an overall score for each examinee may be adequate for summarizing his (or her) performance on the test. Depending on the constraints that must be accommodated and the resources that are available, test developers can choose between using several subscales, which is informative but costly, and using a single scale, which is less expensive but less informative.

The high correlations among the cognitive process skills resulted in the high correlations among the three content areas, which were estimated to be .906, .889, and .846. This is so because the cognitive factors are believed to cross the disciplinary boundaries of the content areas. Although assigning each examinee a set of subscores, each corresponding to a field of science, can be justified for practical purposes, it provides no more information than reporting a single overall score. Besides, students' individual differences in science learning can be more accurately described in terms of cognitive process skills than subject matter knowledge.

5.1.1.2 The exploratory item factor analytic model

- The model/narrative frame

The exploratory item factor analytic model does not impose any predefined structure on the test, except that there are a number of latent factors that control the examinees' responses to all the test items. The test structure is determined by the empirical data. There may be one factor that by itself adequately explains the item

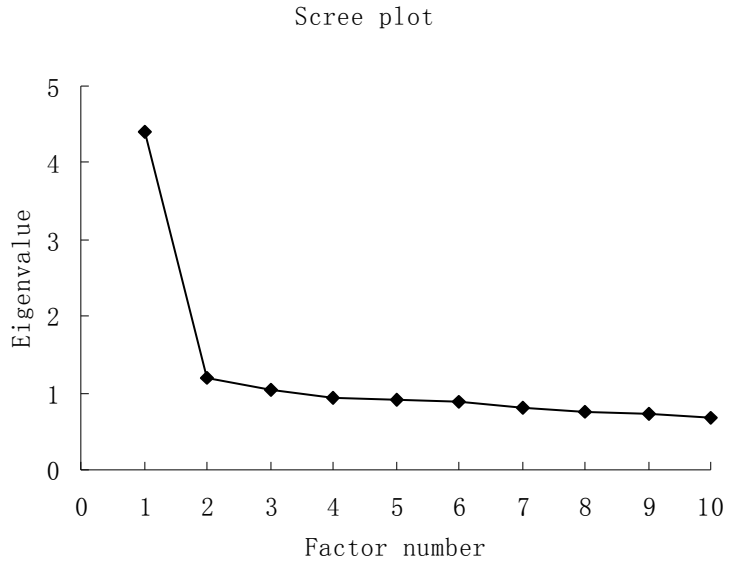
responses, or there may be several factors that together account for the variance of the response patterns. The loadings of the items on the factors are not predetermined, either. Each item may load significantly on one or more factors.

What is of particular interest in this dissertation is to investigate whether the factors that come out in the exploratory item factor analytic model solution correspond to the content domains which, according to NAEP, summarize student performance across all items in the assessment.

- Results in terms of parameters and fit

The exploratory item factor analytic models were estimated in TESTFACT. The scree plot in Figure 6 shows that the first factor explains a large proportion of variance in the items while the remaining factors, as compared to the first factor, are insignificant. Besides, in the TESTFACT output, the first five largest latent roots of the smoothed tetrachoric correlation matrix (with the unit diagonal elements replaced by the communalities) are equal to 3.72, 0.41, 0.21, 0.13, and 0.11. As stated in the manual, the number of factors that underlie persons' responses can be set equal to the number of latent roots that are larger than 1. The result indicates that only the first latent root is greater than 1. Thus, it appears that the one-factor model is sufficient for accounting for the data.

Figure 6: Scree plot from exploratory factor analysis for Block S20 (excluding Item 10)



A range of factor solutions including one-, two-, three-, and four-factor models were estimated and their RMSRs are summarized in Table 6. As mentioned before, the criterion for an acceptable factor solution is that RMSR should be less than .05. It turns out that all the four models have acceptable RMSRs. Selecting a preferable model among the four models can be based on another criterion. According to Tate (2003), an additional factor can be added to the model until the reduction of the RMSR index is less than roughly 10%. Judged by this rule, the three-factor model is preferred.

Table 6: Root mean square residuals (RMSRs) for the one-, two-, three-, and four-factor models for Block S20 (excluding Item 10)

Model	RMSR	Percentage of reduction
1-factor model	0.048	
2-factor model	0.037	23%
3-factor model	0.033	11%
4-factor model	0.032	3%

The assessment of dimensionality can also be based on a test of the chi-square difference statistic since the set of models being compared have a nesting relationship. The chi-square difference statistic is asymptotically distributed as a chi-square distribution, with its degrees of freedom equal to the difference between the degrees of freedom of the two models. Table 7 summarizes the chi-square statistics for the four models. Table 8 displays the tests of the chi-square difference statistics. The results suggest that the two-factor model fits significantly better than the one-factor model, the three-factor model does not fit significantly better than the two-factor model, and the four-factor model does not fit significantly better than the three-factor model. Thus, the two-factor model is selected as the best model by this criterion.

Table 7: Chi-square statistics for the one-, two-, three-, and four-factor models for Block S20 (excluding Item 10)

# of factors	Chi-square	DF	P-value
1	3430.39	1220	0.000
2	3385.79	1206	0.000
3	3372.15	1193	0.000
4	3355.90	1181	0.000

Table 8: Tests of chi-square difference statistics between factor models for Block S20 (excluding Item 10)

	Chi-square difference	DF difference	P-value
2-factor vs. 1-factor	44.60	14	0.000
3-factor vs. 2-factor	13.64	13	0.423
4-factor vs. 3-factor	16.25	12	0.180

Finally, the selection of factors should be based on the pattern of factor loadings of each factor solution. Items that have substantial loadings on a factor are considered its salient indicators and used to interpret the factor. In the one-factor solution almost all of the items loaded substantially on the factor. This implies the existence of a general factor that explains a considerable amount of variance of the items. It also agrees with the finding from examining the scree plot.

Table 9 summarizes the factor loadings of the two-factor solution. A conclusion that can be drawn from the two-factor solution is that the two factors did not represent distinct content domains. As a matter of fact, almost all the items that loaded highly on the first factor were designed to assess abilities in practical reasoning or scientific

investigation. There was one exception. Item 1, which was designed to assess conceptual understanding, loaded substantially on the first factor, too. In contrast, all the items that loaded highly on the second factor were designed to assess conceptual understanding, although not all the items designed to assess conceptual understanding loaded substantially on that factor. Therefore, the first factor was interpreted as representing what students can do in science and the second factor was interpreted as indicating what they know in science.

Table 9: Factor loadings of the two-factor model for Block S20 (excluding Item 10)

Item No.	Content	Process	Factor 1	Factor 2
S20_1	Life	Conceptual understanding	**.455**	-.060
S20_2	Earth	Conceptual understanding	.246	**.380**
S20_3	Physics	Conceptual understanding	.169	**.415**
S20_4	Earth	Conceptual understanding	-.126	.240
S20_5	Earth	Conceptual understanding	.093	**.769**
S20_6	Physics	Practical reasoning	**.592**	-.081
S20_7	Physics	Practical reasoning	**.368**	.230
S20_8	Earth	Practical reasoning	**.586**	.031
S20_9	Earth	Practical reasoning	.182	.174
S20_11	Physics	Conceptual understanding	.237	.140
S20_12	Earth	Conceptual understanding	.269	-.030
S20_13	Life	Practical reasoning	**.625**	-.096
S20_14	Life	Practical reasoning	**.565**	-.047
S20_15	Physics	Scientific investigation	**.606**	.194
S20_16	Physics	Practical reasoning	**.650**	-.084

Note: Numbers in bold are substantial loadings (>.30) on the two factors.

The pattern of factor loadings of the three-factor model did not convey so much

conceptual meaning as that of the two-factor model. The first and third factors in the three-factor solution resembled the two factors in the two-factor solution, and the second factor did not have much conceptual meaning. The four-factor model solution was not interpretable, either.

Based on the application of all the above-mentioned criteria, the two-factor model was considered superior to the other three models in terms of conceptual significance and statistical fit. It should be noted that the solution of the two-factor model discussed above was obtained through a promax rotation of the initial solution. The two factors were allowed to be correlated with each other and the estimated correlation was .650. This echoes the fact that the knowing and doing aspects of science learning are closely related to each other.

- Results in terms of substance

The dichotomization between knowing and doing in science is another meaningful way of conceptualizing science achievement. It agrees with the previous finding that what underlies item responses is not subject-matter knowledge of fields of science, but cognitive proficiencies developed in the course of science learning. Characterizing students by these two aspects of science proficiency can be beneficial to instruction and learning, in that students' factor scores will give us an indication about the within-person and across-person differences with respect to these two factors.

The two-factor model solution suggests that some items may require significant amounts of both aspects of science proficiency. This results from the fact that the

two aspects of science proficiency are closely connected with each other and a requirement of one aspect often necessitates the other. Recognizing the interaction between these two aspects of science proficiency provides the basis for designing appropriate task situations that meet the desired goals of assessment (Baxter & Glaser, 1998)

5.1.1.3 The mixture Rasch model

- The model/narrative frame

The mixture Rasch model conceptualizes persons as coming from one of several latent classes, each of which has a distinct ability distribution. Persons from the same latent class have qualitatively similar response patterns to the test items, and the dissimilarities among their responses are accounted for by variations along the latent dimension associated with that particular class. The mixture Rasch model is often used in exploratory analysis to investigate the number of latent classes and how the item parameters are different from one class to another.

The features of latent classes are identified with items that display varying patterns across classes. Students are characterized by the probabilities of belonging to each latent class and the latent abilities along the dimensions associated with the latent classes. In this study, it is of particular interest to test whether the dimensions specified by the latent classes correspond to fields of science or cognitive process skills. If that is the case, students of different latent classes can be distinguished by these factors that are meaningful to science learning and teaching.

- Results in terms of parameters and fit

A two-class mixture Rasch model was fit to the response data in WinBUGS. Five chains with over-dispersed starting values of class proportions were run. In a preliminary run of 3000 iterations, the problem of label switching was observed. In this study, I used the procedure proposed by Chung, Loken, and Schafer (2004) to solve the problem of label switching. I randomly picked one chain and selected for each class a number of students whose class membership was consistent over the iterations. I added those constraints to the model as priors and started the second run. Chung, Loken, and Schafer (2004) suggested that one prior for one latent class is enough in the two latent class case. However, in our study, adding a prior for each class was far from being enough. Instead, eleven priors were selected for each class. A total of 10,000 iterations were simulated for each of the five chains. The first 5000 iterations were discarded as burn-ins, and the 5000 iterations after burn-in were sampled for each chain. Thus, posterior estimates of the model parameters were calculated from a total of 25,000 iterations. The same procedure was followed in estimating the three-class mixture Rasch model. The only difference was that in the three-class mixture model case, ten priors were selected for each class and 7000 iterations were discarded as burn-ins. Again, 5000 iterations after burn-in were run for each of the five chains, and posterior estimates of the model parameters were based on a total of 25,000 iterations.

One thing that needs to be mentioned is that about 19 percent of the students in the sample did not finish all the items due to the time limit, and about 64 percent of the

students gave incorrect responses to more than half of the items in the block. Thus, for a large number of students, their class membership permuted during the span of iterations and they could not be assigned class labels with an acceptable level of certainty.

The two-class mixture Rasch model solution

Checking convergence is a necessary step in MCMC estimation. It is done by examining whether the simulated Markov chain converges to a stationary distribution, i.e. the posterior distribution of the parameter being monitored. For a model with many parameters, it is impractical to check convergence for every parameter. Instead, a random subset of parameters is selected for convergence checking. Two approaches are generally used in assessing convergence (for a more formal approach to convergence diagnosis, please refer to Brooks& Gelman, 1998). The first approach is to examine trace plots of the sample values versus iteration to see when the simulation appears to have stabilized. Second, we can look at the history plot, which shows the full history of the sample values for the parameter being monitored. Model estimation in WinBUGS often involves running multiple chains simultaneously, with each chain starting from a distinct set of initial values for the parameters being estimated. In that case, if all the chains in the trace plot or history plot appear to be overlapping one another, we have evidence to claim convergence. Figure 7 shows the trace plots for a subset of parameters. For each parameter, all the five chains are mixing well and have converged to a stabilized distribution before 5000 iterations are completed. Figure 8 shows the history plots for the same parameters. Again, for

each parameter, the five chains appear to have converged to a stationary distribution.

Figure 7: Trace plots for a subset of parameters being monitored in the two-class mixture Rasch model for Block S20 (excluding Item 10)

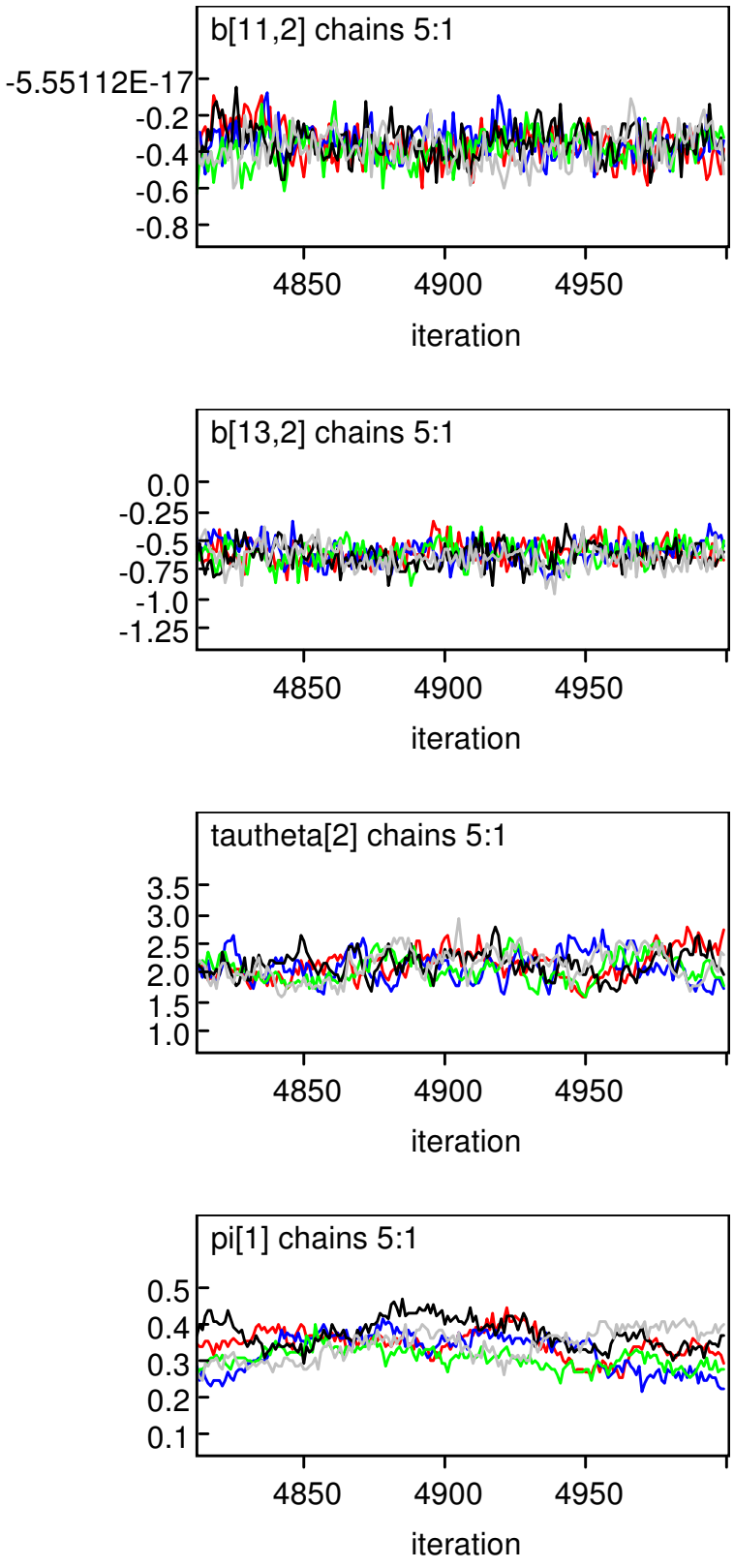


Figure 8: History plots for a subset of parameters being monitored in the two-class mixture Rasch model for Block S20 (excluding Item 10)

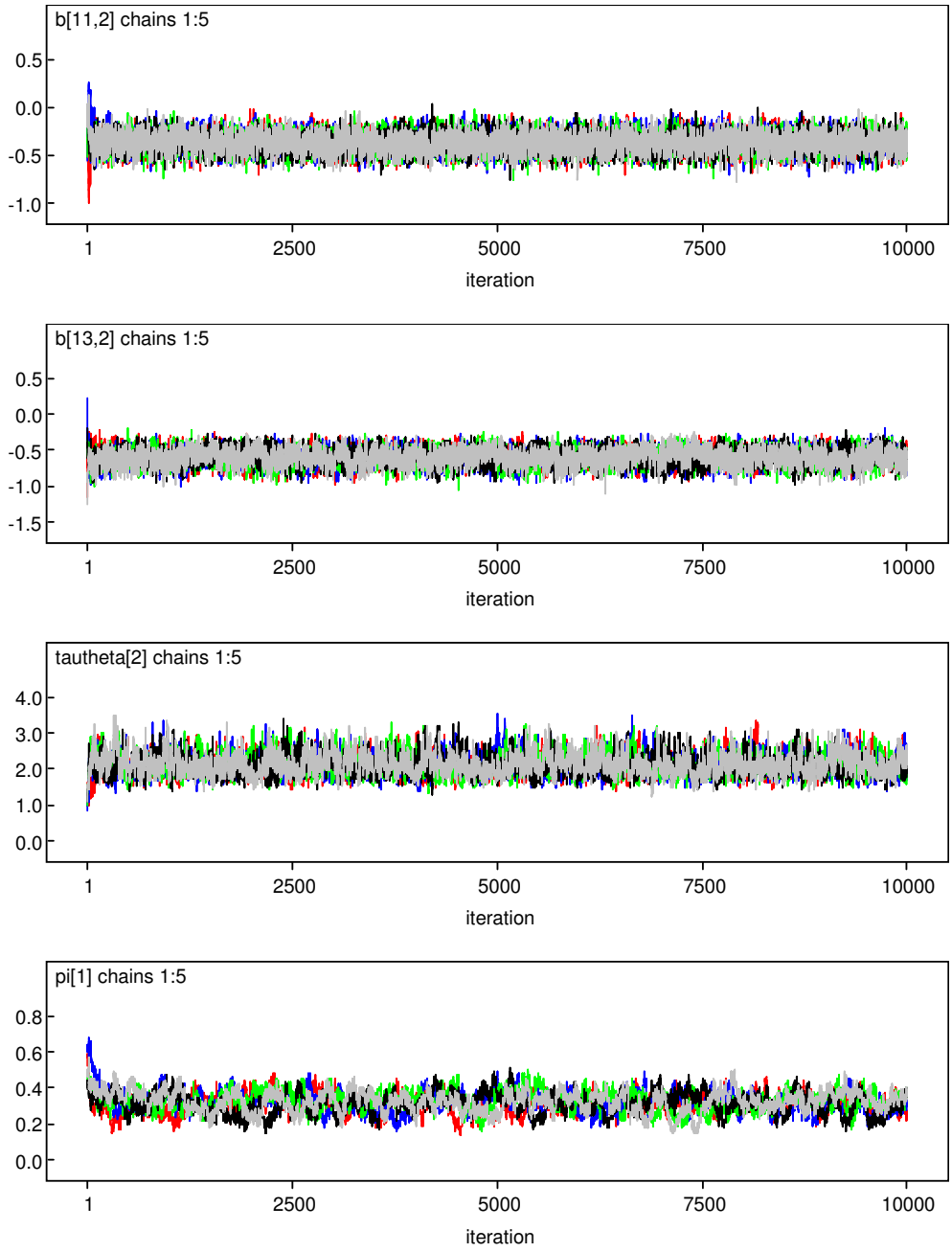


Table 10 summarizes the item difficulties obtained from the 2-class solution. All the items, except Item 4, were more difficult for students in Class 1 than for those in Class 2. This implies that Class 2 was made up of generally more able students, while Class 1 was made up of less able students. This general pattern reiterates the conclusion drawn from the previous analyses that the assessment is essentially

unidimensional. Besides, the finding that Item 4 was more difficult for students in the better-performing group agrees with the one-factor solution in the exploratory item factor analysis. Item 4 loaded negligibly on the factor while almost all the other items had substantial loadings on it. A plausible explanation is that Item 4 was measuring some latent trait other than the science proficiency that was measured by all the other items. Estimates of class proportions indicate that about 31 percent of the sampled students belonged to the first class and about 69 percent belonged to the second class.

Table 10: Item difficulties of the 2-class mixture Rasch model for Block S20 (excluding Item 10)

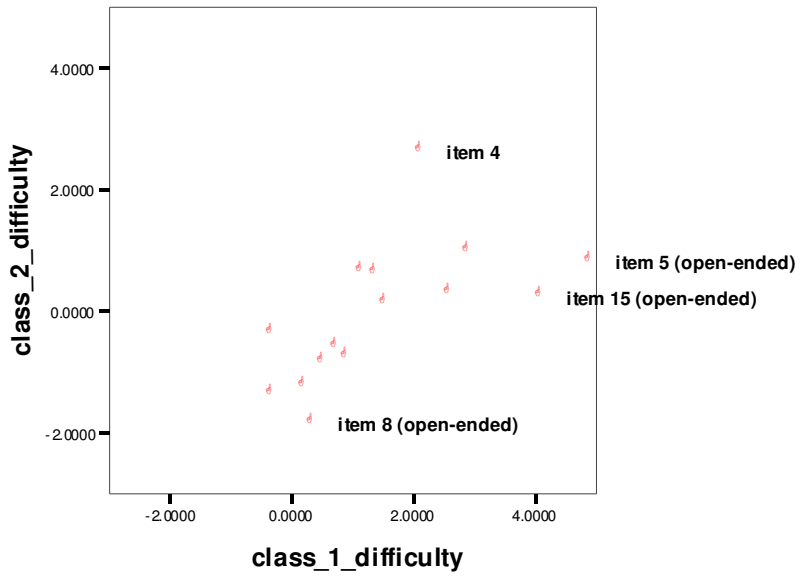
Item	Content	Process	Class 1	Class 2
S20_1	Life	Conceptual understanding	-.3425	-1.361
			(.2007)	(.1058)
S20_2	Earth	Conceptual understanding	.8876	-.7457
			(.1897)	(.1428)
S20_3	Physics	Conceptual understanding	.4909	-.8414
			(.1719)	(.125)
S20_4	Earth	Conceptual understanding	2.099	2.623
			(.2287)	(.1522)
S20_5	Earth	Conceptual understanding	4.876	.8283
			(.851)	(.1557)
S20_6	Physics	Practical reasoning	2.881	.9915
			(.5534)	(.1089)
S20_7	Physics	Practical reasoning	2.578	.3068
			(.6365)	(.1114)
S20_8	Earth	Practical reasoning	.3313	-1.851
			(.2884)	(.1519)
S20_9	Earth	Practical reasoning	1.133	.6668
			(.1671)	(.102)
S20_11	Physics	Conceptual understanding	1.356	.6408
			(.1838)	(.09957)
S20_12	Earth	Conceptual understanding	-.3345	-.3719
			(.1614)	(.09099)
S20_13	Life	Practical reasoning	.1893	-1.231
			(.2288)	(.1125)
S20_14	Life	Practical reasoning	.7343	-.5919
			(.2298)	(.1054)
S20_15	Physics	Scientific investigation	4.078	.2567
			(.8407)	(.1626)
S20_16	Physics	Practical reasoning	1.534	.1323
			(.3351)	(.102)

Note: Numbers in parentheses are standard deviations of the posterior distributions for the estimated parameters.

The scatter plot of item difficulties for the two classes, as displayed in Figure 9, gives a closer look at the items that distinguished between students in Class 1 and those in Class 2. Specifically, items that are far away from the diagonal of the

scatter plot provide some useful information about the characteristics of students in the two classes. From the scatter plot, it can be seen that Items 5, 8 and 15 were particularly hard for students in Class 1. These three items collectively covered the three cognitive skills in the domains of physical science and earth science. They were all open-ended items that required written explanations. This finding confirms the conclusion that students in Class 1 were less successful in learning science than those in Class 2 with respect to both cognitive skills and content knowledge. Besides, a plausible explanation for the poor performance of Class 1 in these items is that they were particularly weak in organizing and explaining their thoughts on scientific procedures, facts, or phenomena.

Figure 9: Scatter plot of item difficulties of the two-class mixture Rasch model for Block S20 (excluding Item 10)



The three-class mixture Rasch model solution

The same procedure was followed in estimating the three-class mixture Rasch

model except that in the three-class case ten priors were selected for each class and 7000 iterations were discarded as burn-ins. Figure 10 shows the trace plots for a subset of model parameters being monitored in the 3-class model. Figure 11 shows the history plots for the same set of parameters. It appears that all the parameters have converged to their stationary distributions by iteration 7000.

Figure 10: Trace plots for a subset of parameters being monitored in the three-class mixture Rasch model for Block S20 (excluding Item 10)

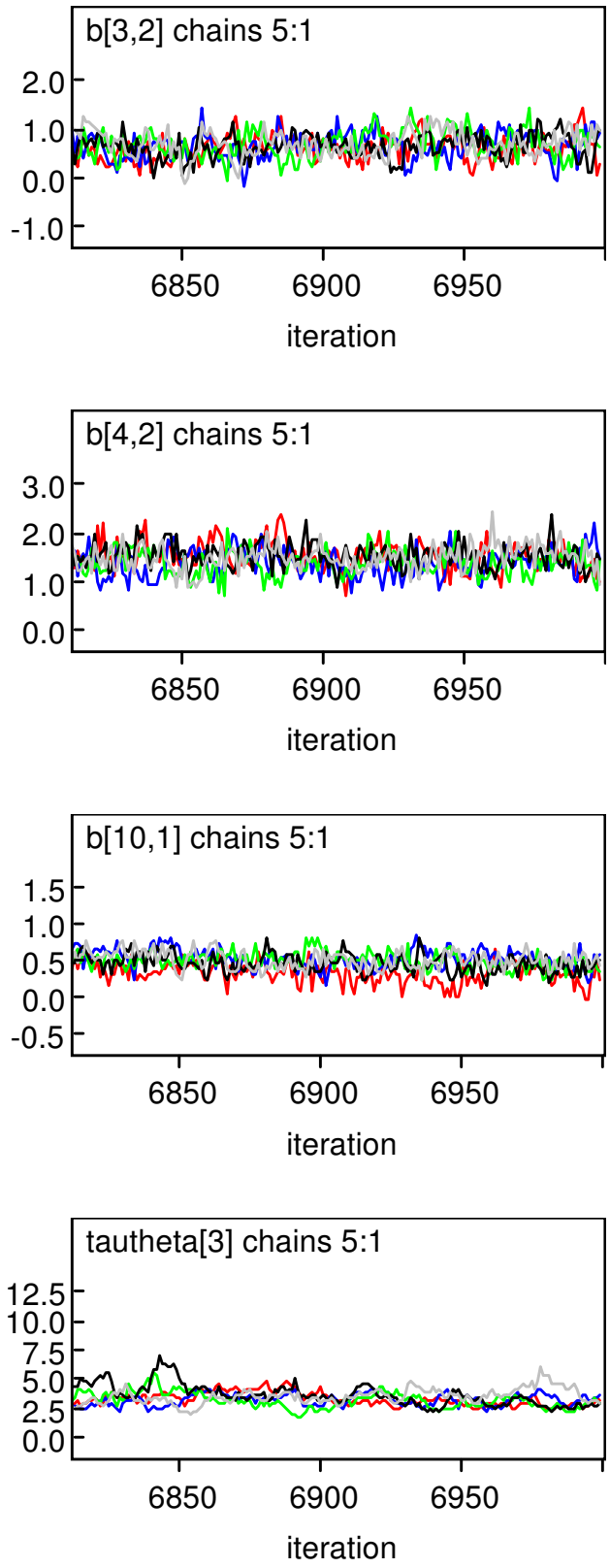


Figure 11: History plots for a subset of parameters being monitored in the three-class mixture Rasch model for Block S20 (excluding Item 10)

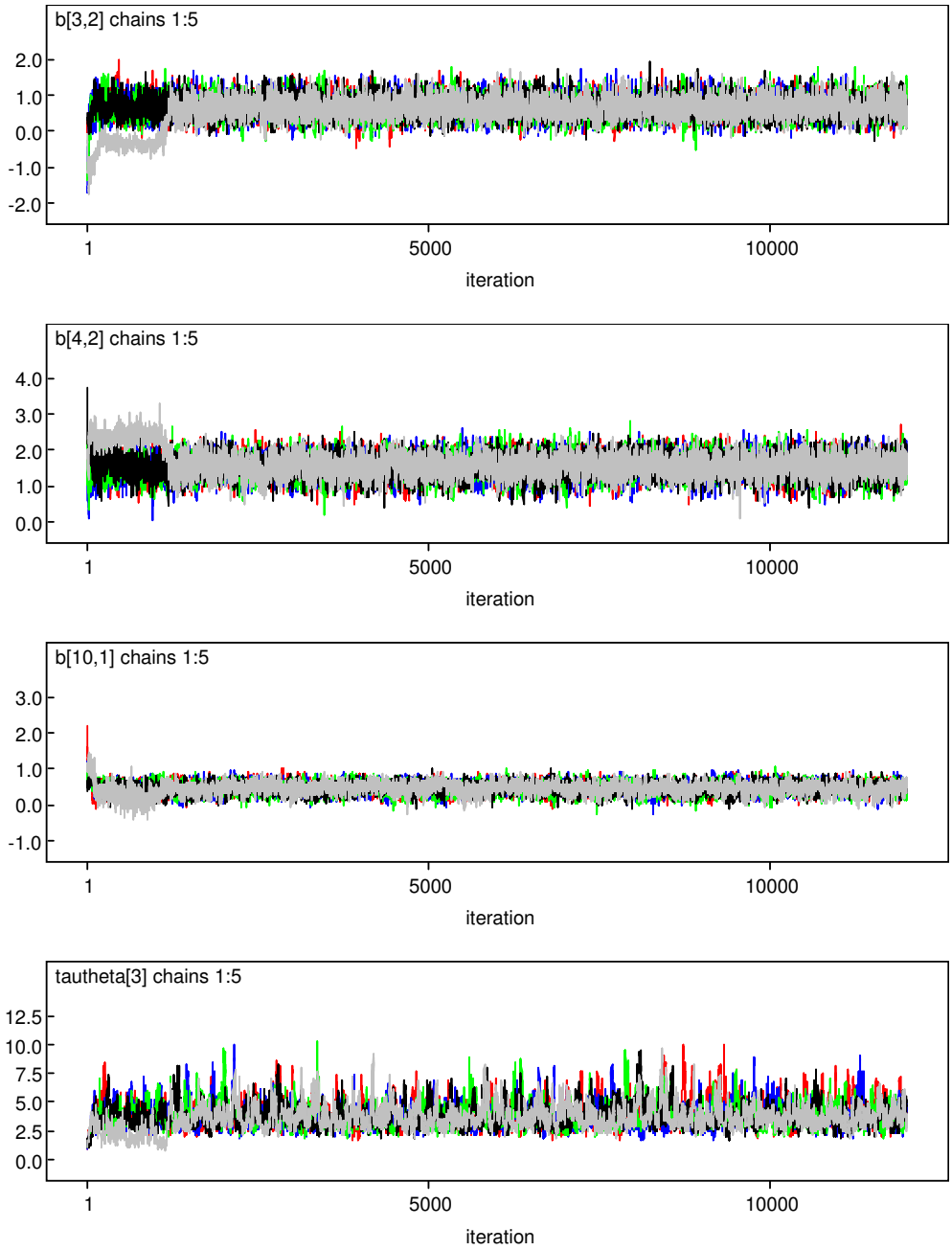


Table 11 summarizes the item difficulties obtained from the 3-class solution. A general pattern that can be inferred is that for every item except Items 4, 9, and 12, the estimated difficulty for Class 2 was greater than that for Class 3, which was, in turn, greater than that for Class 1. This indicates that Class 1 was made up of the most able students, Class 2 of the least able students, and Class 3 of students who stood in

between. Again, this finding suggests the adequacy of the unidimensional model, and the items that did not accord with the general pattern were the items that had the lowest loadings in the one-factor solution of the exploratory item factor analysis. The estimated proportion was .42 for Class 1, .13 for Class 2, and .45 for Class 3.

Table 11: Item difficulties of the 3-class mixture Rasch model for Block S20 (excluding Item 10)

Item	Content	Process	Class 1	Class 2	Class 3
S20_1	Life	Conceptual understanding	-1.321 (.1533)	.6429 (.3583)	-1.26 (.1709)
S20_2	Earth	Conceptual understanding	-1.426 (.2467)	.9829 (.2835)	.4189 (.1955)
S20_3	Physics	Conceptual understanding	-1.366 (.237)	.6769 (.2603)	.06474 (.1499)
S20_4	Earth	Conceptual understanding	2.354 (.1975)	1.517 (.2796)	2.965 (.3416)
S20_5	Earth	Conceptual understanding	.04985 (.266)	4.208 (.7819)	3.739 (.8724)
S20_6	Physics	Practical reasoning	.8849 (.1435)	4.176 (.8689)	1.519 (.1949)
S20_7	Physics	Practical reasoning	.046 (.1487)	3.896 (.8758)	1.141 (.1944)
S20_8	Earth	Practical reasoning	-2.011 (.1928)	1.67 (.6516)	-1.031 (.2007)
S20_9	Earth	Practical reasoning	.398 (.1453)	.7764 (.264)	1.222 (.166)
S20_11	Physics	Conceptual understanding	.4688 (.1433)	1.623 (.307)	1.001 (.1454)
S20_12	Earth	Conceptual understanding	-.3721 (.1417)	.2174 (.2489)	-.5226 (.1538)
S20_13	Life	Practical reasoning	-1.358 (.17)	1.291 (.4034)	-.7938 (.1617)
S20_14	Life	Practical reasoning	-.6725 (.1443)	2.163 (.6384)	-.2347 (.1555)
S20_15	Physics	Scientific investigation	-.2435 (.1917)	4.058 (.8854)	1.768 (.3635)
S20_16	Physics	Practical reasoning	.03917 (.141)	3.208 (.8635)	.5076 (.1678)

Note: Numbers in parentheses are standard deviations of the posterior distributions for the estimated parameters.

A comparison of item difficulties between each two of the three classes is helpful in distinguishing between students who belong to different classes. As displayed in Figure 12, in general, Class 1 was associated with smaller item difficulties than those for Class 2, except for Item 4 which, as discussed above, assessed a latent trait other than the science proficiency that was measured by the rest of the items. Items 5, 7, 8, and 15 were especially hard for students in Class 2. This is consistent with the two-class mixture solution.

Figure 12: Scatter plot of item difficulties between Class 1 and Class 2 of the three-class mixture Rasch model for Block S20 (excluding Item 10)

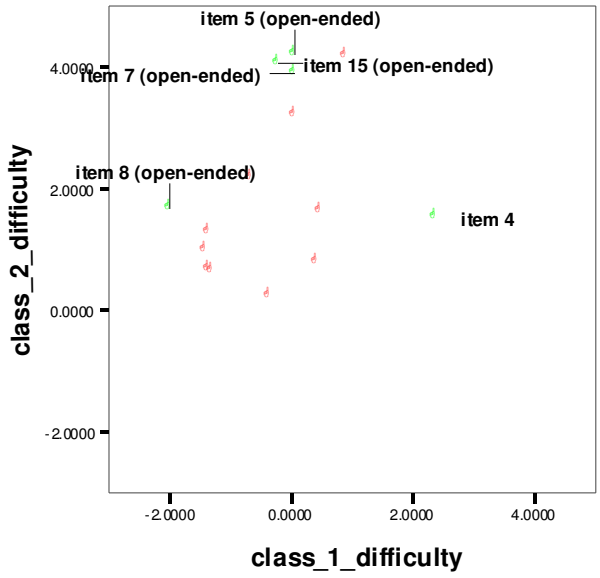
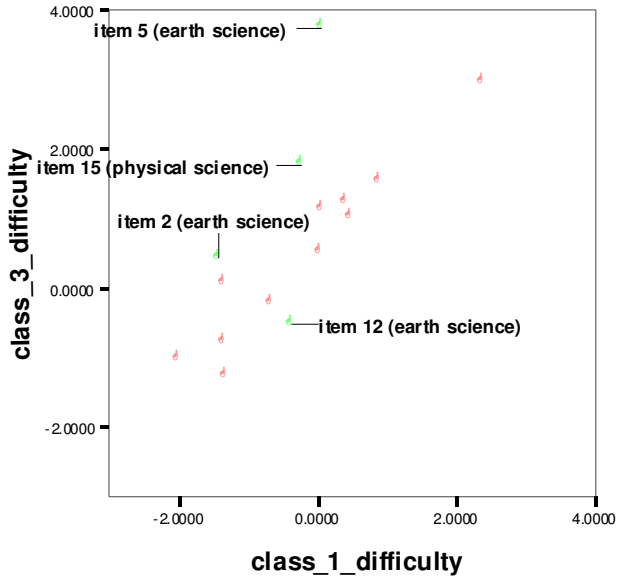


Figure 13 displays the differences in item difficulties between Class 1 and Class 3. In general, item difficulties for Class 1 were less than those for Class 3, except for Item 11. The differences in item difficulty were the largest for Items 2, 5, and 14,

which implies that students in Class 3 had the most difficulty in responding to these items. Items 2 and 5 were associated with earth science and Item 14 with life science. A tentative explanation is that student in Class 3 did not know much in earth science or life science.

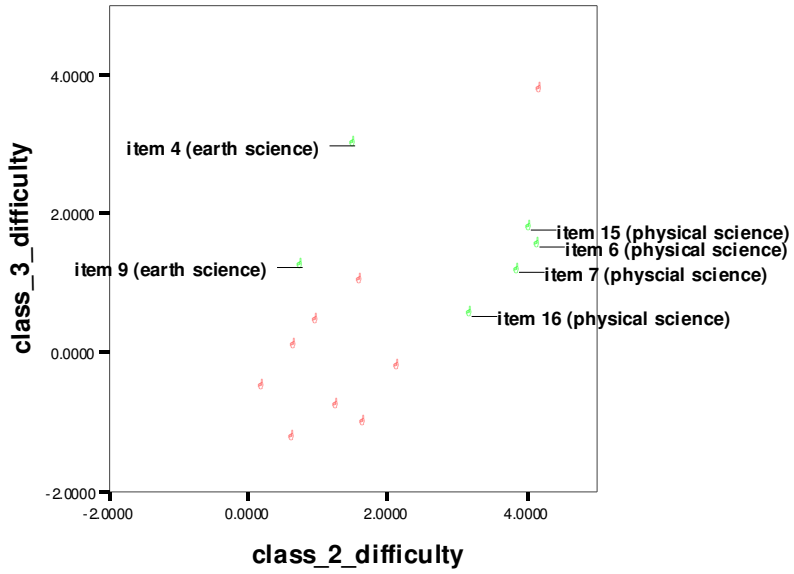
Figure 13: Scatter plot of item difficulties between Class 1 and class 3 of the three-class mixture Rasch model for Block S20 (excluding Item 10)



The items that distinguished between Class 2 and Class 3 would be very helpful in defining the characteristics of the two groups. Figure 14 displays the items that behaved differently between the two classes. Students in Class 3 were, generally speaking, more capable than students in Class 2. However, they did worse in Items 4 and 9, which both assessed knowledge and skills in earth science. This is consistent with the previous finding that students in Class 3 had particular difficulty in solving problems in earth science. On the other hand, Items 6, 7, 15, and 16, which were designed to cover physical science, were especially hard for students in Class 2.

This pattern suggests that Class 2 was made up of students who knew little in physical science while Class 3 was made up of students who found earth science difficult to comprehend.

Figure 14: Scatter plot of item difficulties between Class 2 and class 3 of the three-class mixture Rasch model for Block S20 (excluding Item 10)



A caveat for the interpretations of the latent classes is that they were based on only a few items and the common features of these items were identified in light of the given classification schemes, such as item format, content and cognitive classifications. It is possible that these items shared some other characteristics which made them behave differently across latent classes. However, those characteristics were unknown to us and we could not classify the items in other ways. In addition, the different performances on the identified items between the latent classes might result from some idiosyncratic features of the items rather than meaningful distinctions between the latent classes. Replication studies were

summarized in subsequent sections to cross-validate the results obtained thus far.

- Results in terms of substance

Results of the mixture Rasch models are not as clear-cut as those of the exploratory item factor analytic models. Interpreting latent classes on the basis of only a few items incurs the risk of overgeneralization. Discrepancies in item difficulty between latent classes cannot be unequivocally claimed to reflect the true discrepancies between the latent classes with regard to the aspect of science proficiency being measured. There may be some trivial factors associated with the items that make them behave differently across latent classes.

The three-class solution suggests differences among latent classes that seem to be systematic. Students of different classes exhibit strengths and weaknesses in different content areas. Variations among students in terms of content knowledge are what the assessment is designed to capture, through the application of a different model, though.

The two-class and three-class mixture Rasch models give rise to patterns of item difficulties that are closely connected with the pattern of loadings in the one-factor item factor analytic model. Items that have negligible loadings on the single factor are those that manifest the qualitative differences between the latent classes in the mixture model solutions. How the latent classes differ from each other in terms of the assessed construct and background characteristics is discussed in Section 5.1.4.

5.1.2 Comparison of model fit by information criteria

Three information criterion indices, AIC, CAIC, and BIC, were computed using

the method discussed in Chapter 4, for the models studied in this dissertation. The resulting fit statistics are summarized in Table 12. Judged from "the smaller, the better" rule, the mixture Rasch models fit the response data the best, the item factor models the worst, and the multidimensional between-item models in the middle. Each of the three information criteria points to a different model as the best-fitting model: the three-class mixture model is most favored by AIC, the between-item multidimensional model by CAIC, and the two-class mixture model by BIC. This result makes sense. AIC applies the least penalty on model complexity and tends to pick the more complex model. In contrast, BIC and CAIC apply more penalty on model complexity and tend to pick the simpler model. Therefore, AIC picked among the several models the most complex model, i.e., the 3-class mixture model, and BIC and CAIC picked simpler models. However, if the best-fitting model has to be selected among the models being studied, I would recommend the 2-class mixture model because it has the smallest value on BIC, which is considered the most reliable fit index for comparing mixture models both in terms of correctness and consistency, based on the simulation study by Li, Cohen, Kim, and Cho (2006).

Table 12: Comparison of fit statistics across and within types of models for Block S20 (excluding Item 10)

Models	-2ln(L)	*p*	AIC	CAIC	BIC
Multidimensional between-item model (dimensions defined by content areas)	21154.84	21	21196.84	21496.37	21304.61
Multidimensional between-item model (dimensions defined by cognitive factors)	21123.14	21	21165.14	**21464.67**	21272.91
1-factor item factor analytic model (unidimensional IRT model)	21613.82	30	21673.82	22101.72	21827.77
2-factor item factor analytic model	21537.94	44	21625.94	22253.53	21851.73
2-class mixture Rasch model	20956.62	33	21022.62	21493.31	**21191.97**
3-class mixture Rasch model	20849.14	50	**20949.14**	21662.31	21205.72

Notes: 1. *p* denotes the number of parameters to be estimated in the model and the sample size *N* is 1251 for each model.
2. Models in bold represent the best-fitting models by the different information criteria.

Information criteria take into consideration both degree of fit and model complexity when evaluating models. Based on the information criteria, the mixture Rasch models fit the data so much better than the other two types of models that even after they were penalized by the additional number of parameters, they still came out as the best fitting models. In other words, their superiority in goodness of fit far exceeded the penalties for model complexity. Compared with the other two types of models, the mixture Rasch models seemed to have done a better job in capturing the important features of the response patterns and produced more nuanced interpretations of the examinees' proficiencies.

5.1.3 Comparison of narrative stories for selected examinees

As discussed earlier, the three types of models being studied in the dissertation are based on different conceptions about the NAEP science assessment and fit the test data differently. More importantly, they tell different stories about the characteristics of the examinees. In this section, substantive stories inferred from the model results were compared between models, with both similarities and differences emphasized. Specifically, a small number of response patterns were studied to explore how the same responses elicited different stories through the lenses of different measurement models.

Based on the three sets of analysis results presented above, three models were selected from the three types of measurement models, each representing a different conceptualization about the assessment. Results of between-item multidimensional analysis indicated that the item block was multidimensional and the multiple dimensions were defined as the three cognitive skills: conceptual understanding, practical reasoning, and scientific investigation. Results of exploratory factor analysis suggested that the examinees' responses to the items in this block were governed by two factors, namely, the doing and knowing aspects of science learning. Mixture Rasch model analyses showed that the two- and three-class models exhibited essentially the same patterns, but the two-class model fit better than the three-class model. Thus, the between-item multidimensional model, the 2-factor model, and the 2-class mixture model were the models being compared in the comparison of narrative stories for the selected examinees.

Three examinees were selected, each being representative of a particular type of response patterns. Table 13 summarizes their responses, their latent class membership, their estimated θ values in the between-item multidimensional model, and their factor scores in the two-factor model. An examination of the three response patterns reveals the fact that the first examinee gave incorrect responses to almost all the items except for Item 4. By contrast, the second examinee gave correct responses to all but only a few items. The third examinee gave more correct responses than the first examinee, but the number of incorrect responses was non-negligible.

Table 13: Results of model analyses for the selected examinees for Block S20 (excluding Item 10)

ID	Response Pattern	Class	Between-item Multidimensional Model			2-factor Model	
			θ_1	θ_2	θ_3	f_1	f_2
966	000100000000000	1	-1.715	-2.037	-4.565	-1.869	-0.485
73	111011010011111	2	0.605	0.941	0.958	1.409	0.663
608	11100000000100M	2	-0.905	-1.116	-2.854	-0.682	-0.011

Note: M indicates a missing response.

Results of mixture model analysis show that the first examinee was assigned to Class 1 and its posterior probability of belonging to Class 1 was about 97%. The second examinee was assigned to Class 2 and its posterior probability of belonging to Class 2 was about 99.99%. Obviously, each of these two examinees was representative of a latent class, and their response patterns supported the finding that

Class 2 was composed of examinees who were generally more capable in science than those in Class 1. The third examinee had a response pattern that was not typical of either of the two classes, and was assigned to Class 2 with a posterior probability of 50.02%, a little bit greater than the threshold value of 50%. All the inferences about the third examinee were made on the basis of this class membership. Furthermore, the responses of the three examinees to Item 4 supported the conclusion that Item 4 carried all the evidence about the qualitative difference between the two classes. The first examinee performed better in Item 4 than the other two, although his (or her) performance in the rest of the items was inferior to that of the other two.

Between-item multidimensional analysis produced estimated θ values for each examinee along the three predefined dimensions. Apparently, the first examinee had smaller θ values than the second and third examinee. Another pattern that is easily observable is that the first examinee had lower θ values on the dimensions of practical reasoning and scientific investigation than on the dimension of conceptual understanding. The opposite pattern was observed in the θ values of the second examinee. Based on these θ values, a tentative conclusion can be drawn that the first examinee was especially weak in terms of reasoning and investigation skills, while the second examinee was more balanced in the development of the three kinds of cognitive skills. The pattern of the estimated θ values of the third examinee was similar with that of the first examinee, and a similar statement could be made about the third examinee.

The factor scores obtained from the 2-factor model support the conclusion from

the between-item multidimensional analysis. In the 2-factor model, Factor 1 was interpreted as representing what students can do in science and Factor 2 as representing what students know in science. Obviously, the first examinee had lower scores on Factor 1 than on Factor 2, while the factor scores of the second examinee displayed the opposite pattern. This finding indicates that the first examinee was even weaker in the doing aspect of science learning than in the knowing aspect, while the second examinee was good at both aspects and his (or her) proficiency in the doing aspect was even more advanced than in the knowing aspect of science learning. Again, the third examinee had similar factor scores with those of the first, and, therefore, similar conclusions could be drawn for the two examinees.

5.1.4 Background characteristics of latent classes

In this section, I extracted the demographic and background variables of the examinees and studied how the manifest examinee characteristics were associated with latent class membership. The 2-class mixture model solution was the solution interpreted here because it was identified by BIC as fitting the data better than the 3-class model. The model was estimated in WinBUGS through the MCMC algorithm, which was briefly described in Chapter 4. Basically, in the estimation, a class membership was sampled for each examinee at each iteration proportional to the probability of that examinee belonging to that class, which was conditional upon the item parameters, parameters for the mixing proportions, and abilities of all examinees. The frequency with which each examinee was sampled into each class defined the posterior probability of the examinee's membership in that class, and each examinee

was assigned to whichever latent class that had the highest posterior probability.

Based on the estimation method described above, of the 1251 examinees in the response data, 365 were classified to Class 1 and 886 to Class 2. This is consistent with the estimated class proportions (.31 and .69) reported earlier. As shown in Table 10, the item difficulties estimated for Class 1 are larger than those for Class 2, except for Item 4. As discussed earlier, mixture modeling characterizes person heterogeneity in terms of qualitative differences and quantitative differences. Item 4 carries all the evidence for the qualitative difference between the two classes, while the quantitative differences between the two classes reside in the general pattern of the estimated item difficulties.

An examination of the content of Item 4 provides an answer to the question about how the two classes differ qualitatively from each other. Item 4 asks about the two most common elements in the Earth's crust. Clearly, this item emphasizes the recall of knowledge rather than the more advanced cognitive abilities. However, the knowledge required by this item is unrealistically challenging for 8^{th} graders, according to the commentary by Li (2006). Li claims that even some college students who majored in science and engineering in a prestigious university failed to give a correct answer to this question. It seems that this particular item is actually assessing some peripheral or unscientific content for this sample of examinees. Therefore, the difference in performance on this item between the two latent classes should be attributed to the idiosyncratic features of the item and does not reflect any meaningful systematic difference between the two classes.

Quantitative differences between the two classes are evident if we compare the estimated item difficulties for the two classes. As mentioned above, the item difficulties indicate that generally speaking, Class 2 was a higher-achieving group in science than Class 1. Thirteen demographic and background variables were extracted for the sample and their associations with class membership were studied through significance tests. Table 14 summarizes the variables being studied and the results of the significance tests. Responses of the variables were dichotomized in a manner that fit the research questions of interest, and a z-test for proportions from two independent groups was performed for each variable (Please refer to Appendix B for the associations between the background variables and latent class classifications).

Table 14: Associations between background variables and latent class membership of the 2-class mixture Rasch model solution for Block S20 (excluding Item 10)

Variables	Results of significance tests
Gender	The proportion of males in Class 1 **is not smaller** than that in Class 2.
IEP	The proportion of IEP students in Class 1 **is larger** than that in Class 2.
LEP	The proportion of LEP students in Class 1 **is larger** than that in Class 2.
Race	The proportion of white students in Class 1 **is smaller** than that in Class 2.
Mother's education	The proportion of students whose mothers have high school education or beyond in Class 1 **is smaller** than that in Class 2.
Father's education	The proportion of students whose fathers have high school education or beyond in Class 1 **is smaller** than that in Class 2.
Like science	The proportion of students who like science in Class 1 **is smaller** than that in Class 2.
Good at science	The proportion of students who report themselves as good at science in Class 1 **is smaller** than that in Class 2.
Science is useful	The proportion of students who think science is useful in Class 1 **is smaller** than that in Class 2.
Science is hard	The proportion of students who think science is a hard subject in Class 1 **is not larger** than that in Class 2.
Learning science is memorization	The proportion of students who think learning science is memorization in Class 1 **is not larger** than that in Class 2.
Studying science in school	The proportion of students who study science every day in Class 1 **is smaller** than that in Class 2.
Time on homework	The proportion of students who spend 1 hour or more on science homework in Class 1 **is not smaller** than that in Class 2.

Based on the results of the statistical tests, it seems that Class 2, as compared with Class 1, was made up of more white students and fewer students with either individualized educational plans (IEP) or limited English proficiency (LEP); their parents had more education; they were interested in science, considered themselves as good at science, had strong motivation to study science, and had more opportunities to study science in school. On the other hand, Class 1 was not made up of more female students than Class 2. Besides, Class 1 was not more likely to think that science is a hard subject or that the right way to learn science is memorization, and they did not spend less time on science homework than Class 2.

5.2 Analysis results of block combination 1 (S7 and S4)

Of the four publicly released item blocks for Grade 8, S7 and S4 appeared together in two booklets. Similarly, Blocks S20 and S4 were bundled together in two booklets. In this section and the following section, analyses were performed on these two combinations of item blocks, which elicited responses from more examinees than other combinations of any two of the four released blocks, which appeared in only one booklet. Similar to the analyses of a single block, the three types of models were fit to the response data sequentially and information criteria were computed for each model. Results of analysis are organized under the heading of each model and discussed in terms of estimates of parameters and statistical fit, and substantive meaning.

S7 was a theme block and consisted of 2 multiple-choice and 10 open-ended items. These items covered content from earth science exclusively, and required

proficiencies of conceptual understanding, practical reasoning, and scientific investigation. S4 was made up of hands-on tasks, which required examinees to conduct an investigation using the provided equipment and answer questions related to the investigation. All the 9 items included in the block were open-ended questions. They covered content from physical and earth sciences and required skills from all the three cognitive domains. These two blocks appeared together in two booklets and elicited responses from a total of 419 examinees.

A preliminary exploratory factor analysis was performed on the data matrix, and the factor solutions suggested no odd-behaving items. Subsequently, all the 21 items and their responses were analyzed by the three types of models, following the procedures described in Chapter 4.

5.2.1 Results of model analyses

5.2.1.1. The between-item multidimensional model

In this study, the between-item multidimensional model of the MRCMLM family was used in a confirmatory mode to examine two hypothesized test structures. The first hypothesis states that the test is multidimensional and the latent dimensions can be defined in terms of content areas. In contrast, the second hypothesis says that the test is multidimensional and the dimensions can be defined in terms of cognitive domains. Under each hypothesis, each item was categorized as an indicator of one and only one latent dimension, and the parameter(s) of the item was estimated on that dimension. The association between the items and the hypothesized dimensions was provided in the 1996 NAEP science public release report.

- Results in terms of parameters and fit

As specified in the public release report, the two item blocks collectively covered physical science and earth science. Therefore, under the first hypothesis, the test is two-dimensional, and the latent dimensions correspond to physical and earth sciences. Table 15 summarizes the estimates of item difficulties for the 21 items, along with their standard errors and fit indices. The difficulty of the last item of each dimension was constrained, and standard errors or fit indices were not available. An examination of the diagnostic fit indices for the items suggests that Items 5, 7, 8, and 9 of Block S7 and Items 1 and 5 of Block S4 display modest misfit, and Items 2 and 8 of Block S4 exhibit serious misfit, according to the criteria discussed in Chapter 4.

Table 15: Item difficulty estimates and fit statistics from ConQuest for Blocks S7 and S4 (The dimensions are defined in terms of content areas)

Item	Dimension	Estimate	Standard Error	Unweighted Fit		Weighted Fit	
				MNSQ	T	MNSQ	T
S7_1	Earth	-1.211	0.083	0.96	-0.5	0.98	-0.3
S7_2	Earth	-0.049	0.081	0.93	-1.0	1.01	0.1
S7_3	Earth	-0.291	0.081	0.96	-0.6	1.01	0.2
S7_4	Earth	0.978	0.085	0.93	-1.0	1.00	0.1
S7_5	Earth	4.506	0.111	0.79	-3.3	0.87	**-2.2**
S7_6	Earth	-1.225	0.083	1.14	2.0	1.09	1.5
S7_7	Earth	-0.924	0.082	0.79	-3.3	0.83	**-3.0**
S7_8	Earth	-1.802	0.086	0.82	-2.8	0.84	**-2.8**
S7_9	Earth	-0.036	0.081	0.74	-4.1	0.84	**-2.7**
S7_10	Earth	1.167	0.086	1.00	0.0	1.10	1.5
S7_11	Earth	0.770	0.083	0.80	-3.1	0.90	-1.5
S7_12	Earth	1.628	0.089	0.85	-2.3	0.97	-0.4
S4_1	Earth	-0.675	0.081	1.33	4.3	1.21	**3.3**
S4_2	Physics	0.275	0.082	1.29	3.8	1.29	***4.2***
S4_3	Physics	-0.629	0.080	1.10	1.5	1.07	1.2
S4_4	Physics	0.354*					
S4_5	Earth	-2.773	0.094	0.87	-1.9	0.88	**-2.0**
S4_6	Earth	1.151	0.086	0.88	-1.9	0.94	-0.9
S4_7	Earth	-1.870	0.087	0.91	-1.4	0.89	-1.8
S4_8	Earth	0.249	0.082	0.61	-6.6	0.68	***-5.7***
S4_9	Earth	0.408*					

Notes: 1. * indicates that the item difficulty was constrained.

2. A weighted T statistic with an absolute value larger than 2 suggests moderate misfit. Items in bold are moderately misfit items. A weighted T statistic with an absolute value larger than 4 suggests serious misfit. Items in bold italic are seriously misfit items.

The deviance for the two-dimensional model was equal to 9186.015, and the number of estimated parameters was equal to 24. A unidimensional model was fit to the same response data and its deviance was equal to 9192.948 with 22 estimated parameters. The deviance difference between the two models was equal to 6.933 with 2 degrees of freedom. A significance test of the difference statistic suggests that the two-dimensional model fits significantly better than the one-dimensional model because the p-value ($p = .031$) of the statistic is smaller than the nominal level of .05.

Under the second hypothesis, the test is three-dimensional, and the latent dimensions correspond to conceptual understanding, practical reasoning, and scientific investigation. Table 16 summarizes the estimates of item difficulties along the three dimensions, their standard errors and fit indices. The fit indices for the items suggest that Items 1, 7, 8, and 9 of Block S7 and Item 4 of Block S4 are moderately misfit items and Item 3 of Block S4 is a seriously misfit item.

Table 16: Item difficulty estimates and fit statistics from ConQuest for Blocks S7 and S4 (The dimensions are defined in terms of science process skills)

Item	Dimension	Estimate	Standard Error	Unweighted Fit		Weighted Fit	
				MNSQ	T	MNSQ	T
S7_1	Conceptual understanding	-1.387	0.082	0.86	-2.1	0.86	**-2.4**
S7_2	Conceptual understanding	-0.229	0.080	0.89	-1.7	0.99	-0.1
S7_3	Conceptual understanding	-0.470	0.080	0.88	-1.8	0.93	-1.2
S7_4	Practical reasoning	-0.210	0.089	1.06	0.8	1.07	1.1
S7_5	Practical reasoning	3.221	0.123	0.95	-0.8	0.97	-0.5
S7_6	Scientific investigation	-0.657	0.083	1.14	1.9	1.10	1.6
S7_7	Conceptual understanding	-1.100	0.081	0.79	-3.2	0.85	**-2.6**
S7_8	Scientific investigation	-1.240	0.087	0.85	-2.2	0.84	**-2.6**
S7_9	Scientific investigation	0.541	0.081	0.80	-3.1	0.85	**-2.5**
S7_10	Conceptual understanding	0.981	0.085	0.86	-2.1	0.93	-1.1
S7_11	Conceptual understanding	0.585	0.083	0.81	-2.8	0.90	-1.6
S7_12	Conceptual understanding	1.440	0.088	0.81	-2.9	0.90	-1.5
S4_1	Conceptual understanding	-0.852	0.081	1.09	1.3	1.07	1.1
S4_2	Scientific investigation	1.153	0.083	0.83	-2.6	0.92	-1.2
S4_3	Scientific investigation	0.195	0.081	0.73	-4.4	0.74	***-4.5***
S4_4	Scientific investigation	1.238	0.083	1.29	3.8	1.25	**3.5**
S4_5	Scientific investigation	-2.219	0.094	0.90	-1.5	0.89	-1.8
S4_6	Conceptual understanding	0.965	0.085	0.87	-2.0	0.92	-1.2
S4_7	Practical reasoning	-3.011*					
S4_8	Conceptual understanding	0.067*					
S4_9	Scientific investigation	0.989*					

Notes: 1. * indicates that the item difficulty was constrained.

2. A weighted T statistic with an absolute value larger than 2 suggests moderate misfit. Items in bold are moderately misfit items. A weighted T statistic with an absolute value larger than 4 suggests serious misfit. Items in bold italic are seriously misfit items.

The deviance for the three-dimensional model was equal to 9190.062 with 27 estimated parameters. The deviance difference between this model and the unidimensional model was equal to 2.886 with 5 degrees of freedom. A significance test of the difference statistic suggests that the multidimensional model does not fit significantly better than the one-dimensional model because the p-value ($p = .718$) of the statistic is far greater than the nominal level of .05.

- Results in terms of substance

A comparison of the goodness-of-fit of the three models suggests that between-item multidimensionality in terms of cognitive skills does not provide a better description of the structure of the two item blocks than unidimensionality. In contrast, multidimensionality in terms of content areas makes some sense. This finding contradicts with the one found in the between-item multidimensional model analyses of Block S20. A possible reason for the occurrence of different findings is that items in Blocks S7 and S4, as compared with those in Block S20, require more of content knowledge than of cognitive proficiencies. For these items, correct responses rely more on the subject matter knowledge than on the cognitive proficiencies, and lack of subject matter knowledge is a bigger hindrance than insufficiency in cognitive abilities. As a result, examinees' differences in terms of cognitive abilities can not adequately account for their differential performances in these items, and the cognitive abilities do not come out as significant dimensions.

5.2.1.2 The exploratory item factor analytic model

Similar to what happened in the block-level analysis, one-, two-, three-, and

four-factor models were fit sequentially to the response data and estimated through TESTFACT. The model solutions were compared, and the best-fitting model was selected on the basis of the criteria discussed in Chapter 4.

- Results in terms of parameters and fit

The scree plot in Figure 15 shows the first ten eigenvalues of the tetrachoric correlation matrix for the 21 items. Obviously, the first factor explains a large proportion of variance among the items while the remaining factors, as compared to the first factor, are insignificant.

Figure 15: Scree plot from exploratory factor analysis for Blocks S7 and S4

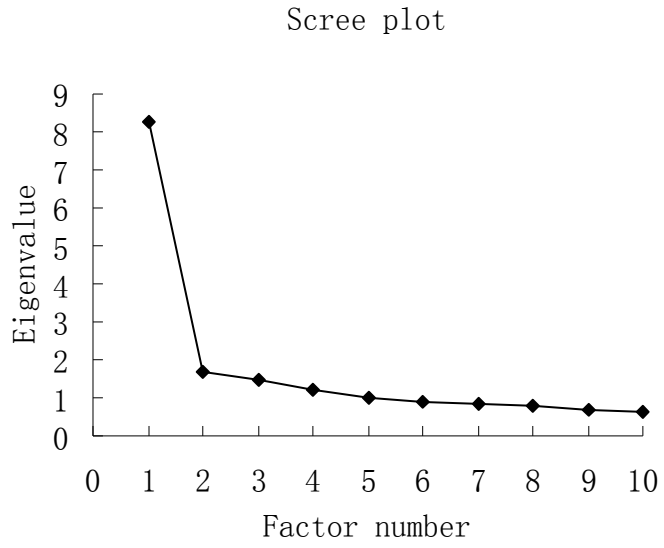


Table 17 summarizes the root mean square residuals (RMSRs) for the four model solutions and the percentage of reduction in the RMSR as an additional factor is added to the current model. Rules for using the RMSR index say that an RMSR less than .05 indicates an acceptable solution and that an additional factor can be added to the current model until the reduction in RMSR is less than 10%. According to these rules, it is safe to say that none of the four models fits the data acceptably well, but the

three-factor model is most preferred.

Table 17: Root mean square residuals (RMSRs) for the one-, two-, and three-factor models for Blocks S7 and S4

Model	RMSR	Percentage of reduction
1-factor model	0.092	
2-factor model	0.076	17%
3-factor model	0.064	14%
4-factor model	0.076	-19%

Table 18 summarizes the chi-square statistics for the four models and Table 19 shows the results of the chi-square difference tests. The results suggest that the four-factor model fits significantly better than the three-factor model, the three-factor model fits significantly better than the two-factor model, and the two-factor model fits significantly better than the one-factor model.

Table 18: Chi-square statistics for the one-, two-, three-, and four-factor models for Blocks S7 and S4

# of factors	Chi-square	DF	P-value
1	3933.36	376	0.000
2	3849.38	356	0.000
3	3815.29	337	0.000
4	3783.98	319	0.000

Table 19: Tests of chi-square difference statistics between factor models for Blocks S7 and S4

	Chi-square difference	DF difference	P-value
2-factor vs. 1-factor	83.98	20	0.000
3-factor vs. 2-factor	34.09	19	0.018
4-factor vs. 3-factor	31.31	18	0.026

Factor loadings are another criterion in judging the meaningfulness of a factor solution. In the one-factor solution, almost all of the items loaded substantially on the factor. This indicates the existence of a general factor that explains a considerable amount of variance of the items.

In the two-factor model, the factor loadings exhibited a pattern close to a "simple structure". Table 20 summarizes the factor loadings of the two-factor solution. Clearly, factor 1 is strongly indicated by the first twelve items, i.e., items of Block S7, and factor 2 is strongly indicated by the last nine items, i.e., items of Block S4, except for only a few exceptions. S7 is a theme-based block, and a close investigation of the items in S7 reveals that they are all questions about the Solar System. S4 is a block of hands-on tasks, and all the items in S4 are related to "salt solutions". Therefore, the two factors are ability and proficiency with regard to the two sets of items. The first factor can be interpreted as representing what students know about the Solar System and how they use the knowledge to understand natural phenomena, and the second factor can be interpreted as indicating students' skills in carrying out investigations with salt solutions and their abilities in explaining the outcomes.

Table 20: Factor loadings of the two-factor model for Blocks S7 and S4

Item No.	Content	Process	Factor 1	Factor 2
S7_1	Earth	Conceptual understanding	**.738**	-.125
S7_2	Earth	Conceptual understanding	**.576**	.024
S7_3	Earth	Conceptual understanding	**.673**	-.028
S7_4	Earth	Practical reasoning	**.516**	.069
S7_5	Earth	Practical reasoning	**.389**	**.309**
S7_6	Earth	Scientific investigation	**.385**	.073
S7_7	Earth	Conceptual understanding	**.802**	-.043
S7_8	Earth	Scientific investigation	**.627**	.177
S7_9	Earth	Scientific investigation	**.426**	**.396**
S7_10	Earth	Conceptual understanding	**.458**	.071
S7_11	Earth	Conceptual understanding	**.611**	.092
S7_12	Earth	Conceptual understanding	**.824**	-.165
S4_1	Earth	Conceptual understanding	.057	.249
S4_2	Physics	Scientific investigation	.022	**.676**
S4_3	Physics	Scientific investigation	.144	**.682**
S4_4	Physics	Scientific investigation	.104	.120
S4_5	Earth	Scientific investigation	**-.314**	**.937**
S4_6	Earth	Conceptual understanding	.188	**.425**
S4_7	Earth	Practical reasoning	-.046	**.623**
S4_8	Earth	Conceptual understanding	.286	**.632**
S4_9	Earth	Scientific investigation	.231	**.612**

Note: Numbers in bold are substantial loadings (>.30) on the two factors.

The pattern of factor loadings in the three-factor model or the four-factor model did not make more sense than that in the two-factor model. Therefore, the two-factor model is retained as the model that best explains the factor structure of the data.

- Results in terms of substance

The clear dichotomization in the factor structure of the two-factor solution reveals the specificity of the kind of proficiency measured by each of the two blocks. Apparently, the two blocks address very different topics. S7 is a theme-based block and requires much knowledge about the Solar System. In contrast, S4 is a hands-on task block and all the items revolve around the topic of salt solutions. Besides, as stated earlier, each block emphasizes more on content knowledge, which is subject-specific, than on cognitive abilities, which cross the boundaries of content areas. Consequently, the commonality that underlies the two blocks is insignificant as compared with their distinction.

5.2.1.3 The mixture Rasch model

Two- and three-class mixture Rasch models were fit sequentially to the response data for Blocks S7 and S4 in WinBUGS. In each model, five chains with over-dispersed starting values of class proportions were run. Similar to what happened in the analysis of Block S20, the problem of label switching was dealt with by imposing constraints on class membership for a few examinees. Specifically, a small number of examinees were pre-assigned to each latent class in each model. For each model, a total of 10,000 iterations were simulated for each of the five chains. The first 5000 iterations were discarded as burn-ins, and the remaining 5000 iterations were sampled for each chain. Thus, posterior estimates of the model parameters were calculated from a total of 25,000 iterations.

- Results in terms of parameters and fit

The two-class mixture Rasch model solution

In estimating the two-class mixture model, ten examinees were assigned to each latent class with certainty to solve the problem of label switching. Figure 16 shows the history plots for a subset of model parameters being monitored. Clearly, all the five chains converged quickly to the posterior distribution in each plot.

Figure 16: History plots for a subset of parameters being monitored in the two-class mixture Rasch model for Blocks S7 and S4

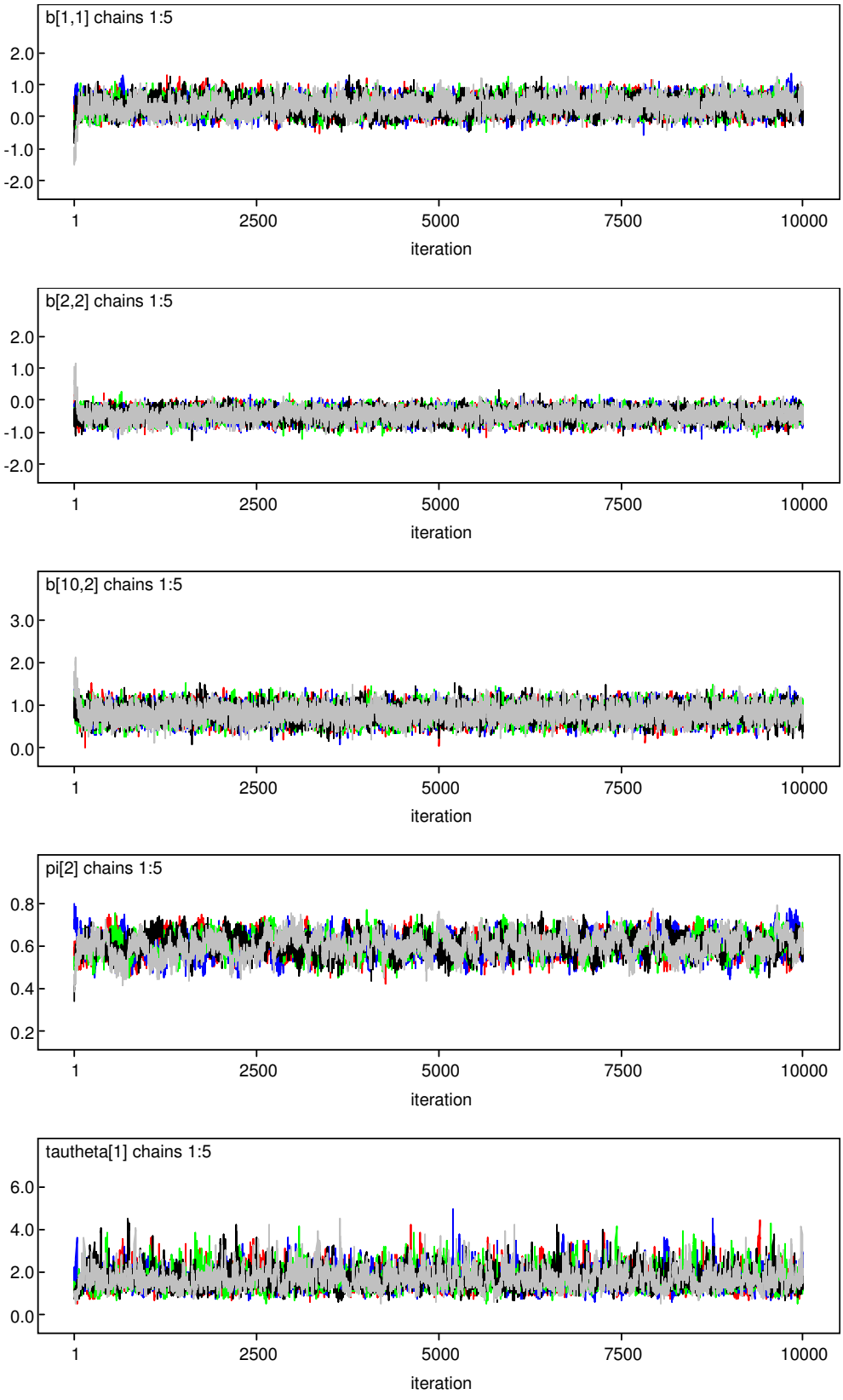


Table 21 summarizes the estimated difficulties of the two-class mixture solution. A general pattern observed is that the item difficulties estimated for Class 1 were greater than those for Class 2, except for Item 4 of Block S4. This means that Class 2 was made up of students who were generally more capable than students in Class 1, given their performance on the two blocks of items. This finding is essentially the same as the one found in the mixture model analysis of Block S20. In addition to item difficulties, posterior estimates of examinees' class membership were obtained. Since each examinee was tested on more items, class membership was estimated with greater confidence.

Table 21: Item difficulties of the 2-class mixture Rasch model for Blocks S7 and S4

Item	Content	Process	Class 1	Class 2
S7_1	Earth	Conceptual understanding	.3486	-1.82
			(.2249)	(.2342)
S7_2	Earth	Conceptual understanding	1.276	-.4551
			(.2476)	(.1735)
S7_3	Earth	Conceptual understanding	1.267	-.8185
			(.2883)	(.1769)
S7_4	Earth	Practical reasoning	2.403	.5043
			(.3679)	(.1714)
S7_5	Earth	Practical reasoning	5.18	3.838
			(.8868)	(.4122)
S7_6	Earth	Scientific investigation	-.2677	-1.238
			(.2076)	(.1843)
S7_7	Earth	Conceptual understanding	1.011	-1.862
			(.2928)	(.2591)
S7_8	Earth	Scientific investigation	-.08091	-2.856
			(.221)	(.4033)
S7_9	Earth	Scientific investigation	1.953	-.7543
			(.3372)	(205)
S7_10	Earth	Conceptual understanding	2.173	.8051
			(.3508)	(.1714)
S7_11	Earth	Conceptual understanding	2.793	.1256
			(.5184)	(.1794)
S7_12	Earth	Conceptual understanding	3.99	.937
			(.7888)	(.1888)
S4_1	Earth	Conceptual understanding	-.3023	-.418
			(.2219)	(.1677)
S4_2	Physics	Scientific investigation	2.198	.01392
			(.3344)	(.1814)
S4_3	Physics	Scientific investigation	1.431	-1.131
			(.2663)	(.2339)
S4_4	Physics	Scientific investigation	.7312	.8134
			(.2147)	(.1742)
S4_5	Earth	Scientific investigation	-1.708	-3.16
			(.259)	(.3544)
S4_6	Earth	Conceptual understanding	2.108	.7578
			(.3133)	(.1764)
S4_7	Earth	Practical reasoning	-1.098	-2.046
			(.2388)	(.2456)
S4_8	Earth	Conceptual understanding	2.585	-.6873
			(.5165)	(.2112)

S4_9	Earth	Scientific investigation	2.183 (.4276)	-.5978 (.2018)

Note: Numbers in parentheses are standard deviations of the posterior distributions for the estimated parameters.

The three-class mixture Rasch model solution

Similarly, the problem of label switching was fixed by preassigning examinees to latent classes. In estimating the three-class mixture model, ten examinees were preassigned to Class 1, ten to Class 3, and five to Class 2. This was so because only five examinees could be classified to Class 2 with high degrees of certainty based on the results of the preliminary run. Figure 17 shows the history plots for a small subset of model parameters being monitored. In each plot, all the five chains converged to the posterior distribution after the completion of a small number of iterations.

Figure 17: History plots for a subset of parameters being monitored in the three-class mixture Rasch model for Blocks S7 and S4

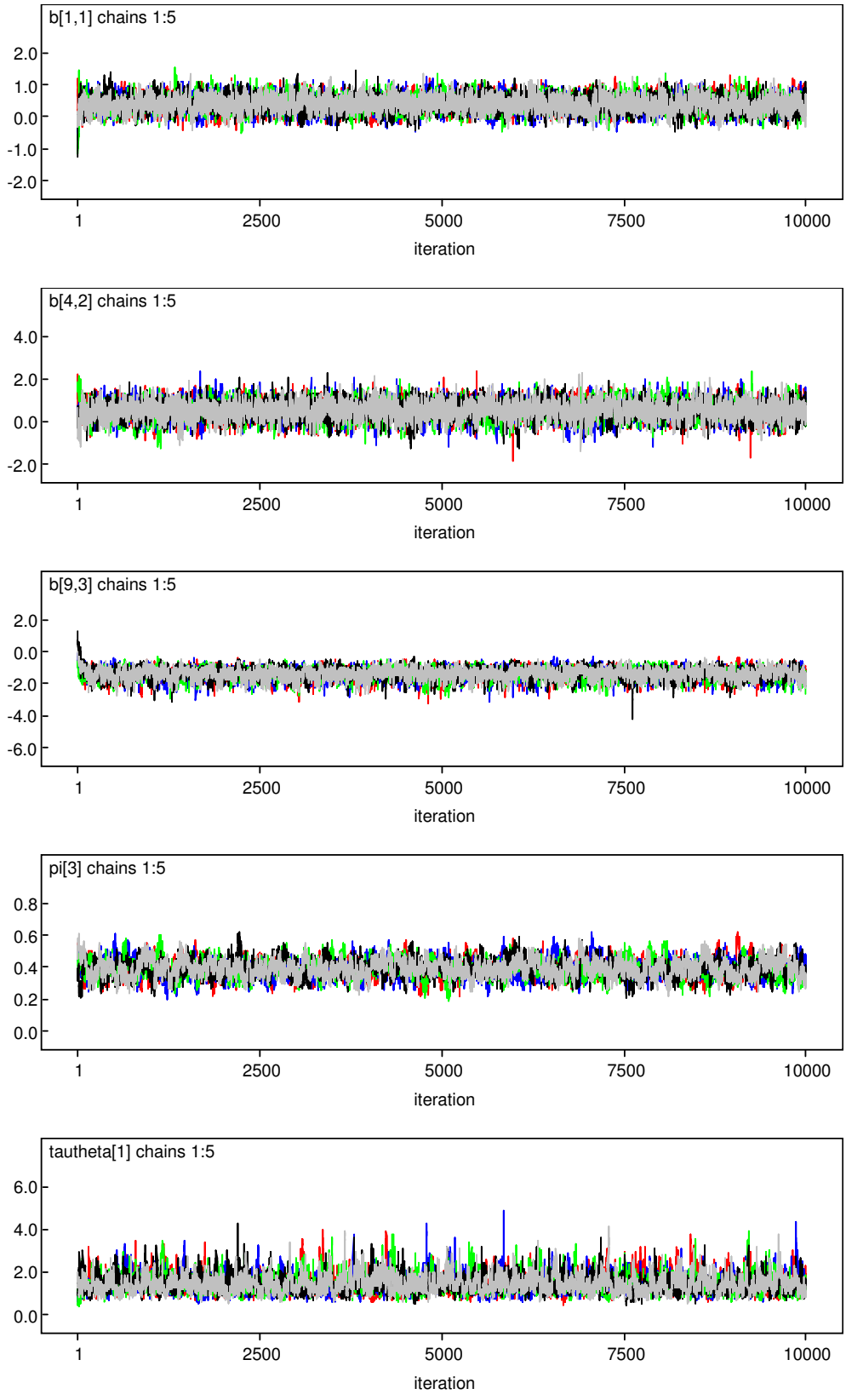


Table 22 summarizes the estimated item difficulties of the three-class mixture solution. For all the items in Block S7 and a couple of items in Block S4, Class 1 had the highest item difficulties among all the three classes. Class 2 had the highest item difficulties on most of the items in Block S4. This finding bears some resemblance to the one from the exploratory factor analysis, in which the items in S7 and items in S4 represent two different factors.

Table 22: Item difficulties of the 3-class mixture Rasch model for Blocks S7 and S4

Item	Content	Process	Class 1	Class 2	Class 3
S7_1	Earth	Conceptual understanding	.3853	-2.687	-1.769
			(.2321)	(.9715)	(.2914)
S7_2	Earth	Conceptual understanding	1.285	-.7737	-.3953
			(.247)	(.5178)	(.2229)
S7_3	Earth	Conceptual understanding	1.295	-1.258	-.7311
			(.2876)	(.5041)	(.2279)
S7_4	Earth	Practical reasoning	2.398	.5139	.4552
			(.3484)	(.3835)	(.2272)
S7_5	Earth	Practical reasoning	5.028	4.129	3.714
			(.85)	(.9381)	(.5392)
S7_6	Earth	Scientific investigation	-.3096	-.9012	-1.416
			(.2104)	(.4108)	(.2682)
S7_7	Earth	Conceptual understanding	1.036	-2.249	-1.934
			(.2994)	(.7299)	(.3187)
S7_8	Earth	Scientific investigation	-.1897	-1.995	-3.494
			(.2241)	(.6024)	(.6243)
S7_9	Earth	Scientific investigation	1.753	.4039	-1.406
			(.3217)	(.3953)	(.3413)
S7_10	Earth	Conceptual understanding	2.083	.8941	.7904
			(.3422)	(.4068)	(.2279)
S7_11	Earth	Conceptual understanding	2.718	.1528	.05573
			(.4488)	(.4381)	(.229)
S7_12	Earth	Conceptual understanding	4.05	.7667	.973
			(.8218)	(.4502)	(.2534)
S4_1	Earth	Conceptual understanding	-.3551	.2294	-.6792
			(.2183)	(.4241)	(.2302)
S4_2	Physics	Scientific investigation	1.991	2.111	-.7644
			(.311)	(.7833)	(.3163)
S4_3	Physics	Scientific investigation	1.308	.205	-2.16
			(.2579)	(.4423)	(.6005)
S4_4	Physics	Scientific investigation	.7553	.501	.9643
			(.2131)	(.4546)	(.2576)
S4_5	Earth	Scientific investigation	-1.858	-1.778	-4.513
			(.2671)	(.469)	(.8818)
S4_6	Earth	Conceptual understanding	2.035	1.698	.4242
			(.3103)	(.5669)	(.2324)
S4_7	Earth	Practical reasoning	-1.2	-.8961	-2.836
			(.2409)	(.4617)	(.4752)
S4_8	Earth	Conceptual understanding	2.273	.5105	-1.319
			(.4308)	(.4639)	(.333)

S4_9	Earth	Scientific investigation	2.07 (.3935)	.4861 (.5308)	-1.167 (.3169)

Note: Numbers in parentheses are standard deviations of the posterior distributions for the estimated parameters.

- Results in terms of substance

Results of mixture Rasch models suggest that all the items, except for only a few, in the two item blocks were consistently harder for a class of examinees and easier for the other classes. This implies that a unidimensional model may be adequate in describing examinees' performances in the assessment. The exceptional items carry the evidence about the qualitative differences between the classes of examinees. The quantitative differences between the latent classes, on the other hand, are reflected in the relative magnitude of the estimated item difficulties.

Again, the items that were harder for the "higher-achieving" class but easier for the "lower-achieving" class were the items that had insignificant loadings in the one-factor model solution. This implies that these items were only remotely related to the factor, while all the other items were strong indicators of that factor. A plausible explanation is that the majority of items were assessing the science proficiency that was designed to be assessed, but a handful of items were poorly written that they were actually assessing a peripheral or nonscientific construct.

5.2.2 Comparison of model fit by information criteria

Table 23 summarizes the fit statistics for the models applied in the above analyses. Generally speaking, the mixture Rasch models fit the response data the best, the item factor models the worst, and the multidimensional between-item models in the middle.

This is consistent with the results shown in Table 12. Each of the three information criteria points to a different model as the best-fitting model: the three-class mixture model is most favored by AIC, the between-item multidimensional model by CAIC, and the two-class mixture model by BIC. This, again, agrees with the finding in Table 12.

Table 23: Comparison of fit statistics across and within types of models for Blocks S7 and S4

Models	-2ln(L)	*p*	AIC	CAIC	BIC
Multidimensional between-item model (dimensions defined by content areas)	8972.764	24	9020.764	**9310.582**	9117.673
Multidimensional between-item model (dimensions defined by cognitive factors)	8972.552	27	9026.552	9352.597	9135.575
1-factor item factor analytic model (unidimensional IRT model)	9092.656	42	9176.656	9683.837	9346.247
2-factor item factor analytic model	9137.572	62	9261.572	10010.268	9511.920
2-class mixture Rasch model	8793.726	45	8883.726	9427.134	**9065.430**
3-class mixture Rasch model	8724.194	68	**8860.194**	9681.344	9134.769

Notes: 1. *p* denotes the number of parameters to be estimated in the model and the sample size *N* is 419 for each model.
2. Models in bold represent the best-fitting models by the different information criteria.

5.3 Analysis results of block combination 2 (S20 and S4)

As mentioned before, Blocks S20 and S4 were bundled together in two booklets and elicited more responses than other combinations of item blocks that only appeared once together. In this section, analyses were performed on this combination of item blocks. The same procedure was followed in the analyses: the three types of models

were fit to the response data sequentially and information criteria were computed for each model.

S20 was a regular paper-and-pencil block and consisted of 8 multiple-choice and 8 constructed-response items. These items covered content from physical science, earth science, and life science, and required proficiencies of conceptual understanding, practical reasoning, and scientific investigation. S4 was a hands-on task block. All the 9 items included in the block were open-ended questions. Collectively, they covered content from physical and earth sciences and required skills from all the three cognitive domains. These two blocks elicited responses from a total of 420 examinees. It should be noted that the examinees being studied in this section are a subsample of the examinees studied in Section 5.1.

A preliminary exploratory factor analysis was performed on the response data, and the factor model solutions suggested that Item 10 from block S20 again behaved oddly. It was dropped in the analyses thereafter. In total, there were 24 items and 420 examinees in the response data analyzed in this section.

5.3.1 Results of model analyses

5.3.1.1. The between-item multidimensional model

Again, the between-item multidimensional model of the MRCMLM family was used to test two hypotheses. The first hypothesis states that the test is multidimensional and the latent dimensions are defined in terms of content areas. The second hypothesis says that the test is multidimensional and the dimensions are defined in terms of cognitive domains. Under each hypothesis, each item was

categorized as an indicator of one and only one latent dimension, and the parameter(s) of the item was estimated on that dimension. Information about which dimension each item was written to assess was provided in the 1996 NAEP science public release report.

- Results in terms of parameters and fit

Under the first hypothesis, the test is three-dimensional and the three dimensions correspond to the three fields of science. Table 24 summarizes the estimates of item difficulties for the 24 items, along with their standard errors and diagnostic fit indices. Again, the difficulty of the last item of each dimension was constrained, and standard errors or fit indices were not available. The fit indices for the items suggest that Items 3, 7, 8 and 15 of Block S20 and Items 2 of Block S4 display modest misfit, and Items 5 of Block S20 and Item 8 of Block S4 exhibit serious misfit.

Table 24: Item difficulty estimates and fit statistics from ConQuest for Blocks S20 (excluding Item 10) and S4 (The dimensions are defined in terms of content areas)

Item	Content	Estimate	Standard Error	Unweighted Fit		Weighted Fit	
				MNSQ	T	MNSQ	T
S20_1	Life	-0.449	0.080	1.02	0.2	0.99	-0.2
S20_2	Earth	-0.520	0.081	0.88	-1.7	0.89	-1.9
S20_3	Physics	-1.226	0.081	1.11	1.6	1.12	**2.1**
S20_4	Earth	2.555	0.101	1.07	1.0	0.93	-0.8
S20_5	Earth	1.226	0.088	0.66	-5.6	0.72	***-4.5***
S20_6	Physics	0.673	0.086	1.09	1.2	1.06	0.9
S20_7	Physics	0.227	0.083	1.16	2.3	1.17	**2.5**
S20_8	Earth	-1.372	0.085	0.86	-2.1	0.87	**-2.3**
S20_9	Earth	0.538	0.083	1.16	2.2	1.03	0.5
S20_11	Physics	0.174	0.082	1.19	2.6	1.08	1.2
S20_12	Earth	-0.640	0.081	1.09	1.2	1.06	1.0
S20_13	Life	-0.130	0.079	1.01	0.1	1.02	0.3
S20_14	Life	0.579*					
S20_15	Physics	0.507	0.084	1.13	1.8	1.17	**2.5**
S20_16	Physics	0.214	0.083	1.11	1.6	1.09	1.4
S4_1	Earth	-0.377	0.081	0.91	-1.4	0.90	-1.6
S4_2	Physics	-0.159	0.081	1.17	2.3	1.17	**2.6**
S4_3	Physics	-0.747	0.080	0.99	-0.2	1.01	0.2
S4_4	Physics	0.336*					
S4_5	Earth	-2.552	0.095	0.88	-1.9	0.88	-1.8
S4_6	Earth	1.178	0.087	0.94	-0.9	0.98	-0.3
S4_7	Earth	-1.575	0.086	0.89	-1.6	0.89	-1.8
S4_8	Earth	0.590	0.083	0.51	-8.6	0.60	***-7.2***
S4_9	Earth	0.949*					

Notes: 1. * indicates that the item difficulty was constrained.

2. A weighted T statistic with an absolute value larger than 2 suggests moderate misfit. Items in bold are moderately misfit items. A weighted T statistic with an absolute value larger than 4 suggests serious misfit. Items in bold italic are seriously misfit items.

The deviance for the three-dimensional model was equal to 11134.302, and the number of estimated parameters was equal to 30. A unidimensional model was fit to the same response data and its deviance was equal to 11139.051 with 25 estimated parameters. The deviance difference between the two models was equal to 4.749 with 5 degrees of freedom. A significance test of the difference statistic suggests that the three-dimensional model does not fit significantly better than the one-dimensional model because the p-value ($p = .447$) of the statistic is greater than the nominal level of .05.

Under the second hypothesis, the test is three-dimensional, and the latent dimensions correspond to the three cognitive abilities. Table 25 summarizes the estimates of item difficulties along the three dimensions, their standard errors and fit indices. The fit indices for the items suggest that Items 2, 5, and 15 of Block S20 and Item 4 of Block S4 are moderately misfit items and Item 3 of Block S4 is a seriously misfit item.

Table 25: Item difficulty estimates and fit statistics from ConQuest for Blocks S20 (excluding Item 10) and S4 (The dimensions are defined in terms of cognitive domains)

Item	Content	Estimate	Standard	Unweighted Fit		Weighted Fit	
			Error	MNSQ	T	MNSQ	T
S20_1	Conceptual understanding	-1.523	0.081	0.94	-0.9	0.94	-1.0
S20_2	Conceptual understanding	-0.752	0.078	0.84	-2.4	0.86	**-2.6**
S20_3	Conceptual understanding	-0.986	0.078	0.94	-0.9	0.94	-1.1
S20_4	Conceptual understanding	2.228	0.096	0.94	-0.9	0.93	-1.0
S20_5	Conceptual understanding	0.935	0.084	0.72	-4.4	0.81	**-3.0**
S20_6	Practical reasoning	1.281	0.087	1.01	0.2	1.00	-0.0
S20_7	Practical reasoning	0.838	0.084	1.01	0.2	1.00	0.0
S20_8	Practical reasoning	-1.217	0.085	1.06	0.9	1.12	1.8
S20_9	Practical reasoning	0.680	0.083	1.04	0.6	1.03	0.5
S20_11	Conceptual understanding	0.371	0.080	0.95	-0.7	0.99	-0.2
S20_12	Conceptual understanding	-0.868	0.078	1.08	1.1	1.09	1.5
S20_13	Practical reasoning	-0.845	0.083	1.02	0.4	1.07	1.1
S20_14	Practical reasoning	-0.143	0.081	1.04	0.5	1.04	0.7
S20_15	Scientific investigation	1.047	0.090	0.74	-4.2	0.84	**-2.3**
S20_16	Practical reasoning	0.824	0.084	1.08	1.1	1.08	1.3
S4_1	Conceptual understanding	-0.614	0.078	0.85	-2.2	0.89	-1.9
S4_2	Scientific investigation	0.306	0.086	0.97	-0.4	0.97	-0.5
S4_3	Scientific investigation	-0.351	0.085	0.67	-5.4	0.75	***-4.3***
S4_4	Scientific investigation	0.857	0.089	1.26	3.5	1.20	**2.8**
S4_5	Scientific investigation	-2.873	0.100	0.95	-0.7	0.95	-0.7
S4_6	Conceptual understanding	0.888	0.083	0.85	-2.2	0.91	-1.4
S4_7	Practical reasoning	-1.419*					
S4_8	Conceptual understanding	0.320*					
S4_9	Scientific investigation	1.014*					

Notes: 1. * indicates that the item difficulty was constrained.
2. A weighted T statistic with an absolute value larger than 2 suggests moderate misfit. Items in bold are moderately misfit items. A weighted T statistic with an absolute value larger than 4 suggests serious misfit. Items in bold italic are seriously misfit items.

The deviance for the three-dimensional model was equal to 11103.413 with 30 estimated parameters. The deviance difference between this model and the unidimensional model was equal to 35.638 with 5 degrees of freedom. A significance test of the difference statistic suggests that the multidimensional model fits significantly better than the one-dimensional model because the p-value ($p = .000$) of the statistic is smaller than the nominal level of .05.

- Results in terms of substance

The multidimensional model analyses suggest that multidimensionality in terms of cognitive factors makes more sense than multidimensionality in terms of content areas. This is consistent with the result found in the analyses of Block S20, but different from that found in the analyses of Blocks S7 and S4.

Whether the meaningful multiple dimensions correspond to content areas or cognitive factors reflects the relative demands of the two kinds of proficiencies in the items, which are largely decided at the item writing stage. If most of the items require more of the recall of knowledge than of the application of more advanced science process skills, content areas will come out as significant dimensions. On the contrary, if a majority of the items have high demands for science process skills but low or minimal demands for content knowledge, science process skills will end up as significant dimensions.

5.3.1.2 The exploratory item factor analytic model

To find the factor structure that best describes the response data, one-, two-, three-, and four-factor models were fit sequentially to the data and estimated through TESTFACT. The model solutions were compared, and the best-fitting model was selected on the basis of the criteria discussed in Chapter 4.

- Results in terms of parameters and fit

Figure 18 is the scree plot that shows the first ten eigenvalues of the tetrachoric correlation matrix for the 24 items. Similar to what was found in the other scree plots, the first factor explains a large proportion of variance among the items while the remaining factors, as compared to the first factor, are insignificant.

Figure 18: Scree plot from exploratory factor analysis for Block S20 (excluding Item 10) and S4

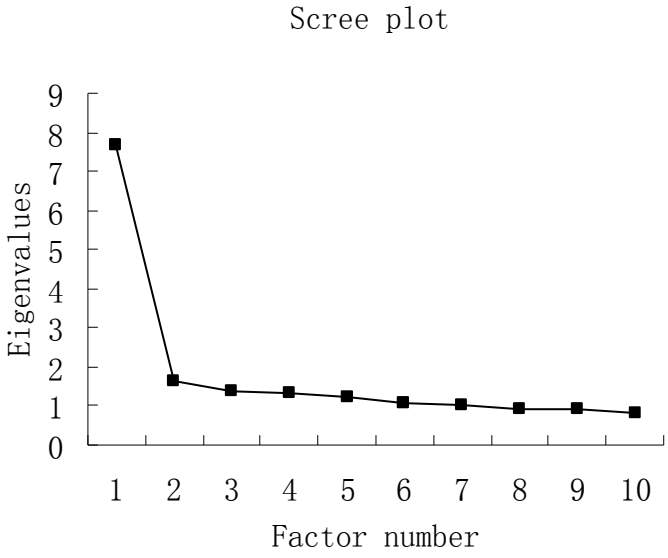


Table 26 summarizes the root mean square residuals (RMSRs) for the four model solutions and the percentage of reduction in the RMSR as an additional factor is added to the current model. Comparing the RMSRs of the four models to the threshold value of .05, it is safe to say that none of them fits the data acceptably well. The percentage of reduction in the RMSR suggests that the two-factor model is most preferred.

Table 26: Root mean square residuals (RMSRs) for the one-, two-, and three-factor models for Blocks S20 (excluding Item 10) and S4

Model	RMSR	Percentage of reduction
1-factor model	0. 080	
2-factor model	0.071	11%
3-factor model	0.065	8%
4-factor model	0.062	5%

Table 27 summarizes the chi-square statistics for the four models and Table 28 shows the results of the chi-square difference tests. The results suggest that the four-factor model fits significantly better than the three-factor model, which fits significantly better than the two-factor model, which, in turn, fits significantly better than the one-factor model.

Table 27: Chi-square statistics for the one-, two-, three-, and four-factor models for Blocks S20 (excluding Item 10) and S4

# of factors	Chi-square	DF	P-value
1	5872.69	371	0.000
2	5822.59	348	0.000
3	5765.84	326	0.000
4	5715.12	305	0.000

Table 28: Tests of chi-square difference statistics between factor models for Blocks S20 (excluding Item 10) and S4

	Chi-square difference	DF difference	P-value
2-factor vs. 1-factor	50.10	23	0.001
3-factor vs. 2-factor	56.75	22	0.000
4-factor vs. 3-factor	50.72	21	0.000

The last criterion by which we judge the meaningfulness of the factor model solution is factor loadings. An examination of the four model solutions suggests that the one- and two-factor solutions make more sense than the three- and four-model solutions. In the one-factor solution, almost all of the items loaded substantially on the factor, except for Item 4 of Block S20 and Item 4 of Block S4. This again indicates the existence of a general factor that explains a considerable amount of variance of the items.

In the two-factor model, the majority of items are strong indicators of factor 1 and only Items 2 through 5 of Block S20 are strong indicators of factor 2, as shown in Table 29. The items that have substantial loadings on factor1 covered content of all

the three fields of science and skills in the three cognitive domains. In contrast, the items that load substantially on factor 2 were designed to assess conceptual understanding. A close look at the content of the four items suggests that, as compared with the rest of the items, they have higher demands for content knowledge. In other words, a lack of subject matter knowledge has a more direct impact on the chances of correctly responding to these items than inadequacies in cognitive skills. Based on the examination of the item content, the first factor can be interpreted as the general science proficiency and the second factor as the recall of content knowledge.

Table 29: Factor loadings of the two-factor model for Blocks S20 (excluding Item 10) and S4

Item No.	Content	Process	Factor 1	Factor 2
S20_1	Life	Conceptual understanding	**.358**	.087
S20_2	Earth	Conceptual understanding	-.091	**.807**
S20_3	Physics	Conceptual understanding	.107	**.418**
S20_4	Earth	Conceptual understanding	-.176	**.405**
S20_5	Earth	Conceptual understanding	.197	**.647**
S20_6	Physics	Practical reasoning	**.418**	.204
S20_7	Physics	Practical reasoning	**.434**	.224
S20_8	Earth	Practical reasoning	**.435**	.129
S20_9	Earth	Practical reasoning	.207	.186
S20_11	Physics	Conceptual understanding	**.314**	.020
S20_12	Earth	Conceptual understanding	**.360**	-.049
S20_13	Life	Practical reasoning	**.533**	.025
S20_14	Life	Practical reasoning	**.510**	-.014
S20_15	Physics	Scientific investigation	**.525**	.287
S20_16	Physics	Practical reasoning	**.508**	.007
S4_1	Earth	Conceptual understanding	**.463**	.050
S4_2	Physics	Scientific investigation	**.525**	.147
S4_3	Physics	Scientific investigation	**.661**	.166
S4_4	Physics	Scientific investigation	**.429**	-.153
S4_5	Earth	Scientific investigation	**.892**	**-.353**
S4_6	Earth	Conceptual understanding	**.547**	-.022
S4_7	Earth	Practical reasoning	**.514**	.008
S4_8	Earth	Conceptual understanding	**.610**	.280
S4_9	Earth	Scientific investigation	**.508**	**.414**

Note: Numbers in bold are substantial loadings (>.30) on the two factors.

- Results in terms of substance

The results of exploratory factor analysis suggest that a unidimensional model may be adequate for explaining the examinees' performances on the items, while the two-factor model shows more distinction among the items. Specifically, a general factor of science proficiency is sufficient to account for examinees' performances on the items, but the differential content and process demands of the items result in the dichotomization of the general factor.

5.3.1.3 The mixture Rasch model

Similarly, two- and three-class mixture Rasch models were fit to the response data for Blocks S20 and S4 in WinBUGS. In each model, five chains with over-dispersed starting values of class proportions were run. Again, the problem of label switching was dealt with by pre-assigning a small number of examines to each latent class. For each model, a total of 10,000 iterations were simulated for each of the five chains. The first 5000 iterations were discarded as burn-ins, and the remaining 5000 iterations were sampled for each chain. Thus, posterior estimates of the model parameters were calculated from a total of 25,000 iterations.

- Results in terms of parameters and fit

The two-class mixture Rasch model solution

In estimating the two-class mixture model, the problem of label switching was fixed by assigning ten examinees to each latent class with certainty. Figure 19 shows the history plots for a subset of model parameters being monitored. Clearly, all the five chains converged quickly to the posterior distribution in each plot.

Figure 19: History plots for a subset of parameters being monitored in the two-class mixture Rasch model for Blocks S20 (excluding Item 10) and S4

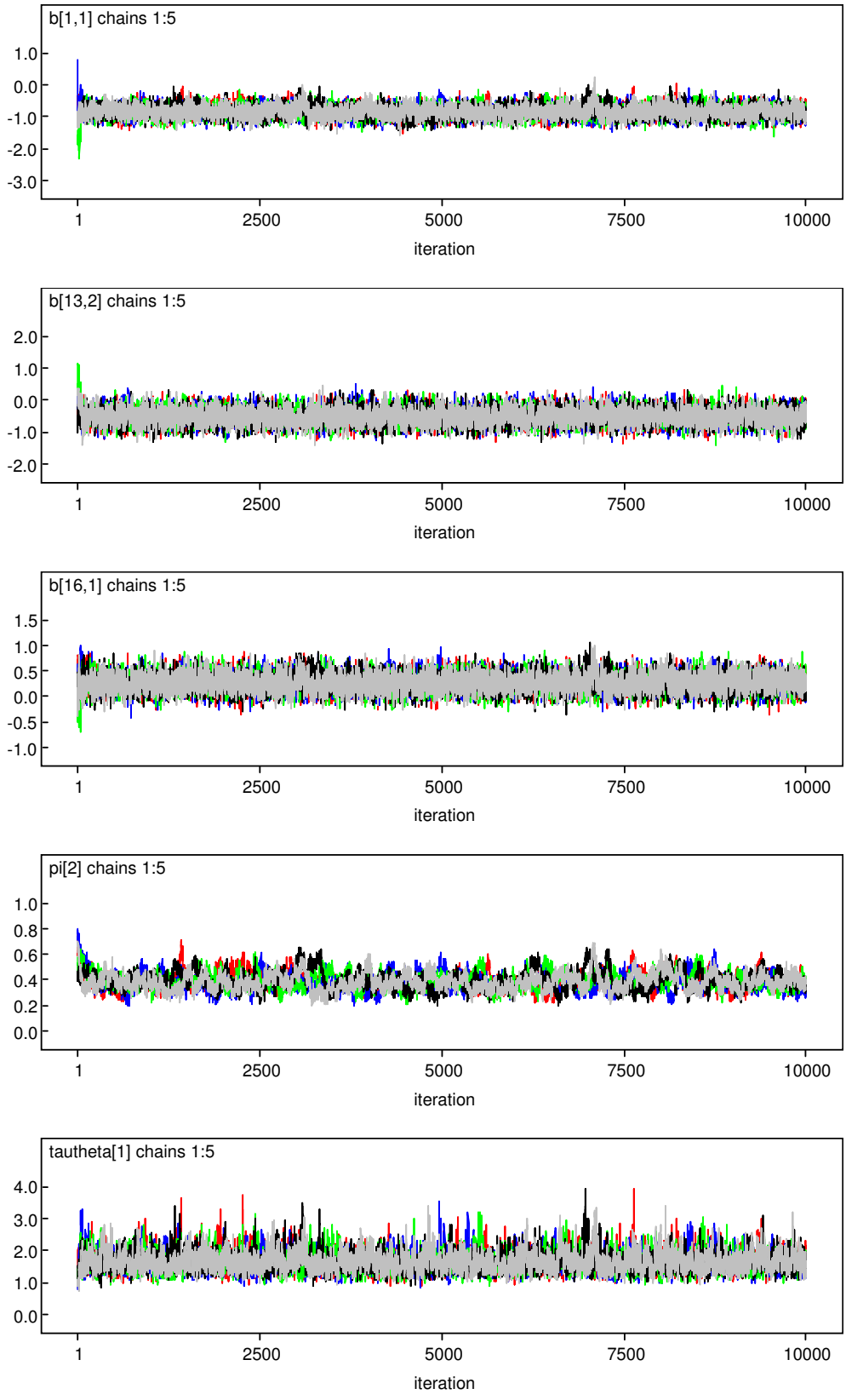


Table 30 summarizes the estimated difficulties of the two-class mixture solution. A general pattern observed is that the item difficulties estimated for Class 1 are greater than those for Class 2, except for Item 4 of Block S4. This means that Class 2 was made up of students who were generally more capable than students in Class 1, given their performance on the two blocks of items. This finding is essentially the same as the one found in the mixture model analysis of Block S20.

Table 30: Item difficulties of the 2-class mixture Rasch model for Blocks S20 (excluding Item 10) and S4

Item	Content	Process	Class 1	Class 2
S20_1	Life	Conceptual understanding	-.8338 (.1787)	-1.246 (.2436)
S20_2	Earth	Conceptual understanding	.3265 (.1719)	-1.389 (.3467)
S20_3	Physics	Conceptual understanding	-.1452 (.1584)	-1.098 (.2851)
S20_4	Earth	Conceptual understanding	2.709 (.2667)	2.368 (.3289)
S20_5	Earth	Conceptual understanding	2.783 (.3671)	-.01154 (.3161)
S20_6	Physics	Practical reasoning	1.959 (.2905)	.452 (.2215)
S20_7	Physics	Practical reasoning	1.577 (.2311)	-.1024 (.2428)
S20_8	Earth	Practical reasoning	-.6906 (.1749)	-1.763 (.2908)
S20_9	Earth	Practical reasoning	.9916 (.1707)	.2848 (.2297)
S20_11	Physics	Conceptual understanding	.9005 (.168)	.6767 (.2266)
S20_12	Earth	Conceptual understanding	-.1893 (.161)	-.6588 (.2252)
S20_13	Life	Practical reasoning	-.3195 (.1768)	-1.484 (.2758)
S20_14	Life	Practical reasoning	.1182 (.1806)	-.5074 (.2249)

S20_15	Physics	Scientific investigation	1.979	-.2034
			(.289)	(.2809)
S20_16	Physics	Practical reasoning	.8648	.1029
			(.2177)	(.2296)
S4_1	Earth	Conceptual understanding	.275	-.8061
			(.1623)	(.2725)
S4_2	Physics	Scientific investigation	1.438	-.8166
			(.2848)	(.2769)
S4_3	Physics	Scientific investigation	.9869	-2.094
			(.2865)	(.4568)
S4_4	Physics	Scientific investigation	1.018	.8515
			(.1781)	(.2362)
S4_5	Earth	Scientific investigation	-2.082	-3.199
			(.235)	(.4708)
S4_6	Earth	Conceptual understanding	1.851	.5444
			(.2288)	(.2431)
S4_7	Earth	Practical reasoning	-1.009	-2.141
			(.1833)	(.3571)
S4_8	Earth	Conceptual understanding	1.804	-.8376
			(.3532)	(.3289)
S4_9	Earth	Scientific investigation	2.52	-.9149
			(.455)	(.4672)

Note: Numbers in parentheses are standard deviations of the posterior distributions for the estimated parameters.

The three-class mixture Rasch model solution

In estimating the three-class mixture model, six examinees were preassigned to Class 1, six to Class 2, and ten to Class 3. The number of priors varied across latent classes because fewer examinees could be classified to Class 1 or Class 2 with high degrees of confidence based on results of the preliminary run. Figure 20 shows the history plots for a small subset of model parameters being monitored. In each plot, all the five chains converged to the posterior distribution after the completion of a small number of iterations.

Figure 20: History plots for a subset of parameters being monitored in the three-class mixture Rasch model for Blocks S20 (excluding Item 10) and S4

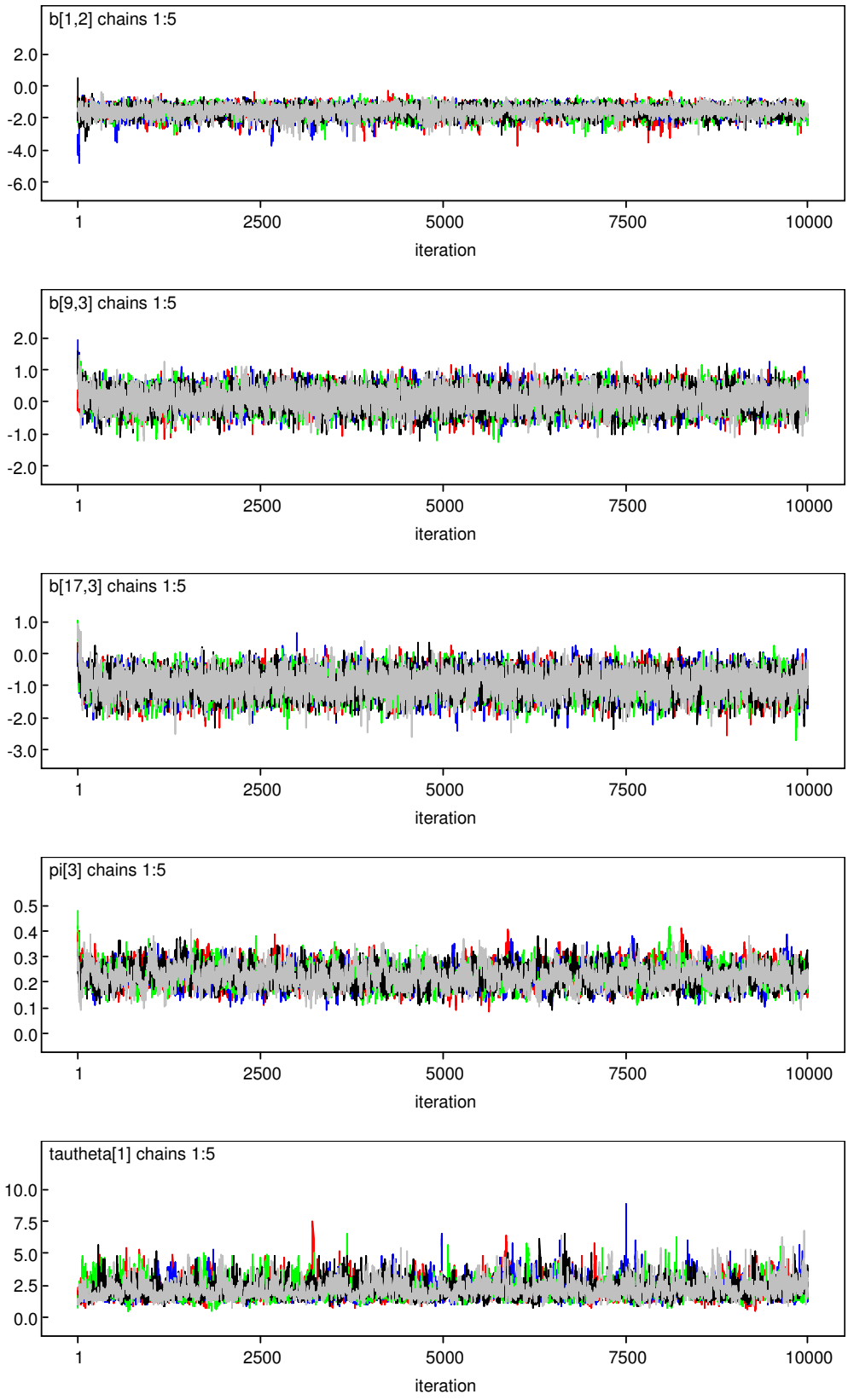


Table 31 summarizes the estimated item difficulties of the three-class mixture solution. For all the items except Item 4 of Block S20, Class 1 has the highest item difficulties among all the three classes. Class 2 has higher item difficulties than Class 2 on all but six items. The six items cover all the three content areas and all the three cognitive dimensions. No tentative conclusions can be made on the basis of the six items about Class 2 and Class 3 with regard to their qualitative differences.

Table 31: Item difficulties of the 3-class mixture Rasch model for Blocks S20 (excluding Item 10) and S4

Item No.	Content	Process	Class 1	Class 2	Class 3
S20_1	Life	Conceptual understanding	-.2606 (.3452)	-1.575 (.2998)	-1.093 (.3212)
S20_2	Earth	Conceptual understanding	.4281 (.2587)	-.06561 (.2556)	-2.269 (.5777)
S20_3	Physics	Conceptual understanding	-.07908 (.2636)	-.1669 (.2488)	-2.065 (.5915)
S20_4	Earth	Conceptual understanding	2.339 (.3724)	3.157 (.5691)	2.07 (.3875)
S20_5	Earth	Conceptual understanding	2.887 (.4925)	1.832 (.3767)	-.7151 (.376)
S20_6	Physics	Practical reasoning	3.021 (.6332)	.7477 (.3474)	.6263 (.3078)
S20_7	Physics	Practical reasoning	2.443 (.6634)	.6519 (.2635)	-.4064 (.3258)
S20_8	Earth	Practical reasoning	-.2426 (.3408)	-1.313 (.261)	-1.986 (.4246)
S20_9	Earth	Practical reasoning	.9471 (.283)	.8002 (.2573)	.07646 (.2991)
S20_11	Physics	Conceptual understanding	1.379 (.4262)	.5166 (.2331)	.6567 (.3054)
S20_12	Earth	Conceptual understanding	.04394 (.2425)	-.4606 (.2286)	-.8256 (.317)
S20_13	Life	Practical reasoning	.179 (.3305)	-1.117 (.3103)	-1.521 (.3699)
S20_14	Life	Practical reasoning	.6603 (.3275)	-.6699 (.3243)	-.2704 (.3115)

S20_15	Physics	Scientific investigation	2.827	.9086	-.5131
			(.8082)	(.3071)	(.311)
S20_16	Physics	Practical reasoning	1.759	-.09723	.4429
			(.6428)	(.2872)	(.3312)
S4_1	Earth	Conceptual understanding	.5013	-.02669	-1.359
			(.2841)	(.2437)	(.3982)
S4_2	Physics	Scientific investigation	2.084	.2249	-.9223
			(.4707)	(.3414)	(.3354)
S4_3	Physics	Scientific investigation	2.508	-.6763	-2.51
			(.8592)	(.4839)	(.5897)
S4_4	Physics	Scientific investigation	1.625	.4526	1.067
			(.4462)	(.2442)	(.3386)
S4_5	Earth	Scientific investigation	-1.536	-3.474	-2.954
			(.3331)	(.8341)	(.5526)
S4_6	Earth	Conceptual understanding	2.018	1.282	.3535
			(.3547)	(.2794)	(.3032)
S4_7	Earth	Practical reasoning	-.9024	-1.377	-2.36
			(.3512)	(.3057)	(.4665)
S4_8	Earth	Conceptual understanding	3.26	.5295	-1.549
			(.9238)	(.3455)	(.4508)
S4_9	Earth	Scientific investigation	3.35	1.153	-1.997
			(.9059)	(.3868)	(.6613)

Note: Numbers in parentheses are standard deviations of the posterior distributions for the estimated parameters.

- Results in terms of substance

The mixture model solutions suggest that the examinees' quantitative differences are dominant as compared with their qualitative differences. This is based on the fact that all but one item in the two blocks were consistently harder for a class of examinees than for (the) other classes. Besides, the exceptional item that carries the evidence about the qualitative differences between the classes of examinees is the item that had low internal consistency with the rest of the items, or in other words, the item that failed to assess the construct of science proficiency that all the other items assessed. As discussed in Section 5.1.4, this item was most likely to be assessing

some peripheral or unscientific construct for this sample of examinees.

5.3.2 Computation of information criteria

Table 32 summarizes the fit statistics for the six models being compared in the analyses. These statistics exhibit the same pattern as that shown in Table 12 or 23. Generally speaking, the mixture Rasch models fit the response data the best, the item factor models the worst, and the multidimensional between-item models in the middle. Each of the three information criteria points to a different model as the best-fitting model: the three-class mixture model is most favored by AIC, the between-item multidimensional model by CAIC, and the two-class mixture model by BIC.

Table 32: Comparison of fit statistics across and within types of models for Blocks S20 (excluding Item 10) and S4

Models	-2ln(L)	*p*	AIC	CAIC	BIC
Multidimensional between-item model (dimensions defined by content areas)	10859.528	30	10919.528	11281.943	11040.736
Multidimensional between-item model (dimensions defined by cognitive factors)	10839.338	30	10899.338	**11261.753**	11020.546
1-factor item factor analytic model (unidimensional IRT model)	10979.650	48	11075.650	11655.515	11269.582
2-factor item factor analytic model	10981.762	71	11123.762	11981.478	11410.620
2-class mixture Rasch model	10706.128	51	10808.128	11424.234	**11014.181**
3-class mixture Rasch model	10618.348	77	**10772.348**	11702.547	11083.448

Notes: 1. *p* denotes the number of parameters to be estimated in the model and the sample size *N* is 420 for each model.
2. Models in bold represent the best-fitting models by the different information criteria.

5.4 Synthesis of analysis results across the three data sets

In the above three sections, each of three data sets was analyzed by three types of models and the analysis results for each data set were compared within and across models. In this section, the analysis results of the three data sets are synthesized, with common and unique findings specified and discussed. The purpose of the synthesis is to make a generalization about how well the three types of models were able to describe the data structure of the 1996 NAEP science assessment and how the multidimensionality in the assessment can be accounted for based on the results from the models.

All three sets of between-item multidimensional analysis indicated that the between-item multidimensional model fit the response data better than the unidimensional model, although the definition of the multiple dimensions was not constant across the three data sets. How the multiple dimensions were defined largely depended on the content and cognitive demands of the items in the data set. If the items were designed around one or more specific content strands, like the items in Blocks S7 and S4, and the recall of content knowledge was essential to correct responses to the items, multidimensionality would be better defined in terms of content areas than cognitive factors. In contrast, if the items required more of the application of science process skills than the recall of content knowledge, like the items in Block S20, multidimensionality in terms of cognitive factors would explain the data better than multidimensionality in terms of content areas.

For all three data sets, results of exploratory factor analysis suggested that the

one- and two-factor models were more interpretable than the more complex models. In the one-factor model solution, all the items had substantial loadings on the factor, except for only one or two items. This indicated the existence of a general factor. The scree plot for each data set also supported this finding. The items that had negligible loadings on the factor were items that had low internal consistency with the rest of the items, or, in other words, items that did a poor job in assessing the targeted science proficiency that the other items were all assessing.

For each of the three data sets, the two-factor model provided a more subtle description of the data structure than the one-factor model, although the two factors were interpreted differently for each data set. What the factors represented in each data set was strongly influenced by the items included in the data set. In Block S20, some items were assessing examinees' understanding of the important concepts in science, while other items required scientific reasoning or investigation skills. Therefore, in the two-factor model solution for the first data set, the two factors were interpreted as the doing and knowing aspects of science learning. In the second data set which consisted of two very different item blocks, each block of items required knowledge of very specific subject matter and, probably, different levels of science process skills. As a result, the two factors corresponded to the two blocks, with each factor representing the knowledge and skills demanded by each block. The third data set ended up with a different version of factor interpretation. The two factors were interpreted as general science proficiency and the ability to recall relevant content knowledge. The first factor was strongly indicated by the majority of the

items in the two blocks, while the second factor was closely associated with a small number of items that required specific content knowledge.

Results of mixture Rasch models for all three data sets implied the prevalence of unidimensionality. All the items, except for one or two, were harder for a class of examinees but easier for the other class(es). This indicated that the quantitative differences between latent classes outweighed their qualitative differences and that the quantitative differences between latent classes could be roughly measured on a single scale. The exceptional items that were harder for the "higher-performing" class but easier for the "lower-performing" class(es) bore evidence of the qualitative differences between latent classes. These items turned out to be the items that had insignificant loadings on the single factor in the one-factor model solution. These items were so poorly written that were not assessing the targeted science proficiency that all the other items were assessing. Therefore, the qualitative differences between latent classes that were based on these items did not have much psychometric meaning. Besides, the between-class differences that were inferred from Figures 12 through 14 in the three-class mixture Rasch model analysis for Block S20 could not be generalized to the other two data sets. Most likely, they resulted from the peculiarity of the items or the sample of examinees. In a word, although the mixture models fit the data sets better than the other two types of models, they did not provide a stronger story about the multidimensionality in the assessment.

Comparisons of information criteria across models for the three data sets all pointed to the same finding: the mixture models fit the data better than the

between-item multidimensional models, which, in turn, fit the data better than the exploratory factor analytic models. As discussed above, the mixture Rasch models did not explain the multidimensionality in the assessment better than the other two types of models, despite their superiority in statistical fit. Furthermore, the finding that the between-item multidimensional models fit the data better than the exploratory factor models suggested the plausibility of the design rationale of the NAEP science assessment.

In addition, finding different models under different information criteria implies that differences among the models in terms of statistical fit were not large enough. Therefore, comparing the models in terms of their substantive meaning has more relevance in identifying the distinctions among the models.

In general, the analysis results of the three data sets shared many commonalities and the differences were largely accounted for. It is expected that the common patterns observed in the analyses and the conclusions drawn from the results can be generalized to the 1996 NAEP science assessment for the 8^{th} grade, and probably to the assessments for the other grades, too, but further study is warranted.

Chapter 6: Conclusions

Measurement models are the lenses through which we view patterns in the assessment data. Our viewpoint is determined by how we conceptualize the knowledge, skills, and abilities in the targeted subject domain, what kinds of inferences we want to make about students, and how we think about the structure of the assessment. In this dissertation, three types of measurement models were compared in the analysis of the NAEP science assessment. They had different assumptions about the assessment structure, targeted at different aspects of students' proficiencies, and conveyed different messages about how students learn science. A comparison of the analysis results across the three data sets revealed some common patterns, which reflected the systematic features of the NAEP science assessment. The main findings from analyzing the three data sets are summarized in the following section.

6.1 Summary of main findings

Confirmatory analyses of the predefined test structures led to the conclusion that the NAEP science assessment was multidimensional. The proposed multidimensional model fit the response data significantly better than the unidimensional model. The definition of the multiple dimensions varied across data sets. The cause of the variation is considered to be the relative content and cognitive demands of the items. If the cognitive demands outweighed the content demands for the majority of the items, the multiple dimensions would correspond to the three cognitive dimensions. Otherwise, the multiple dimensions would correspond to the

three content areas.

Exploratory factor analyses of the three data sets pointed to the same conclusion that the examinees' responses were better explained by the one- and two-factor models than the more complex models. The chi-square difference test indicated that the two-factor model fit significantly better than the one-factor model, but other indices of statistical fit favored the one-factor model. The one-factor model suggested the existence of the general science proficiency that accounted for the examinees' responses. The two-factor model gave more clues about the underlying structure of the assessment and the characteristics of the examinees. The two factors, however, were interpreted differently across the three data sets. The variation in factor interpretation was again due to the different requirements of the items with regard to content knowledge and cognitive proficiencies.

Comparison of the information criteria across the three types of models for the three data sets ended up with the same finding: the 3-class mixture Rasch model was most favored by AIC, the between-item multidimensional model by CAIC, and the 2-class mixture Rasch model by BIC. This conclusion is consistent with expectation. Among the three information criteria, AIC applies the least penalty on model complexity, CAIC applies the most penalty, and BIC applies more penalty than AIC but less than CAIC. As a result, AIC points to the most complex model, and CAIC to the simplest model.

Comparison of narrative stories of the three types of models highlighted the fact that the between-item multidimensional models and the exploratory item factor

analytic models gave plausible stories of the assessment and the performances of the examinees, while the mixture Rasch models did not provide a strong and useful explanation of the multidimensionality in the assessment. Although results of fit statistics showed that the mixture models fit the data better than the other two types of models, they did not have as much substantive meaning as the other two types of models.

6.2 Responses to the meta-questions

In addition to answering the specific questions related to the nature of the multidimensionality in the NAEP science assessment, this dissertation is meant to illustrate the idea of the assessment triangle with a real-world assessment and answer some meta-questions that are of significance to all assessment applications. In the first three chapters, the four research questions are addressed with full details and responses to the questions are then illustrated with results of data analysis. This section is the gist of the discussion in response to the research questions listed in Chapter 1.

Substantive theories, measurement models, and patterns in the data are the three building blocks of an assessment. They should integrate with each other in a consistent manner so as to achieve the goals of the assessment. Measurement models are statistical frameworks motivated by substantive theories and convey narrative stories about the subject domain being assessed. Decisions about what proficiencies are to be inferred about, what aspects of performance are important to the targeted inferences, and what are the relations between the targeted proficiencies

and observed performances are made based on the substantive theories. These decisions are translated into variables, distributions, and formulas in measurement models. Through the structure of measurement models, observed performance data are used to inform inferences about examinees' targeted proficiencies.

As illustrated by the three types of models being applied in this study, different measurement models were inspired by different substantive theories and conceptions about knowing and learning in the subject domain of interest. The important attributes of which the targeted proficiency is conceived to be composed or the significant stages through which the proficiency is developed are used to describe persons or classify items. The interactions between persons and items are characterized in different measurement models in terms of different variables, distributions, and equations. Different types of performance data may be required by different measurement models. In this study, however, the three types of models analyzed the same sets of data, but from different perspectives. Each model highlighted certain patterns of the data at the expense of hiding other potentially interesting patterns. Consequently, the analysis results generated by the three types of models shared some commonalities but also exhibited considerable differences.

As argued repeatedly in the presentation, selection of an appropriate model should be based on substantive and statistical considerations. However, the substantive aspect of model evaluation is often overlooked in assessment applications. Statistical fit is often mistakenly used as the sole criterion in model comparison and selection. As discussed in the first two chapters, statistical fit should be combined

with considerations of substantive meaningfulness in the evaluation of competing models. More importantly, the model used in the analysis of an assessment should be able to serve the purpose of the assessment and make the types of inferences that the test developers want to make about the examinees. Only in this way can the measurement model achieve consistency with substantive theories and observed data.

This dissertation is meant to illustrate how substantive theories, patterns in the data, and measurement models should be connected to the other two in a meaningful way in order to produce an effective assessment and generate sound inferences (National Research Council, 2001). Basically, assessment design is an iterative process of clarifying the specifics of the three building blocks and strengthening the connections among them (Wilson, 2005). In this dissertation, I demonstrated how each model is interpreted as a framework by which test performance is predicted in a way consistent with the underlying theory, how the predictions about patterns in the response data are different from each other, and how the different predictions can be synthesized to provide a fuller and richer account of students' proficiencies, and, more importantly, to inspire another iteration of assessment design, in which we refine the substantive model, create new tasks, collect more data, and analyze the data with an appropriate statistical model. Each iteration of work represents an effort of strengthening the linkage among the elements of an assessment and enhancing the cohesion and effectiveness of the underlying argument.

6.3 Implications to assessment design and analysis

The design framework of the NAEP science assessment indicates that the

assessment tasks were designed to measure both content knowledge and science process skills. However, results of analyses, as discussed in Chapter 5, revealed that the requirements for content knowledge and process skills varied from task to task. Many combinations of content knowledge and process skills can be possibly involved in science assessment tasks. To better conceptualize the content and process demands, Baxter and Glaser (1998) proposed a content-process space, based on which assessment items can be roughly divided into four types, namely, "content rich-process open, content lean-process constrained, content lean-process open, and content rich-process constrained" (p. 38). The location of an assessment task in this space determines the cognitive activities involved for successful task completion, and observed patterns of performance reflect examinees' development stages with regard to content knowledge and process skills.

Baxter and Glaser's framework helps task designers translate their assessment goals into content and process demands and design tasks that are aligned with those demands. With the desired content and process requirements in mind, tasks are designed in a way that gives examinees ample opportunities to engage in the appropriate cognitive activities that demonstrate how much content knowledge and process skills they have possessed. Furthermore, Baxter and Glaser (1998) argued that "recognition of the interrelationships among the subject matter and cognitive features of assessment situations provides a basis for selecting or revising situations to meet specified objectives" (p. 38). Iterations between theory and model-based analysis of task performances provide useful information about what changes should

be made on the content and process demands of the tasks and how task situations can be modified so that they involve the appropriate combination of content knowledge and process skills.

The goal of science learning and instruction is to develop a knowledge structure that links both content knowledge and science process skills with the conditions under which the content knowledge and process skills are to be used. This well-developed knowledge structure is what we call science proficiency or literacy. The goal of science assessments, especially those science assessments for general purposes, is to gauge examinees' science proficiencies using the right type of tasks. However, the difficulty of assessing both content knowledge and process skills in a content rich-process open task resides in the fact that students' failure to complete the task can be accounted for by their lack of the required content knowledge or their insufficiency with the requisite science process skills. Therefore, tasks should be designed with a focus on the distinguishing features of differential competence and achievement. For example, to assess students' cognitive skills, minimum prior knowledge with the content domain should be required in the task so that students' unfamiliarity with the content domain will not be a hindrance to successful task performance. On the other hand, if tasks are meant to assess knowledge generation or recall, they should be designed in a way that gives students explicit direction or guidance with regard to what procedures are to be carried out and how to carry out those procedures.

In Chapter 5, results of mixture Rasch model analyses pointed out items that behaved differently between classes and items that behaved similarly between classes.

For example, in Figure 12, the items off the diagonal of the scatter plot are items that provide the most information about the differences between the two classes, while the items along the diagonal are items that have similar properties between classes. An implication to science assessment task design based on these results is that items can be created that stress the differences between the off-the-diagonal and along-the-diagonal items. For example, ten items can be created that resemble the characteristics of the items off the diagonal and ten items can be created that mimic the items on the diagonal. Students' performances on the two types of items are expected to reveal systematic differences between latent classes and what make the classes different from each other are significant for science learning and instruction.

This dissertation centers on the interrelationships among the substantive, statistical, and operational aspects of an assessment. Questions such as how an assessment depends on the interrelationships, how the interrelationships can be maintained, and how to check the strength of the interrelationships are addressed by analyzing students' response data in an existing assessment through application of three types of measurement models. The discussion of the interrelationships and the analysis of the real-world assessment are meant to furnish researchers or practitioners with a better understanding of the principles in assessment design and analysis. The principles can be applied to inform efforts in the design of new assessments or the modification and analysis of existing assessments.

In new assessment design, having the three elements of the assessment triangle and their interrelationships laid out at the outset of the project helps the designers

organize their thoughts and plan out all the subsequent operational procedures. Decisions about what proficiencies are to be measured, how tasks can be designed, and what measurement models are to be used to analyze the data will set the tone for the entire assessment project, because these decisions will influence all aspects of the assessment's design and use, including content, format, scoring, reporting, and use of results. In addition, the cycle of formulating the substantive theory, designing assessment tasks, and analyzing data through measurement models may need to be iterated in order to achieve a close approximation to the assessment triangle.

The principles of assessment design can also be used to understand or modify existing assessments. Making explicit the building elements of an assessment helps to clarify the set of assumptions underlying the assessment and identify potential inconsistencies among substantive theories, measurement models, and patterns in the data. With an existing assessment, many aspects of the assessment are already in shape and can not be revised on a large scale. However, in case of inconsistencies, the elements of the assessment triangle need to be re-examined and fine-tuned to strengthen their interrelationships. All these examinations and fine-tunings are intended to make the existing assessment meet its goals within the context of the existing constraints.

In this dissertation, three different types of measurement models were used to analyze the same sets of response data from the 1996 NAEP science assessment. Each type of models is inspired by one school of thinking about how students learn science, and each type of models highlights a certain pattern in the data sets.

Different models bring different features of the data sets into attention and produce different analysis results. More complex models could be used to analyze the same data sets and might revoke more features that reside in the data sets. However, there is no unified model that can tackle all the patterns in the data. Highlighting certain patterns in the model analysis often necessitates hiding other patterns in the data, which may be potentially meaningful. In the real world, it is not practical to analyze the same data sets with multiple models, like what was done in this dissertation. The goal is to apply a measurement model that is coherent with the substantive theories and assessment data and serves the assessment purpose sufficiently well.

6.4 Limitations of the study and future work

Due to the confidentiality of the NAEP science assessment data, analyses were conducted solely with responses to the publicly released item blocks. Since there were only four item blocks released for public use and not any three of them appeared together as a booklet, no analysis was done at the booklet level. The number of items included in each of the three data sets analyzed in this presentation was relatively small. This created a problem, especially in the multidimensional between-item models, in which some dimensions were related to a very small number of items. As a result, the accuracy of the estimates of person parameters along those dimensions was less than satisfactory.

In addition, due to the matrix sampling design feature of the NAEP science assessment, analyzing more items would involve having fewer examinees in the response data. This problem was evident in the analyses of two item blocks when

the sample size was reduced to around 420. A small sample size like that would have an adverse impact on the estimation of parameters and interpretation of significance tests.

In this study, all the analyses were done on dichotomous responses, due to the unavailability of the software program POLYFACT, which is able to perform exploratory factor analysis on polytomous items. In the dichotomization of polytomous responses, potentially interesting information might be lost. This limitation can be overcome by securing the software or by programming in other software packages.

In this study, guessing was not accounted for in any of the three types of models. The reason for not modeling the influence of guessing was that the first type of model, namely the MRCMLM model, is a Rasch-type model and unable to accommodate guessing. Leaving guessing out of the models could achieve a better consistency among the three types of models. However, it should be noted that students were very likely to have resorted to the guessing strategy in the NAEP science assessment when they did not know the correct answer to an item or when the administration time was about to expire.

In the mixture Rasch model analyses, interpretation of the latent classes was constrained by the given classification schemes of the items. In other words, the latent classes were interpreted on the basis of the common characteristics of the items that were provided by NAEP. The items that behaved differently across latent classes might share some other types of characteristics and those characteristics might

be potentially meaningful. However, those characteristics could not be studied in this dissertation due to the constraints on the item classification schemes.

In this study, three types of models were selected to analyze the same sets of response data. The reason why these models were selected was that each type of models was consistent with a strand of theory about science learning and each type of models delivered a substantively meaningful story. There exist other types of models that are substantively meaningful and worth studying. For example, the multi-trait multi-method (MTMM) style model is of interest substantively but requires an unrealistically large number of items. Due to the limited availability of the released NAEP science assessment items, the MTMM model was not studied in this dissertation.

Science assessments are often designed to be multidimensional since they cover a variety of content areas and involve a spectrum of cognitive abilities. An extension to the study on test multidimensionality is a study on subscale score reporting. Specifically, reporting subscale scores for the subdomains in science is in line with the current conception about science learning and has substantial pedagogical meaning. Research efforts should be directed to areas such as the development of models that yield precise and reliable subdomain scores, the implementation of efficient estimation methods, and the application of these methods to computerized adaptive testing.

Appendix A: Items in Block S20 for Grade 8

1. A certain organism has many cells, each containing a nucleus. If the organism makes its own food, it would be classified as

A. a bacterium

B. a fungus

C. a plant

D. an animal

2. If the locations of earthquakes over the past ten years were plotted on a world map, which of the following would be observed?

A. Earthquakes occur with the same frequency everywhere on Earth.

B. Earthquakes generally occur along the edges of tectonic plates.

C. Earthquakes most frequently occur near the middle of continents.

D. Earthquakes do not seem to occur in any consistent pattern.

3. Which of the following energy sources is the best example of a nonrenewable resource?

A. Coal

B. Wind

C. Water

D. Sunlight

4. The two most common elements in the Earth's crust are

A. oxygen and silicon

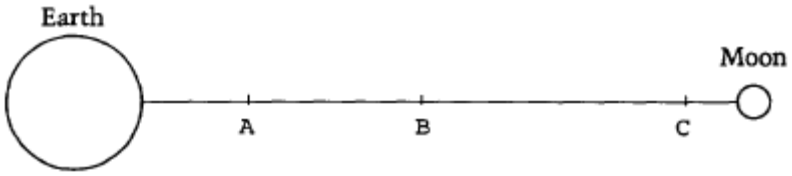
B. oxygen and hydrogen

C. carbon and iron

D. carbon and sulfur

5. A space station is to be located between the Earth and the Moon at the place where

the Earth's gravitational pull is equal to the Moon's gravitational pull. On the diagram below, circle the letter indicating the approximate location of the space station. Explain your answer.



6. Many young people in their twenties have a significant hearing loss in the high-frequency range. Name one factor that contributes to this loss of hearing. Name two ways people could prevent this loss of hearing.

7. When operating ordinary incandescent lightbulbs produce a lot of heat in addition to light. Fluorescent lightbulbs produce much less heat when operating. If you wanted to conserve electricity, which type of bulb should you use? Explain your answer.

8. Maria's house is near a stream. She wants to put her vegetable garden close to the edge of the stream. Discuss one advantage and one disadvantage of putting the garden there.

9. Mrs. Sanchez grows crops on her farm in a hilly region where soil erosion is a big problem. Which of the following would normally help most to protect the soil on her farm from eroding?

A. Rotating her crops on a yearly basis

B. Using contour plowing

C. Irrigating her crops more frequently

D. Using more chemical pesticides

10. What does a mitochondrion do in a cell?

A. It controls the transport of substances leaving and entering the cell.

B. It contains the information to control the cell.

C. It produces a form of energy that the cell can use.

D. It breaks down waste products in the cell.

11. An insulated bottle keeps a cold liquid in the bottle cold by

A. destroying any heat that enters the bottle

B. keeping cold energy within the bottle

C. trapping dissolved air in the liquid

D. slowing the transfer of heat into the bottle

12. According to current scientific theory, as the Solar System formed, matter in the solar nebula came together to form planets. The force most responsible for these formations was

A. gravitational

B. electrical

C. magnetic

D. nuclear

13. A group of students took potato salad made with mayonnaise to a picnic on a very hot day. Explain how eating the potato salad could cause food poisoning. Describe something that could be done to the potato salad to prevent the people who eat it from getting food poisoning.

14. When a population of mice is infected with parasites, many of the mice die from the parasitic infection, but some mice appear as healthy as they were before being infected. Some people are considering using these parasites to control the mouse population in people's homes. Give one advantage and one disadvantage of using these parasites instead of mouse traps or poisons to limit the population of mice.

Questions 15-16 refer to an experiment your teacher asks you to perform to compare the heating rate of soil with that of water. To do this, you are given the following materials.

2 heat lamps

2 bins

2 thermometers

1 sample of soil

1 sample of water

1 timer

You are instructed to heat a sample of soil and a sample of water with heat lamps, measuring the temperature of each sample once a minute for 8 minutes.

15. There are many experimental variables that must be controlled for in order to perform this experiment accurately. Name three of these variables.

16. Suppose that the experiment yielded the results shown in the table below.

Time (min)	0	1	2	3	4	5	6	7	8
Soil temp (℃)	20	21	22.5	24	26	27.5	29.5	30.5	32
Water temp (℃)	20	21.5	23	23.5	24	25.5	26	27.5	28.5

At a beach that has white sand, you measure the temperature of the sand the temperature of the seawater at 9:00 a.m. You find that both have a temperature of 16℃. If it is clear and sunny all morning, what do the data from the experiment predict about the temperature of the white sand compared to the temperature of the seawater at noon? Explain your answer. Explain why the prediction based on the data might be wrong.

Appendix B: Latent class membership and examinees' background variables (for the sample of examinees who responded to Block S20)

1. Gender

	Latent Class		
Gender	1	2	Total
Male	182	440	622
	49.9%	49.7%	49.7%
Female	183	446	629
	50.1%	50.3%	50.3%
Total	365	886	1251
	29.2%	70.8%	

2. Individualized education plan (IEP)

	Latent Class		
IEP	1	2	Total
Yes	36	40	76
	9.9%	4.5%	6.1%
No	329	846	1175
	90.1%	95.5%	93.9%
Total	365	886	1251
	29.2%	70.8%	

3. Limited English proficiency (LEP)

	Latent Class		
LEP	1	2	Total
Yes	23	9	32
	6.3%	1.0%	2.6%
No	342	877	1219
	93.7%	99.0%	97.4%
Total	365	886	1251
	29.2%	70.8%	

4. Race/Ethnicity

	Latent Class		
Race	1	2	Total
White	118	548	666
	32.3%	61.9%	53.2%
Non-white	238	322	560
	65.2%	36.3%	44.8%
Omitted	9	16	25
	2.5%	1.8%	2.0%
Total	365	886	1251
	29.2%	70.8%	

5. Mother's highest education

	Latent Class		
Mother's education	1	2	Total
Didn't finish high school	57	86	143
	15.6%	9.7%	11.4%
High school or more	215	692	907
	58.9%	78.1%	72.5%
Omitted	93	108	201
	25.5%	12.2%	16.1%
Total	365	886	1251
	29.2%	70.8%	

6. Father's highest education

	Latent Class		
Father's education	1	2	Total
Didn't finish high school	59	89	148
	16.2%	10.0%	11.8%
High school or more	195	638	833
	53.4%	72.0%	66.6%
Omitted	111	159	268
	30.4%	17.9%	21.4%
Total	365	886	1251
	29.2%	70.8%	

7. Agree/disagree: I like science.

	Latent Class		
I like science	1	2	Total
Agree	150	475	625
	41.1%	53.6%	50.0%
Disagree/Not sure	206	401	607
	56.4%	45.2%	48.5%
Omitted	9	10	19
	2.5%	1.1%	1.5%
Total	365	886	1251
	29.2%	70.8%	

8. Agree/disagree: I am good at science.

	Latent Class		
I am good science	1	2	Total
Agree	98	453	551
	26.8%	51.1%	44.0%
Disagree/Not sure	255	422	677
	56.4%	47.6%	54.1%
Omitted	12	11	23
	2.5%	1.2%	1.8%
Total	365	886	1251
	29.2%	70.8%	

9. Agree/disagree: Science is useful for everyday problems.

	Latent Class		
Science is useful	1	2	Total
Agree	106	368	474
	29.0%	41.5%	37.9%
Disagree/Not sure	249	505	754
	68.2%	57.0%	60.3%
Omitted	10	12	22
	2.7%	1.4%	1.8%
Total	365	886	1251
	29.2%	70.8%	

10. Agree/disagree: Science is a hard subject.

	Latent Class		
Science is hard	1	2	Total
Agree	145	315	460
	39.7%	35.6%	36.8%
Disagree/Not sure	208	560	768
	57.0%	63.2%	61.4%
Omitted	12	11	23
	3.3%	1.2%	1.8%
Total	365	886	1251
	29.2%	70.8%	

11. Agree/disagree: Learning science is mostly memorization.

Learning science is memorization	Latent Class 1	Latent Class 2	Total
Agree	118	300	418
	32.3%	33.9%	33.4%
Disagree/Not sure	234	569	803
	64.1%	64.2%	64.2%
Omitted	13	17	30
	3.6%	1.9%	2.4%
Total	365	886	1251
	29.2%	70.8%	

12. How often do you study science in school?

Study science in school	Latent Class 1	Latent Class 2	Total
Everyday	229	644	873
	62.7%	72.7%	69.8%
Less frequent than everyday	120	227	347
	32.9%	25.6%	27.7%
Omitted	16	15	31
	4.4%	1.7%	2.5%
Total	365	886	1251
	29.2%	70.8%	

13. How much time per week do you spend on doing science homework?

Time on science homework	Latent Class 1	Latent Class 2	Total
Less than 1 hour	231	544	775
	63.3%	61.4%	62.0%
1 hour or more	116	323	439
	31.8%	36.5%	35.1%
Omitted	18	19	37
	4.9%	2.1%	3%
Total	365	886	1251
	29.2%	70.8%	

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255-278.

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol III; pp. 143-166). Norwood, NJ: Ablex.

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

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

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report.* Washington, D.C.: National Center for Education Statistics.

Ansley, T. N., & Forsyth, R. A. (1990). An investigation of the nature of the interaction of reading and computational abilities in solving mathematics word problems. *Applied Measurement in Education, 4,* 319-329.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: Issues and Practice, 17(3), 37-45.

Bhansali, R. J., & Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion.

Biometrika, 64, 547-551.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

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

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39,* 331-348.

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

Brooks, S. P., & Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434-455.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5-32.

Chung, H., Loken, E., & Schafer, J. L. (2004). Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician, 58,* 152-158.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42,* 133-148.

Congdon, P. (2003). *Applied Bayesian modeling.* New York: Wiley

Dayton, C. M. (1998). *Latent class scaling analysis.* Thousand Oaks, CA: Sage.

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

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 36,* 359-374.

Greeno, J. G.. (1983). Conceptual entities. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 227-252). Hillsdale, NJ: Lawrence Erlbaum Associates.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-47). New York: Simon and Schuster Macmillan.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp.74-94). Vancouver, BC: Educational Research Institute of British Columbia.

Jöreskog, K. G.., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software International.

Junker, B. W. (1993). Conditional association, essential independence and monotone

unidimensional item response models. *The Annals of Statistics, 21,* 1359-1378.

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

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2006). *Model selection methods for mixture dichotomous IRT models.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Li, J. (2006). Not ready for science tests: Questions that need answers before the federal mandate kicks in. *Education Week, 25* (33), 40.

Lin, T., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22,* 249-264.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30,* 23-40.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report ONR 91-2). Iowa City, IA: American College Testing.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3-31.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk, 2,* 237-258.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th Ed). Phoenix, AZ: Greenwood.

Mislevy, R. J., & Huang, C-W. (2006). Measurement models as narrative structures. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models.* New York: Springer.

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-66.

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

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41-71.

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 489-531). Dordrecht, the Netherlands: Kluwer Academic Press.

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

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551-560.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus: The comprehensive modeling program for applied researchers: User's guide.* Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Muthén, L. K. (2001). *Mplus: Statistical analysis with latent variables.* Los Angeles, CA: Statmodel.

National Assessment Governing Board. (2000). *Science framework for the 1996 and 2000 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

National Center for Education Statistics. (1999). *NAEP 1996 state assessment program in science secondary-use data user guide.* Washington, DC: U.S. Department of Education.

National Research Council. (1996). *National Science Education Standards: Observe, interact, change, learn.* Retrieved from http://books.nap.edu/readingroom/books/nses/.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pelligrino, N. Chudowsky, and R. Glaser (Eds.), Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146-178.

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

Reckase, M. D. (1997). A linear logistic multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 217-286). New York: Springer-Verlag.

Reise, S. P., & Gomel, J. N. (1995). Modeling qualitative variation within latent trait dimensions: Application of mixed-measurement to personality assessment. *Multivariate Behavioral Research, 30,* 341-358.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society,* Series B, *59*, 731-792.

Rijmen, F., & De Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture Rasch model. *Psychometrika, 70,* 481-496.

Robert, C. P. (1996). Mixtures of distributions: Inference and estimation. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 441-464). Boca Raton, FL: Chapman & Hall.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271-282.

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

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response*

theory. Thousand Oaks, CA: Sage.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS (Version 1.4)* [Computer software]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions.* Unpublished doctoral dissertation. Oxford: Magdalen College.

Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.), *Handbook of Research on Science Teaching and Learning* (pp 284-300). New York: MacMillan.

Stoel, R.D., Galindo-Garre, F., Dolan, C.,& Van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11*, 439-455.

Stone, C. A., & Yeh, C-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychological Measurement, 66,* 193-214.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331-354.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393-408.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27,* 159-203.

Toulmin, S. E. (1958). *The use of argument*. Cambridge: Cambridge University Press.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276-289.

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

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software International.

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

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

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

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models* (Research Report RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K. & Everson, H. T. (1995). *Modeling the mixture of IRT and pattern responses by a modified hybrid model* (Research Report RR-95-16). Princeton, NJ:

Educational Testing Service.