The Dissertation Committee for Adem Ekmekci Certifies that this is the approved version of the following dissertation:

Mathematical Literacy Assessment Design:

A Dimensionality Analysis of Programme for International Student Assessment (PISA) Mathematics Framework

Committee:

Guadalupe Carmona-Dominguez, Supervisor

Anthony Petrosino

Catherine Riegle-Crumb

Walter Stroup

Daniel Powers

Tiffany Whittaker

Mathematical Literacy Assessment Design:

A Dimensionality Analysis of Programme for International Student Assessment (PISA) Mathematics Framework

by

Adem Ekmekci, B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2013

Dedication

This dissertation is dedicated to the following people:

- To Betul, whose love, understanding, and support kept me alive and energetic throughout my graduate life,
- To Omer H., whose boundless energy and joy allowed me to enjoy the life,
- To Ahmet A. K., whose recent arrival in our family brought joy and bounties,
- To my parents, Hatice and Ali, who have done everything they could for me without any expectations,
- To Allah (c.c.), above all, for his blessings and for everything else he has granted me.

I would also like to express my thanks to the following people for their encouragement and support:

- To Filiz Ablam, Hayriye Ablam, Alim Abim, and Kardesim Arif,
- To my parents-in-law, Hatice and M. Emin,
- To my brother-in-law, Dr. Salih.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Lupita Carmona, who has been there for me from the very beginning of my graduate life. There is so much to say, but in short, thank you for taking a chance with me and for your excellent mentorship, wise guidance, brilliant ideas, and endless support throughout.

I would also like to thank the other members of my committee. I have been privileged to have Dr. Tony Petrosino and Dr. Walter Stroup not only on my committee but also in my graduate program. Their expertise inspired me and helped me build my own professional knowledge base. Dr. Petrosino was always available to answer any questions, academic or not. Dr. Stroup's critical comments and thought-provoking feedback helped me refine my thoughts. I am also very grateful to have Dr. Riegle-Crumb, Dr. Powers, and Dr. Whittaker in my committee. Their input and constructive feedback were invaluably significant to the production of this work.

I am very thankful to have known Dr. Yetkin Yildirim, who has been a friend, a colleague, and a mentor to me, and Dr. Serdal Kirmizialtin, who has been a sincere friend (I am going to miss our coffee breaks). I also extend my thanks to Kim Hughes, Director of the UTeach Institute, for her support during most of my graduate life and to my co-workers at the Institute for their excellent fellowship. Moreover, I would like to express my gratitude to my dear friends and colleagues, Dr. Steven Greenstein, Dr. Brian Fortney, Dr. Teddy Chao, Luz Maldonado, Dr. Mehmet C. Ayar, Dr. Mehmet S. Corlu, Dr. Serkan Ozel, Gladys Krause, and Devi Sivam.

I would also like to acknowledge and thank all the faculty, students and staff in the STEM Education Program and the Department of Curriculum and Instruction at UT-Austin, especially Dr. Jill Marshall and Dr. Susan Empson, who have provided valuable help to me. Last but not the least, I am very grateful to all of my friends, especially the ones from the American Turkish community for making it feel like home here.

Mathematical Literacy Assessment Design:

A Dimensionality Analysis of Programme for International Student Assessment (PISA) Mathematics Framework

Adem Ekmekci, Ph.D.

The University of Texas at Austin, 2013

Supervisor: Guadalupe Carmona-Dominguez

The National Research Council (NRC) outlines an assessment design framework in *Knowing What Students Know*. This framework proposes the integration of three components in assessment design that can be represented by a triangle, with each corner representing: *cognition*, or model of student learning in the domain; *observation*, or evidence of competencies; and *interpretation*, or making sense of this evidence. This *triangle* representation signifies the idea of a need for interconnectedness, consistency, and integrated development of the three elements, as opposed to having them as isolated from each other. Based on the recommendations for research outlined in the NRC's assessment report, this dissertation aims to conduct a dimensionality analysis of Programme for International Student Assessment (PISA) mathematics items. PISA assesses 15-year olds' skills and competencies in reading, math, and science literacy, implementing an assessment every three years since 2000. PISA's mathematics assessment framework, as proposed by the Organisation for Economic Co-operation and

Development (OECD), has a multidimensional structure: *content, processes, and context*, each having three to four sub-dimensions. The goal of this dissertation is to show how and to what extent this complex multidimensional nature of assessment framework is reflected on the actual tests by investigating the dimensional structure of the PISA 2003, 2006, and 2009 mathematics items through the student responses from all participating OECD countries, and analyzing the correspondence between the mathematics framework and the actual items change over time through these three implementation cycles.

Focusing on the cognition and interpretation components of the assessment triangle and the relationship between the two, the results provide evidence addressing construct validity of PISA mathematics assessment. Confirmatory factor analysis (CFA) and structural equation modeling (SEM) were used for a dimensionality analysis of the PISA mathematics items in three different cycles: 2003, 2006, and 2009. Seven CFA models including a unidimensional model, three correlated factor (1-level) models, and three higher order factor (2-level) models were applied to the PISA mathematics items for each cycle. Although the results did not contradict the multidimensionality, stronger evidence was found to support the unidimensionality of the PISA mathematics items. The findings also showed that the dimensional structure of the PISA mathematics items were very stable across different cycles.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Mathematical Literacy

Definition and Important Concepts

This dissertation study is an investigation of mathematical literacy assessment design from an international perspective. The term mathematical literacy in this dissertation refers to successful mathematics learning and sufficient mathematics knowledge and skills to function well in society in very broad terms. Other terms used in the literature to imply this meaning to some extent are quantitative literacy, numeracy, mathemacy, mastery of mathematics, mathematical proficiency, and mathematical competence (Kilpatrick, 2001). Different terms might be preferred over others in different educational systems. For example, in the U.K. and Australia the term *numeracy* is used (Stacey, 2010), whereas some prefer *quantitative literacy* in the U.S (Steen, 2001). I use the term mathematical literacy (Organization for Economic Cooperation and Development [OECD], 2003) in this dissertation as an umbrella term more or less representing all of these terms more or less. In addition, terms such as modeling, mathematisation, and mathematizing are the mathematical conceptions used in the literature that are closely related to mathematical literacy (Jablonka, & Gellert, 2007; Lesh & Carmona, 2003).

Mathematics educators around the world view mathematical literacy as a multidimensional construct composed of distinguishable but related components rather than single, general mathematics ability. Some math educators (e.g., Kilpatrick,

Swafford, & Findell, 2001) focus on proficiencies or competencies when defining mathematical literacy, while others (e.g., Ojose, 2011) describe knowledge and skills. Some others (e.g., Steen, 2001) situate mathematical literacy according to its connection to real life situations (i.e., context). There is also a content-wise decomposition of mathematical literacy (Steen, 2001). So, there appears to be more than one dimension and more than one approach in composing mathematical literacy as discussed in the mathematics education field.

The OECD defines mathematical literacy as

> an individual's capacity to identify, and understand, the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned, and reflective citizen. (OECD, 2003, p.24)

The OECD also views mathematical literacy as a multidimensional construct in terms of three important aspects: content, process, and context. The first component, "content," is divided into 4 dimensions (overarching ideas): quantity, space and shape, change and relationships, and uncertainty. "Processes" consist of three competency clusters: reproduction, connections, and reflection. Lastly, "situations" are defined in terms of 4 dimensions: personal, educational/occupational, public, and scientific (OECD, 2009a).

Background and Importance

The concept of mathematical literacy gained crucial importance especially in the 80's. Since then, the standards that had been once considered for literacy (being able to

read and write) also began to be considered for mathematical literacy (Jablonka, 2003; Moses & Cobb, 2001). That is, mathematical literacy is as critical in today's society as reading literacy was 40-50 years ago. There is no doubt about how critical the mathematics domain is for the workforce, especially for technical careers that drive the economy. Nor would anyone disagree with emphasis put on mathematics as a school subject. However, what mathematics literacy implies is a different question.

What is meant by mathematical literacy is more than knowing mathematics as a school subject. It has been considered the keystone of public mathematics education as well as an indicator for the quality of educational programs at local, national, and international levels. According to mathematics educators and educational psychologists, being mathematically literate means much more than being ready for the workforce and well equipped to tackle everyday problems (e.g., Ojose, 2011). Mathematical literacy is an essential part to critical education and democracy and a necessity for both personal and national empowerment (Skovsmose, 1994). As Moses and Cobb (2001) put it, "mathematics literacy [is] fundamental to this generation." Applying Paulo Freire's (1970) critical pedagogy theory in mathematics education, mathematical and statistical literacy plays an important role in social and economic development through democratization and liberalism (Frankenstein, 1992). In other words, to be critical citizens, it is necessary to be mathematically literate. Mathematical literacy opens space for civil rights and leads to social change (Moses & Cobb, 2001). Moreover, mathematical literacy could also serve for different assets such as cultural identity, environmental awareness, and developing human capital when different approaches are

taken (Jablonka, 2003). For example, through mathematical modeling, people are equipped with necessary mathematical tools and abilities to succeed in their lives. Ethno-mathematics could help protect cultural assets through connecting the informal math used to solve day-to-day problems and school mathematics. To sum up, we could say that mathematics literacy is not less important than reading and writing in today's world (Moses & Cobb, 2001).

National and international education and assessment organizations, in addition to educational theorists, also acknowledge the critical role mathematical literacy plays in individual and societal life. This view of mathematical literacy differs a lot from traditional school mathematics in many ways. Mathematical literacy means more than having basic mathematical knowledge. It requires students to be able to apply it to solve real world problems (Ojose, 2011). In order for mathematical literacy to make individual reflective citizens and critical thinkers, it needs to pursue flexible transfer of mathematics knowledge and skills and successful application of them in different situations. The National Council of Teachers of Mathematics (NCTM) envisioned and enunciated in late 90's that this would be possible through modeling, estimating, analyzing, reasoning, formulating, and interpreting mathematical problems in variety of contexts (NCTM, 2000).

The most prominent national assessment in the U.S., the National Assessment of Educational Progress (NAEP) assesses student knowledge and skills in science and mathematics at grades four, eight, and twelve. This shows how it is considered very important to evaluate mathematical literacy in the U.S. Internationally, two major

comparative student assessments take place: the Programme for International Student Assessment (PISA), and Trends in International Mathematics and Science Study (TIMSS). Mathematics is one of the domains included in these assessments. Therefore, it would be fair to say that it is unequivocal, both nationally and internationally, that mathematics domain plays a major role and mathematical literacy is seen as crucial in preparing youngsters for real life.

The International Association for the Evaluation of Educational Achievement (IEA) sees mathematics literacy as one of the fundamental educational goals around the world (Mullis et al., 2009). The OECD (2003) states that mathematical literacy is one of the keys to develop human capital, personal, social, and economic well-being, and democratic participation in the social life.

Given the relevance of mathematical literacy in our society, it follows that assessing mathematical literacy is an important facet to mathematics education. Through its assessment, it is possible to understand what mathematical literacy would mean in terms of student achievement. Large-scale, standardized assessments of student performance in mathematics provide information about students' mathematical literacy. In addition, how different levels of mathematical literacy relate to other student characteristics is often explored through these assessments (Anderson et al., 2007).

Assessing Mathematical Literacy

Assessment of mathematical literacy is a complex task. The assessment design framework proposed by the NRC sets the ground. Several important factors interplay in the design of the assessment and each should be paid a great amount of attention. For

example, developing items to assess mathematical literacy is a very important process intertwined with other components of assessment design. From the validity perspective, the items should be measuring what they are intended to measure (Loevinger, 1957). Therefore, item development needs to be very carefully designed.

Large-scale assessments could be valuable only if they are well designed and appropriately used (NRC, 2001). Two important questions serve as the baseline for a good assessment design. The first one is: what views of mathematical literacy are these large-scale assessments designed to reflect? Secondly, what relationships do they have with teaching and learning?

As mentioned earlier, mathematical literacy is often defined and viewed as a multidimensional construct. When it comes to its assessment, especially in large-scale context, this multidimensional conception of mathematical literacy cannot be disregarded. Then, there are important questions to be answered about current large-scale assessment practices with regards to this conception. For example, what conceptions of mathematical literacy do large-scale assessments reflect? What can we say about dimensionality of large-scale assessments? What is the connection between the dimensionality of large-scale assessments and that of mathematical literacy? These questions about large-scale assessment designs for mathematical literacy are important to answer and clarify because it might not be possible otherwise to draw valid inferences from their results.

PISA's Assessment Framework for Mathematical Literacy

The Programme for International Student Assessment (PISA), coordinated by the OECD, is an international assessment that includes a mathematical literacy assessment. PISA is the focus of this dissertation and offers a unique opportunity to evaluate 15-year-olds' mathematical literacy. PISA has been assessing youngsters' skills and competencies in reading, math, and science every three years since 2000. PISA developed assessment frameworks defining reading, mathematical, and scientific literacy and explaining what competencies and dimensions are assessed for each literacy domain. The OECD's definition of mathematical literacy is given earlier. Based on this definition, PISA assesses mathematical literacy in a multidimensional way. Three important aspects of mathematical literacy are content, process, and context. The first component, "content," is divided into 4 dimensions (overarching ideas): quantity, space and shape, change and relationships, and uncertainty. "Processes" consist of three competency clusters: reproduction, connections, and reflection. Lastly, "situations" are defined in terms of 4 dimensions: personal, educational/occupational, public, and scientific (OECD, 2009a).

Content Dimensions

The overarching idea of quantity requires an understanding of numeric phenomena, relationships and patterns encompassing understanding operations, number sense, computations, arithmetic, and estimations. Space and shape content focuses on the understanding of spatial and geometric phenomena and relationships. This content area relates to geometric patterns, differences and similarities of shapes, and relative positions of objects. Change and relationships content focuses on understanding of fundamental

types of change occurring in natural phenomena, representing those changes in a comprehensible form, and functional relationships and dependency among variables of change. Uncertainty content relates to probabilistic and statistical phenomena and relationships. PISA recognizes the importance of uncertainty as viewing the data as numbers in a context and developing an understanding of random events.

Process Dimensions

Process dimension, also known as competency clusters, is composed of reproduction, connections, and reflection sub-dimensions. PISA classifies underlying mathematical skills in these three competency clusters. Reproduction competency deals with factual knowledge, equivalency, recalling mathematical objects and properties, performing routine procedures, standard algorithms, and technical skills. The connections cluster involves a degree of interpretation and linkages. The focus of connections is on linking the different strands and domains within mathematics and integrating information in order to solve problems that allow different strategies and mathematical tools. The reflection cluster relates to analysis and interpretation of mathematics embedded in the situation, development of models and strategies, and making generalizations and proofs.

Context Dimensions

The third dimension, context or situations, consists of personal, educational and occupational, public, and scientific categories. Every mathematical problem is situated within a context in PISA. Personal contexts include day-to-day activities to provide immediate and personal relevance to students. Educational and occupational contexts provide school and work situations which students might encounter while at school or in

the work environment. Public contexts involve situations in which individual interacts with the outside world. Lastly, scientific context presents scientific or explicitly mathematical problems.

Rationale

Although PISA uses a multidimensional mathematical literacy framework, it is not known if the results reflect this multidimensionality. The purpose of this dissertation is to investigate to what extent the PISA mathematics items reflect the complex dimensionality of the original assessment framework. Current practices use expert opinions, which serve as content-wise validation of the PISA mathematics items. However, no study has been undertaken to demonstrate the dimensionality of the actual results of mathematics assessment items that would provide evidence for the construct validity of the PISA mathematics assessment. This dissertation offers to fill in the gap.

Why is it important to investigate this? Is the above-mentioned gap significant enough to need to be filled? Yes, it is because it follows the NRC's recommendations for research, policy, and practice in assessment design for school science and mathematics.

The Committee on the Foundations of Assessment of the NRC outlines an assessment design framework for educators and psychometricians in its 2001 report entitled *"Knowing What Students Know: The Science and Design of Educational Assessment."* This report has explored recent advances in cognitive sciences and their implications for improving assessment of science and mathematics education (NRC, 2001). In the report, the NRC presents the need to link theories of cognition with psychometric perspectives. This report proposes the assessment triangle, where each

corner of the triangle represents: cognition, or model of student learning in the domain; observation, or evidence of understanding; and interpretation, or making sense of this evidence (NRC, 2001).

This "Triangle" representation signifies the idea of the interconnectedness of the three elements as opposed to having them as isolated from each other, which often times is found problematic in most of assessment designs. Of the five important recommendations for research outlined in the NRC's assessment report, one urges for in-depth analyses of these three elements and their coordination. This study will shed light on the C-I (cognition-interpretation) linkage of a prestigious worldwide assessment, PISA.

The NRC urges for "in-depth analyses of the critical elements (cognition, observation, and interpretation) underlying the design of existing assessments that have attempted to integrate cognitive and measurement principles" (NRC, 2001) in this report as a part of future research agenda on educational assessment in science and mathematics. It is also recommended in the report that

> Developers of assessment instruments for classroom or large-scale use should pay explicit attention to all three elements of the assessment triangle (cognition, observation, and interpretation) and their coordination...Considerable time and effort should be devoted to a theory-driven design and validation process before assessments are put into operational use. (NRC, 2001, p.305)

Purpose

This dissertation study has three purposes: (1) to assess the dimensionality of an international assessment for mathematical literacy from an assessment design perspective, (2) to investigate the statistical, structural (factorial) correspondence between PISA mathematics items and PISA mathematical literacy framework, and (3) to explore the interplay between unidimensionality assumptions and multidimensionality expectations. The first purpose will inform the re-conceptualization of mathematical literacy as a multidimensional construct. The second purpose relates to the coordination of cognition and interpretation components of the assessment triangle in PISA's case. Although the OECD carefully developed the mathematics items so that they are offered in a way that reflects the assessment framework (cognition), no statistical validation process for the dimensionality of the PISA mathematics items utilizing the student responses (interpretation) exists. Lastly, PISA uses statistical methods for scaling and interpretation of scores that assumes a unidimensional test structure. However, the intended assessment framework assumes mathematical literacy to be a complex, intertwined, and multidimensional construct. This study will help understand the interplay between these two competing sides.

Research Questions

In this dissertation study, I will investigate the coordination between PISA's mathematical literacy framework and the PISA mathematics items utilizing the student responses through following research questions:

1. What is the correspondence between the dimensional structure of the PISA mathematics items and PISA's mathematical literacy assessment framework in terms of the content, process, and context dimensions?
2. What is the best representation for the dimensional structure of the PISA mathematics items for implementation cycles 2003, 2006, and 2009??
3. How does the dimensional structure of the PISA mathematics items change over time?

Chapter 2: Literature Review

There is no doubt that assessments constitute an important part in today's education. A large variety of assessments exist depending on the context, content area, format, and purpose. Whether in-class or large-scale, measuring achievement or aptitude, in mathematics or reading, formative or summative, some theoretical principles apply to all assessments (NRC, 2001). For example, it is clear through evidence that eliciting knowledge and skills possessed are examples of such principles that underlie all assessments. Every assessment design is essentially based on evidentiary (Mislevy, 1994) and inferential (Messick, 1994) notions. As such, the NRC (2001) identifies three foundational components that should underpin all types of assessments: cognition, observation, and interpretation.

This triad is referred to as the *assessment triangle.* This chapter will commence with detailed explanations of these important elements of assessments outlined by the NRC. The connection among these elements and how they form a coherent whole are then described. This assessment design framework sets the theoretical background for this dissertation. The NRC's recommendations for research as they relate to large-scale assessments and to this dissertation are also summarized. Then, the conceptual framework for this dissertation study is provided. Two very important concepts related to this study, validity and dimensionality are defined. Next, this chapter gives an overview of studies exploring the dimensionality of large-scale assessments and methodologies used in studying their validity. What follows finally is how this study differs from earlier ones.

The Triad for Assessment Design

> The NRC's (2001) report on assessment conveys a clear message:
>
> Every assessment, regardless of its purpose, rests on three pillars: a model of how students represent knowledge and develop competence in the subject domain, tasks or situations that allow one to observe students' performance, and an interpretation method for drawing inferences from the performance evidence thus obtained. In the context of large-scale assessment, the interpretation method is usually a statistical model that characterizes expected data patterns, given varying levels of student competence. (p. 2)

The three important components are cognition, observation, and interpretation. As mentioned above, this triad is referred as the *assessment triangle* (Figure 2.1).

The *cognition* component of the assessment triangle refers to cognitive models of learning. How people learn and develop knowledge and skills in a particular subject area should be the starting point in designing an assessment. Theories of learning informed by educational and learning sciences should be the guide for the designer to identify the set of knowledge and competencies to be targeted and measured by the assessment (Chudowsky & Pellegrino, 2003). As the learning sciences develop to incorporate new theories and models of cognition, cognitive components of the assessment should be modified accordingly to reflect the most recent theories of how people learn and come to understand. The cognition component of an assessment includes different aspects of student knowledge and skills that are drawn from a larger set of theories of learning in a particular domain and that are specified as the targets for assessment. For example, these

targets could be sub-domains such as numbers and probability in mathematics or processes such as reproduction and transfer. In the context of large-scale assessments, these targets are often specified in the assessment frameworks of tests.

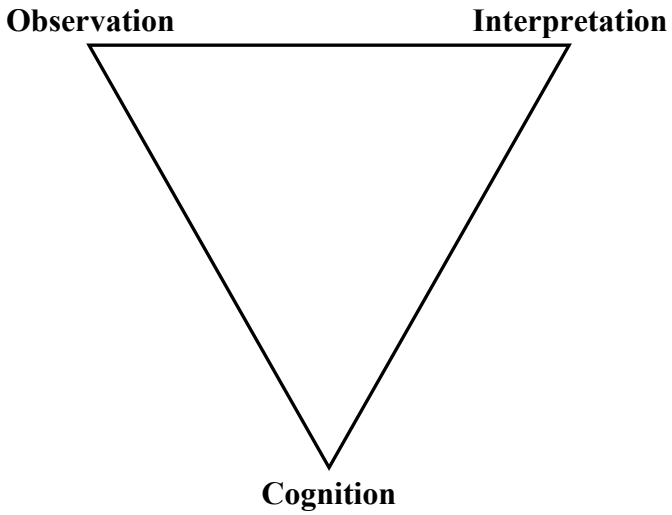


Figure 2.1. Assessment Triangle

Observation represents tasks and/or situations that would demonstrate the learners' performance. It includes questions, problems, tasks, projects, and any prompt given to learners that would reveal what they know and can do in a subject domain. In practical terms, this might mean a written exam or questions responded to orally. This refers to the collection of data as evidence of learning, knowledge and skills. In a large-scale context, formal examination of learners with a set of questions represents *observation* component of the assessment.

Interpretation refers to making sense of the observed performance. That is, interpretation is the process of reasoning from the evidence collected through the observation phase. This process includes transforming the data about learner performance

into assessment results. The interpretation component could also be considered as the collection of models and tools that are used to draw inferences about learners' knowledge and skills. In the large-scale assessments context, the interpretation component generally refers to statistical assumptions and models that are used to characterize response patterns and levels of learning.

These three components are very important in an assessment design. What is more important, though, is that they should be interconnected and in accordance with each other. Otherwise, inferences drawn from assessments might not be as meaningful, if not misleading or inaccurate. There are three linkages in the assessment triangle, each of which are explained below: C-O (cognition-observation), O-I (observation-interpretation), and C-I (cognition-interpretation).

Cognition-Observation linkage assures that tasks are designed with the knowledge and skills in mind that those tasks will demonstrate. Theories of how people learn and develop skills in a particular domain should inform the types of tasks that would reveal evidence about those skills. This is not a one-way relationship, however: learning about what tasks could effectively demonstrate what knowledge and skills (and how those tasks could demonstrate those skills) helps the assessment designer revisit and modify the original assessment framework informed by cognitive theories.

Observation-Interpretation linkage connects the interpretation methods and designing tasks to observe performance. Interpretation models are needed to understand what constitutes evidence for learners' performance in a given task. There are various interpretational models including both statistical and qualitative models. Each

interpretation model could offer different opportunities and has its limitations as well. Knowing what interpretation models are available and what they could offer helps the assessment designer in developing tasks that are effective and efficient in observing targeted performance areas.

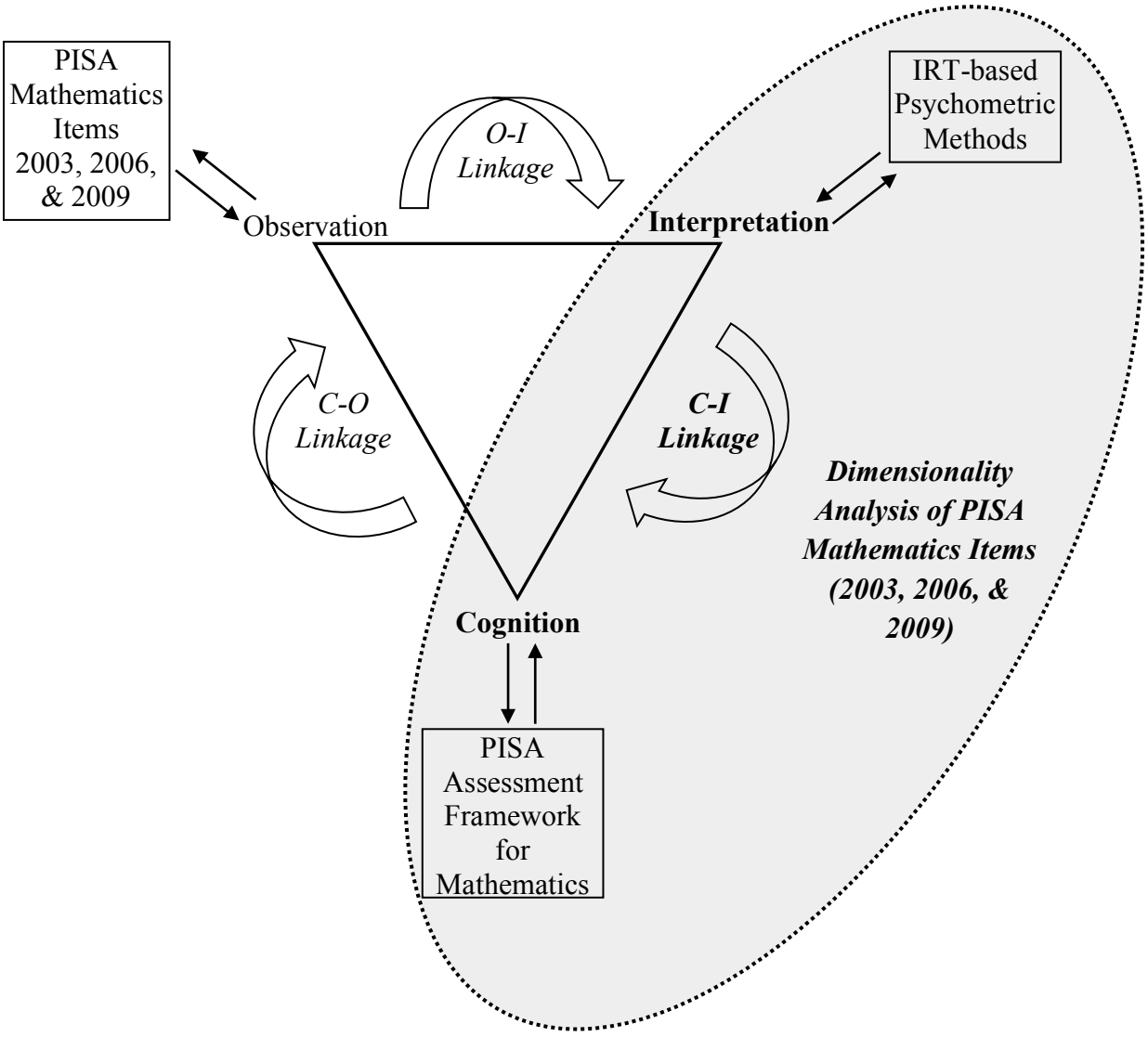


Figure 2.2. Conceptual Framework

Cognition-Interpretation linkage serves to align the types of interpretation models with the cognitive theories. Knowing about how people learn helps determine the appropriate interpretation models. In addition, when drawing inferences about learner performance using different interpretation models, either statistical or qualitative, learner knowledge and skills targeted by the assessment framework should be taken into consideration. Conversely, although the assessment framework is mostly guided by cognitive theories, available interpretation models also inform this framework (i.e. what types of knowledge and skills could be targeted based on the methods of interpreting assessment results).

Conceptual Framework

Applying the NRC's assessment triangle to PISA assessment design sets the conceptual framework for this dissertation study. Figure 2.2 demonstrates how the assessment triangle applies to PISA context. Within this conceptual framework the focus of this dissertation study is the C-I linkage through dimensionality analysis of PISA mathematics items from 2003, 2006, and 2009.

In the context of PISA, the assessment framework for mathematical literacy corresponds to the cognition component of the assessment triangle. According to the OECD this mathematics assessment framework is very detailed and robust. The observation component refers to mathematics items developed to evaluate students' knowledge and skills in reference to this assessment framework. This includes careful design of items, developing scoring schema for each item, administering the items, and scoring student responses to the items. Lastly, interpretation encompasses producing the

final scores and making inferences about students' mathematical literacy in terms of the assessment framework and student, school, and country characteristics. Item Response Theory (IRT) and statistical methods employed to scale the results are parts of the interpretation component.

C-I linkage within PISA assessment design is the focus of this dissertation study. Statistical models and assumptions that are used to interpret student responses form the interpretation component. Their synchrony with the multidimensional aspects of the assessment framework implies the relationship between interpretation and cognition, i.e., C-I linkage. Through dimensionality analysis this relationship will be explored in this dissertation study. This linkage is important in studying and providing evidence for the validity of this assessment concept. The topic of validity is described later in the chapter.

Recommendations for Research, Policy and Practice

The NRC report on assessment design provides a contemporary theoretical framework assessment design. Whether in-class or large-scale, all types of assessment could and should fit in this framework to be accurate, efficient, and effective. The council also provides important recommendations and implications for research, policy, and practice on assessment design within this framework. Among a dozen recommendations, two are identified as the most relevant to this study.

The first one is about in-depth analysis of the critical components of the assessment triangle (cognition, observation, and interpretation.) There are various types of assessments ranging from in-class to large-scale, and from formative to summative designed in rigorous ways. One crucial aspect in the advancement of assessment design

entails studying the design and operational characteristics of these different types of assessments. The NRC urges for further in-depth analyses of exemplary and important assessment practices. “Important” means having serious impact on educational practices. In one way, it could be understood as high-stakes attached to assessments. Similarly, it could also be the case that it is the impact of results having directive role in changing educational policies that makes the assessment very important. In either case, the assessment becomes the key that might change lives. Therefore, it is very necessary to analyze any assessment from this assessment design perspective.

Analyzing an assessment from the NRC’s perspective on assessment design is a complex procedure rather than a single-step, straightforward one. First, the assessment should be reverse-engineered to its basic assumptions and underlying components. Secondly, these foundational components should be analyzed from the assessment triangle perspective. Analyzing the synchrony between the cognition, observation, and interpretation components of the assessment then follows the mapping to the assessment framework.

The second most relevant recommendation is to develop large-scale assessments that encompass a broad range of knowledge and competencies. Both lead to clear ways of reporting results. Research in efforts to improve large-scale assessments is worth both time and the resources to be invested (Chudowsky & Pellegrino, 2003). Assessment designers must reflect deeply on how and what large-scale assessments measure. Confirmatory analyses using responses to assessments reveal information about the degree of validity of assessments and help find ways to improve the assessment tests.

Validity: Concept and Sources

Validity is a fundamental concept in test development and evaluation. Validity is defined in *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p.10). Messick (1989) refers to validity in his influential book chapter as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences, interpretations and actions based on the test score" (p.13).

Although the literature has discussed validity from multiple perspectives (e.g., Borsboom, 2005; Cizek, Rosenberg, & Koons, 2008; Cronbach, 1971; Kane, 1992; Lissitz & Samuelsen; Loveinger, 1957; Messick, 1989) as an argument and as a concept, the definitions of validity given in the above paragraph are commonly accepted and agreed upon. This study is guided by these definitions. There are two important elements in defining validity, which are "evidentiary" notion and "theoretical" rationale. In addition, there seems to be an agreement on the nature of validity: it is of a single nature, of a unified concept (Messick, 1989). Validation is viewed as an ongoing process in which different sources of validity evidence are collected, summarized, analyzed, and evaluated (Cizek, Rosenberg, & Koons, 2008).

The other important aspect in validity is how or for what purposes these two elements, evidentiary and theoretical, are used. This is where evaluative, interpretational, and inferential notions of assessment tests come into play. Messick (1989) highlights the "actions" or practical consequences of the test as an inseparable from validity concept.

This argument, though, has not been left without criticism. Some thinks that consequential issues are central to assessment design but irrelevant to the concept of validity (Borsboom, 2005; Cizek, Rosenberg, & Koons, 2008). From the scientific point of view, construct validity represents the whole notion of validity and other types of validity are essentially *ad hoc* (Loveinger, 1957).

Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) identify five main sources of validity evidence: (1) assessment test content, (2) responses to assessment items, (3) internal structure of the assessment test, (4) relationship to external variables, and (5) consequences of test results. A framework of knowledge and skills that defines the mastery level in a domain is needed to organize the content of an assessment. Student responses to assessment items provide evidence for content and internal structure of the assessment (Loveinger, 1957). The internal structure of the assessment can provide evidence of the underlying cognitive model. Different external variables such as school, teacher, and student characteristics could be used to confirm the validity of an assessment for its intended purpose. Consequential validity (Messick, 1989) is an overarching look at the issues with practical consequences of an assessment. Expert opinion and professional judgment are needed when making decisions about what sources of evidence can best support the validity of assessments (AERA, APA, NCME, 1999).

This dissertation study takes on the evidentiary notion of the validity for PISA assessment. Internal structure (referred to as *dimensionality* in this dissertation) of PISA mathematics tests will be assessed. The results of dimensionality analysis of PISA will be

collected, summarized, and evaluated as one of five sources of validity evidence (AERA, APA, NCME, 1999). This work will contribute to validation (Cizek, Rosenberg, & Koons, 2008) of PISA assessment, providing evidence for its construct validity (Loevinger, 1957).

Dimensional Structure of Assessments

There are various methods for analyzing the dimensionality of tests. All serve as a validation process for tests, which are based on particular methodological and theoretical assumptions. The validation process is a scientific inquiry that should never be disregarded (Messick, 1989). For example, tests that are designed with Item Response Models require the test structure to be unidimensional. That is, the test is supposed to measure only one construct. Checking to see if responses to test items form a unidimensional structure is a part of the test validation process. Before getting into the methods of analyzing validity and test dimensionality, a little background on validity concept and different sources of validity evidence is needed.

Test Dimensionality: Definition and Concepts

Test dimensionality could be informally defined as "the minimum number of examinee abilities measured by the test items" (Tate, 2002, p.182). If assessment items form a unidimensional structure, then this set of items are said to be measuring one attribute of a construct. Dimensionality relates to central issues in development and use of large-scale assessments such as content validity, construct validity, score reliability, and test fairness. For example, unidimensionality is the basic assumption of measurement models (Hattie, 1985) and is required for construct validity (Rubio, Berg-Weger, & Tebb,

2001). Therefore, the dimensional structure of a test provides one type of validity evidence based upon the internal structure of a test. Loevinger (1957) claims that the construct validity is the only type of validity that is appropriate for tests and structural representation (i.e., dimensionality) of tests are central to construct validity. Some tests are designed to be unidimensional, while other tests are developed to measure several factors. However, it is sometimes the case that a test that is intended to be unidimensional may unintentionally be measuring more than one latent variable. Conversely, it is sometimes the case that some construct-irrelevant factors such as item types and formats could introduce multidimensionality to the assessment structure. Finally, it could be the case that the assessment is designed to be unidimensional but due to the planned content structure the assessment ends up multidimensional.

Item Response Theory (IRT) based tests rely on the assumption that the test structure is unidimensional. According to psychometricians, this assumption is always violated to some degree (Deng, Wells, & Hambleton, 2008; Tate 2002). On the other hand, the consequences of violating this assumption might have important implications on various phases of the test development process including comparisons across years and gathering validity evidence (Burg, 2007). However, the sources of multidimensionality are important to evaluate consequences. For example, the consequences of unidimensionality violation may not be as serious if the multidimensionality stems from the cognitive framework in the subject domain. When construct irrelevant factors cause the violation by introducing multidimensionality to the structure of the test, then consequences could be very complicated (Tate, 2002).

Assessing Test Dimensionality

Since the violation of assumptions made on dimensionality of an assessment has implications on the validity of items, and its consequences for interpretations of the assessment results are important, analyzing the dimensional structure of a set of items is a crucial task. Rigorous and careful assessment design and development process is necessary but may not be sufficient to produce an assessment structurally congruent to the planned framework. Thus, it is crucial to confirm the intended test structure empirically and to identify any construct-irrelevant sources of multidimensionality if any exist.

There is no standard form of test dimensionality procedures commonly accepted and used by researchers and psychometricians. Various methods have been used to test dimensionality including indices based on answer patterns, reliability indices, dimensionality fit statistics, principal component analyses, factor analyses, and structural equation modeling. Interestingly, studies comparing different procedures of assessing dimensionality have not concluded that one technique yields better results than others. Thus, no single method is considered preferable to the others; they all seem to be more or less equally useful. Researchers still debate and study in order to answer the question of how to best assess test dimensionality.

To provide a brief historical overview of techniques used for assessing dimensionality, it could be said that until the late 1980's various fit indices had been used as ad hoc procedures to confirm if a test was unidimensional or not. Hattie (1985) is conducted the last and the most comprehensive review of methods for unidimensionality

assessment. He reports 87 indices to test unidimensionality of a measure in his review. However, others criticized his review (see Tate, 2003), stating that it provided limited or no rationale and empirical support for the methods reported. Over the last two decades, though, methods for dimensionality assessment showed a great improvement. Several of the methods identified by Hattie (1985) have been further improved and used with the advances in technology. New methods emerged and simulation studies of the quality of proposed methods become possible with improving computer software. Some applications of several methods to real test data have been reported. However, most of the literature on assessing dimensionality during this time is relatively narrow, no more comprehensive than comparing some of the available methods at the most (Tate, 2003). So, there is not a comprehensive review that reports dimensionality assessment methods currently available with robust rationale and empirical support. Rather, studies on test dimensionality compare some of the methods currently available with each other (Burg, 2007; de Champlain, 1992; Deng, Well, & Hambleton, 2008; Tate, 2003; Wei, 2008).

To reiterate, assessing the dimensionality of a test could be done in many different ways. The methods for assessing dimensionality could be mainly categorized into two main families: parametric (linear, non-linear, IRT-based) and non-parametric (Tate, 2003). Review of each method and the rationale for each, as well as the purpose of commonly used methods are provided in chapter three. There are some considerations when selecting a model such as sample size and variable type (i.e., categorical vs. dichotomous) (Tate, 2002). These are discussed in detail to justify the model selection for this dissertation study in the methodology chapter (Chapter 3.

Studies on Test Dimensionality of PISA

There are various studies analyzing the test structure of achievement tests such as NAEP (e.g. Abedi, 1997; Burg, 2007; Griffo, 2011; Stone & Yeh, 2006; Wei, 2008; Zwick, 1987). However, there is a very limited number of studies to date that provide empirical evidence for dimensional structure of PISA items. Schwab (2007) investigated the relationship between the multidimensional structure of science as the cognitive domain and its assessment using IRT-based parametric techniques. She found that multidimensional models of the internal structure of the science items from PISA 2003 did not reflect the complex structure of PISA's cognitive framework for scientific literacy. Ekmekci and Carmona (2012) investigated the US students' responses to PISA 2003 mathematics items and found unidimensionality in the PISA 2003 mathematics items for the US population. Thus, the multidimensional structure of mathematical literacy detailed in the assessment framework was not reflected in the mathematics items. These are the only two studies exploring the dimensionality of PISA tests.

Somerville (2012) developed a new IRT-based method for differential item functioning (DIF) analysis as an extension of a generalized full-information item bifactor analysis model. He used PISA 2009 mathematics items to confirm the utility of his new model. He concluded that all but one mathematics items in 2009 showed insignificant DIF. This would mean that PISA 2009 mathematics items would be fair to different groups of students with similar education indices (Somerville, 2012)

There are a few other studies utilizing different variations of nonparametric DIF or LD studies in PISA context (e.g., Le, 2009; Yildirim & Berberoglu, 2009). To give a

few examples, Le (2009) investigated the relationships between gender differential item functioning (DIF) across countries and test languages for science items. He focused on different dimensions of science items: item format, focus, context, competency, and scientific knowledge. He found that gender DIF for science depended on item formats and content domains. Males were found to be more advantageous than females for some dimensions while the opposite is true for other dimensions. Yildirim and Berberoglu (2009) investigated PISA 2003 mathematics items for their fairness across different language and cultural groups. They concluded that cognitive skills measured by the items, translation errors, and use of quantitative words caused DIF.

Although findings from these studies provide a potentially valuable contribution to the development of tests for international use, none of them directly target the investigation into the validity of PISA by assessing the dimensionality of its items with the exception of Schwab's (2007) study, which focuses on science domain. No studies have been conducted to date to assess the dimensionality of PISA mathematics items.

Conclusion

As the review of literature shows there is a need to study the dimensionality of PISA's mathematical literacy assessment. This investigation is an important contribution to the study of its validity. Moreover, as the conceptual framework for this dissertation study demonstrates (see Figure 2.2), assessing dimensionality of PISA mathematics items is needed to understand the relationship between the important components (assessment triangle) of PISA assessment design for mathematical literacy. Prior studies have set the ground but have left a gap in assessing dimensionality of PISA. This study has the

potential to fill in this gap. The significance of this study comes from the need to provide evidence for validation process of PISA mathematical literacy assessment.

Chapter 3: Methodology

Data Sources

This study will entail a secondary-analysis of the dataset from the OECD's PISA database. The data includes student responses to individual mathematics items from the PISA 2003, 2006, and 2009 cycles.

As explained in the second chapter, PISA assesses mathematical literacy in a multidimensional way in terms of three important concepts: content, processes, and situations (or context). The first component, "content," is divided into 4 dimensions (overarching ideas): quantity, space and shape, change and relationships, and uncertainty. "Processes" consist of three competency clusters: reproduction, connections, and reflection. Lastly, "situations" are defined in terms of 4 dimensions: personal, educational/occupational, public, and scientific (OECD, 2009a). The number of items in each dimension by each implementation year is given in Table 3.1, Table 3.2, and Table 3.3.

For each cycle, the OECD provides data files at two levels: student-level and school-level. Student level data include two files: a cognitive item response data file, namely student responses for each item, and a student questionnaire data file. The student questionnaire is designed to collect information about their home, family, and school background. The school questionnaire is designed for school principals to provide information about various aspects including demographics of the school, staffing, environment, resources, and educational practices.

Identification variables for the country, school and student are common in all data files. The cognitive item response data file provides student responses for each item included in the test. The student questionnaire data file includes student responses for background questions; students' overall performance scores in mathematics, science, and reading; student weights; and country weights. This study will make use of cognitive item response and student questionnaire data files.

Instrument

According to the OECD (2009a), the mathematics domain portion in PISA is designed to explore students' capacity to analyze, reason, solve, and interpret mathematical problems in a variety of situations involving mathematical concepts such as quantity, spatial, probability, and change. The PISA assessment framework defines mathematical literacy as the individuals' capacity to identify and understand the role of mathematics in real life (OECD, 2003). A mathematically literate individual is expected to use mathematics in ways that meet the needs of their individual lives as a constructive, concerned, and reflective citizen. This definition of mathematical literacy underlies mathematical knowledge and skills that students posses and are able to utilize to solve problems they would encounter in their lives (OECD, 2009a).

PISA was first administered in 2000 in 32 countries (28 OECD, 4 non-OECD), and is implemented in all OECD and partner countries every three years. For each cycle, there is a domain that is emphasized by including a larger pool of items related to that particular content area, alternating between Reading, Mathematics, and Science. It started with reading as the major domain in 2003. Mathematics was the major domain in 2003

for the first time. PISA developed its mathematical literacy framework in a complete and comprehensive form for 2003 cycle. This is why 2003 is considered the earliest time point that the results from following cycles could be compared to (OECD, 2009b). 30 OECD and 11 non-OECD countries participated in PISA in 2003. Scientific literacy was the major domain in PISA 2006, in which 30 OECD and 27 non-OECD countries participated. In 2009, reading was the major domain again. 34 OECD and 41 non-OECD countries participated PISA 2009. The major domain of 2012 was mathematics but the results of this cycle won't be released until December 2013.

Table 3.1. *Number of mathematics items by content area and cycle*

	Cycles		
Content	2003*	2006**	2009***
Quantity	22	13	11
Space and Shape	20	11	8
Change and Relationships	22	13	9
Uncertainty	20	11	7
Total	84	48	35

Major domain is: * *Mathematics*; ** *Science*; *** *Reading.*

Mathematics Items by Content

In its assessment framework, PISA considers mathematical literacy from three perspectives: content, process, and context. Mathematical content is organized into four overarching ideas: quantity, space and shape, change and relationships, and uncertainty. The number of items in each content area by cycles is given in Table 3.1.

Table 3.2. *Number of mathematics items by process (competency cluster) and cycle*

	Cycles		
Process	2003*	2006**	2009***
Reproduction	26	11	9
Connections	39	24	18
Reflection	19	13	8
Total	84	48	35

Major domain is: * *Mathematics*; ** *Science*; *** *Reading.*

Mathematics Items by Process (Competency Cluster)

Process aspect, also known as competency clusters, is composed of reproduction, connections, and reflection sub-dimensions. PISA classifies underlying mathematical skills in these three competency clusters. The number of items in each competency cluster by cycles is given in Table 3.2.

Mathematics Items by Context

The third component, context or situations, consists of personal, educational and occupational, public, and scientific categories. Every mathematical problem is situated within a context in PISA. The number of items in each content area by cycles is given in

Appendix A provides the full classifications for individual items.

Item Formats

Mathematics assessment items are constructed in four different formats. Simple multiple-choice items have four responses from which students need to select the best answer. Complex multiple-choice items have several statements, each of which rquire

students to choose one of several possible responses (e.g., yes/no, true/false, correct/incorrect). Short closed-constructed response items require constructing a numeric response within very limited constraints, or only require a word or short phrase as the answer. Open-constructed extended response items require more extensive writing and often require some explanation or justification. All of multiple-choice items are scored as credit/no-credit basis. Scoring of a few of the short response and open response items allow for partial credits, which are worth half of a full credit with the exception of one item where two partial credits (worth as one third and two thirds of the full credit) are given.

Table 3.3. *Number of mathematics items by context and cycle*

	Cycles		
Context	2003*	2006**	2009***
Personal	18	9	4
Educational and Occupational	20	8	5
Public	28	18	13
Scientific	18	13	13
Total	84	48	35

Major domain is: * *Mathematics*; ** *Science*; *** *Reading.*

In 2003, there were a total of 85 mathematics items included in the test. The PISA Governing Board (PGB) had to exclude one item from the results because of some technical problems with it. So, 84 items remained. Of these, 18 (21%), 11 (13%), 41

(49%), and 14 (17%) are in simple multiple-choice, complex multiple-choice, short closed-constructed, and open-constructed extended response format, respectively.

In the years 2006 and 2009, no new mathematics items were developed. So, PGB selected items for these cycles from 84-item pool used in 2003. PISA 2006 test included a total of 48 mathematics items. The number (and percentages) of different item formats is 13 (27%), 9 (19%), 21 (44%), and 5 (10%) for simple multiple-choice, complex multiple-choice, short closed-constructed, and open-constructed extended response items, respectively.

In 2009, the number of mathematics items included in the test was 35 all of which were also used in both the 2006 and 2003 cycles. The break-down of these items in terms of item format is 10 (29%), 7 (20%), 16 (46%), and 2 (5%) for simple multiple-choice, complex multiple-choice, short closed-constructed, and open-constructed extended response items, respectively.

Participants

Respondents of the student questionnaire and cognitive item test are 15-year-old students from these countries that administered the PISA test in their educational systems in 2003, 2006, and 2009. Tests are typically administered to between 4,500 and 10,000 students in each country. The sampling of 15-year olds is two-stage stratified sample to ensure the appropriate representativeness. First, individual schools that are eligible are chosen with probabilities that are proportional to a measure of size. In the second stage, 35 students that are 15 years old are selected from sampled schools with equal

probability. In schools with less than 35 15-year olds, all of students are selected unless that number is less than 20.

41 countries in total (30 OECD, 11 non-OECD) participated in PISA-2003 with a total number of students of about 275,000. In 2006, when the major domain was scientific literacy, 57 countries participated in PISA. 30 of these were OECD countries and the remaining 27 countries were non-OECD partners. Nearly 400,000 students took the PISA test in 2006. In the last cycle of PISA, 2009, the main domain was in reading literacy. Of 75 participating countries, 34 were OECD members and 41 were non-OECD partner countries. About 520,000 students were given the test.

This dissertation study aims to analyze the dimensionality of the PISA mathematics items in different cycles. This longitudinal aspect (Carmona et al., 2011) of this dissertation will explore if and in what ways the dimensional structure of mathematics items would demonstrate difference and similarities at different time points. This longitudinal aspect requires that the same student profile needs to be kept across different cycles in order to control for external variables related to the economy of participating countries. Thus, the data for this dissertation will include students from the OECD countries that have participated in all three cycles.

That is about 200,000 students from 30 countries for each of the cycles. The 30 OECD countries included in the data are Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, New Zealand, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, The Netherlands, Turkey, and United States.

Analysis

This section starts with a review of methods for assessing test dimensionality. Then, some considerations for model selection follow. Next, the statistical methods to be utilized in this study are specified, including rationale for model selection. Lastly, the relationship between methods of analyses and research questions are briefly discussed.

Methods for Assessing Test Dimensionality

Assessing the dimensionality of a test can be done in many different ways. The goal of this section is to provide a brief overview of some of the more popular procedures that are available today for the empirical assessment of the dimensional structure of a test. The methods could be mainly categorized into two broad classes: parametric and non-parametric (Tate, 2003). Procedures for assessing dimensionality either belong to a family of parametric models or that of nonparametric.

Parametric and nonparametric methods are distinguished by their specification of the item response function (IRF). In item response theory, the probability of success on the item i is usually presented by the IRF $\Pr(X_i = 1, \xi \mid \theta)$. Parametric methods assume a prescriptive parsimonious model for the IRF whereas nonparametric ones only assume a monotonic IRF that gives freedom from dependence on a particular model (Burg, 2007). In practical terms, however, it could be said that nonparametric models are a way of checking if the structure of a test is unidimensional or not. In other words, nonparametric models offer a confirmatory hypothesis testing with the null hypothesis that the one-factor model fits the test structure the best (Tate, 2002). See Table 3.4 for a list of parametric and nonparametric methods of assessing test dimensionality.

Table 3.4. *Parametric and nonparametric procedures of test dimensionality assessment*

Parametric Methods
Linear Factor Analytic Methods
Principal Component Analysis (PCA) (e.g., Zwick, 1987)
Exploratory Factor Analysis (EFA) (e.g., Mislevy, 1986)
Confirmatory Factor Analysis (CFA) (e.g., Muthén, 1993)
Structural Equation Modeling (SEM) (e.g., Rubio, Berg-Weger, & Tebb, 2001)
Item Factor Analytic Methods (Nonlinear)
Nonlinear item factor analysis (e.g., Hambleton & Rovinelli, 1986)
Full-Information factor analysis (Bock, Gibbons, & Muraki, 1988)
IRT-based Methods
Local item dependencies (e.g., Yen 1993)
Nonparametric Methods
Hierarchical cluster analysis of item proximities (Roussos, Stout, & Marden, 1998)
Test of essential dimensionality (Stout, 1987)
DETECT index of dimensionality (Zhang & Stout, 1999)

Parametric procedures

Parametric methods offer a parsimonious and quantitative description of data structure (Tate, 2003). Parametric methods have various types including parallel analyses of principal components (Principal Components Analysis – PCA), classical factor analysis (FA), item factor analyses (modified version of classical FA), and IRT-based approaches.

Linear methods: Methods based on classical factor analytic methods refer to the traditional, linear FA using covariance matrices. Classical FA and PCA for test items uses ϕ (phi) or tetrachoric correlations unlike Pearson's product-moment correlation coefficient r for continuous variables. The goal of FA and PCA are similar: to determine the latent structure underlying a set of variables. However, PCA and factor analyses are not the same. Although the differences between the two have been long discussed and are important, because of space limitation here, it would be enough to say that PCA analyzes variance while FA analyzes covariance (Burg, 2007). There are many questions and concerns related to the use of PCA and FA as tools for assessing dimensionality, including the appropriateness of these methods for dichotomous data and the criterion for determining how many factors or principal components to extract (Abedi, 1997). Studies on FA and PCA seem to unequivocally find the use of indices based on them very problematic for assessing test dimensionality (for the details of those studies, see de Champlain, 1992)

Procedures based on exploratory factor analysis (EFA) have long been used to analyze the structure of measures by determining the number of factors present in a set of items. However, researchers have found that EFA does not afford testing models with high-order factors and that EFA often times underfactor (Anderson & Gerbing, 1988; Rubio, Berg-Weger, & Tebb, 2001). Confirmatory factor analysis (CFA) is used when there is a prior expectation about the structure of a test. When a hypothesized model fails to fit the response structure of the test, it is possible to determine where the model failed and to find the appropriate model that fits the data best (Tate, 2002). CFA is considered

to be a special case of structural equation modeling (SEM). SEM permits testing various models of the structure of a set of items, developing stronger models, and establishing higher order factors previously not possible (Rubio, Berg-Weger, & Tebb, 2001).

Nonlinear methods: Item factor analytic approaches are an extension of classical FA and use a nonlinear relationship between the probability of a correct response to an item and examinee abilities (or other latent factors). Nonlinear item factor analysis and full-information factor analysis are two types of item factor analysis. Both approaches use summary information such as proportions and correlations to explore the relationship between the item responses and the latent factors (Burg, 2007). The full-information model additionally uses all the information available from the entire response matrix rather than just the covariance or correlation matrix (Tate, 2003).

IRT-based methods: In addition to linear and nonlinear factor analytic procedures for assessing test dimensionality, there are also IRT-based parametric methods. The dimensional structure of a test can be thought of in terms of conditional independence (local item independence) in IRT. When conditional independence is violated, it means local (item) dependences (LD) are present. Chen and Thissen (1997) proposed four statistics for detecting LD among items using IRT including X^2 and G^2LD. These methods were found to be useful when testing a relatively small number of selected item pairs and when searching for any problematic pairs of items by identifying outliers in the distribution of all conditional item associations (Tate, 2002).

Nonparametric procedures

Nonparametric procedures for assessing test dimensionality are very useful in situations with small number of items and examinees. They also have the potential to work in some cases where parametric IRT models fail to provide useful information (Tate, 2003). As also stated previously, nonparametric models assume that the IRF is monotonic, offering the freedom from dependence on highly prescriptive models, unlike parametric approaches (Burg, 2007). In other words, nonparametric models do not have to estimate model parameters or be constrained by model specificity since they do not use IRT models. Using a nonparametric method also eliminates the confusion with lack of model-fit by a particular unidimensional parametric family of models when working with potentially multidimensional data (Stout, 1990).

Stout's (1987) DIMTEST, Zhang and Stout's (1999) dimensionality evaluation to enumerate contributing traits (DETECT), and hierarchical cluster analysis (HCA/CCPROX) (Roussos, Stout, & Marden, 1998) procedures are among commonly used nonparametric techniques. All of these three approaches are based on conditional item associations, also known as local item dependencies (LD). They utilize nonparametric computations of conditional item covariances (de Champlain, 1992). For each item pair, the examinees are first stratified according to number of their correct responses on the rest of test items. Then, the covariance of responses for each pair is computed in each stratified group. Lastly, averages of group values are computed to obtain the final conditional item covariance (Roussos, Stout, & Marden, 1998; Stout, 1990; Zhang & Stout, 1999).

Each of the three methods addresses a different aspect of test structure such as essential dimensionality (Stout, 1990), approximate multidimensional structure (Zhang & Stout, 1999), and approximate simple structure (Roussos, Stout, & Marden, 1998). All together they could provide a complete summary of dimensional characteristics of a test (Stout et al., 1996). As also mentioned above, the strength of nonparametric methods over parametric procedures is the freedom from strong assumption of a particular prescriptive model. However, nonparametric procedures "will not provide mathematical models of multidimensional tests" (Tate, 2002, p.201). More simply stated, nonparametric models are a way of checking if the structure of a test is unidimensional or not.

Some Considerations on Method Selection

There is no standard procedure for assessing test dimensionality. There are various methods that have been used in the literature. So, how do we know which method should be used? There are some considerations that should be kept in mind when selecting a method for a specific situation. The following considerations have been given by researchers as a guide in model selection: response format of test items (i.e., dichotomous, categorical), sample size, number of items, existence of a prior expectation of multidimensionality based on the content structure, and the intention of identifying source of unintended multidimensionality (Tate, 2002).

When there is a strong prior expectation about the structure of a test, methods that are based on confirmatory factor analysis, which is considered a special case of structural equation modeling (SEM), are considered the most appropriate techniques (Kline, 2010;

Tate, 2002). However, most of confirmatory methods demand large sample sizes for achievement tests since these methods were originally developed for continuous variables (Joreskog, 1990). In achievement tests, responses to items are either dichotomous or polytomous. The Linear Structural Relations (LISREL) method, for example, requires that the number of examinees should be at least three to five times the number of correlations between items. That is, for a test of 50 items, which means about 600 correlations, there need to be about 2000-3000 examinees in the dataset (Tate, 2002).

Structural Equation Modeling (SEM)

I will investigate the mathematical literacy dimensions of PISA test by drawing on the mathematics items from 2003, 2006, and 2009 cycles. Each mathematics item is treated as a variable in this study. That is, 84, 48, and 35 variables for 2003, 2006, and 2009, respectively.

Since the cognitive assessment framework of PISA mathematics has a multidimensional content structure, there is a strong prior expectation on the PISA mathematics items to be multidimensional. Therefore, confirmatory item factor analytic procedures, a structural equation modeling (SEM) (Byrne, 2011; Kline, 2010) technique, should be utilized to explore if the multidimensional model best fits the response data for mathematics items. This study will then employ confirmatory factor analyses (CFA) to test dimensionality of mathematics items in PISA tests. In CFA, the number of factors and the relationship of factors to all the measures are hypothesized beforehand. In other words, CFA is a hypothesis testing method, unlike data reduction methods such as principal components analysis. Furthermore, as an SEM technique, CFA allows testing

multiple hypothesized models at the same time, and this is a big advantage. The models specify the degree of correlation between the common factors and which of the unique factors will be correlated.

The models for this study are constructed based on the OECD's framework for mathematical literacy. The competing models will be tested through CFA analyses to determine which model fits the data. Goodness of fit indices (GFIs) such as comparative fit index (CFI), the Tucker-Lewis index (TLI), and root mean square error of approximation (RMSEA), provide measures to determine which model best explains the relationship between observed variables and latent factors (Rubio, Berg-Weger, & Tebb, 2001).

Another advantage of SEM is that different factor structures including hierarchical models (higher order models) and oblique (correlated) models can be compared (Anderson & Gerbing, 1988). Second-order models are used to test the factor structure when the domain specific factors are related with each other and when there is a priori hypothesis that a higher-order or a second-order factor can account for the relationship between the lower order factors (Rubio, Berg-Weger, & Tebb, 2001).

Data Analyses and Hypotheses

The initial data analysis will be conducted to produce descriptive statistics including means, standard deviations, and product-moment correlation indices for the mathematics item for each cycle. Frequencies of proficiency levels for mathematics domain and the overall levels will be computed to explain the distribution of data.

Next, seven SEM models will be produced and compared to each other to test the hypotheses about the dimensional structure of PISA mathematics items for three different cycles. These are: single factor model (interpreting the general mathematical literacy as the only latent factor), four-factor content model, three-factor process model, four-factor context model, higher order content model, higher order process model, and higher order context model. To summarize, there is one single-factor model, three correlated-factors models, and three higher-order factor models.

As mentioned above, all of these models are constructed according to two important elements. The first one is the PISA mathematical literacy framework. It has to do with the factors that are expected to measure a construct and the proposed dimensions. PISA proposed three dimensions for mathematical literacy, each of which is constructed by three to four factors (sub-dimensions). This structure is described in detail earlier. However, Table 3.5 provides a good summary.

Secondly, the trends for confirmatory factors analytic procedures from the literature on dimensionality assessment of tests suggest two levels of parsimony: correlated factors or connected factors through a higher order construct (Anderson & Gerbing, 1988; Rubio, Berg-Weger, & Tebb, 2001). Two levels of parsimony for each dimension totals to six different models. Unidimensional models needs to be included by default for many reasons including but not limited to unidimensional assumption for IRT (e.g., Deng, Wells, & Hambleton, 2008; Tate 2002).

Each of the seven models but the single-level model corresponds to one of three dimensions of PISA assessment framework: content, process, and context. Thus, the

rationale for these models and related hypotheses draw on PISA assessment framework for mathematical literacy. Remaining of this chapter is devoted to description of these seven models, their corresponding hypotheses, and their relationship to research questions.

Table 3.5. *Dimensions of pisa mathematics items*

Dimensions →	**Content**	**Process (Competency)**	**Context (Situations)**
Sub-dimensions →	Quantity	Reproduction	Personal
	Space and Shape	Connections	Educational / Occupational
	Change and Relationship	Reflection	Public
	Uncertainty	-	Scientific

Single-factor model (Model 1)

The single factor model (referred to as Model 1 in this dissertation) does not reflect the structure illustrated in table 3.5. Rather, it attributes every factor to a general latent variable. Correlated factors models treat item responses to be structured according to the sub-dimensions shown in table 3.5. Higher order factor models explore the relationships of sub-dimensions to their higher order dimension and to each other through the higher order dimension.

The single factor model will test the hypothesis that there is only one-factor (general mathematics literacy - GML) that represents all mathematics items regardless of

their content, process, and context. Figure 3.1 illustrates this model for 2009 PISA mathematics items.

1-level correlated factors models (Models 2-4)

The four-factor content model (referred to as Model 2 in this dissertation) tests the hypothesis that there are four content factors correlated to each other. These content factors, according to PISA assessment framework, are quantity (QT), space and shape (SS), change and relationship (CR), and uncertainty (UN). For example, in 2009, eleven items are expected to load on quantity factor, eight on space and shape, nine on change and relationship, and seven on uncertainty. Figure 3.2 shows the model specification for 2009 items.

The three-factor process model (referred to as Model 3 in this dissertation) tests the hypothesis that there are three process factors correlated to each other. These process factors, according to PISA assessment framework, are reproduction (REP), connections (CON), and reflection (REF). To illustrate, nine items are expected to load on reproduction factor, 18 on connections, and eight on reflection in 2009 cycle. The model specification for 2009 items is similar to Figure 3.2.

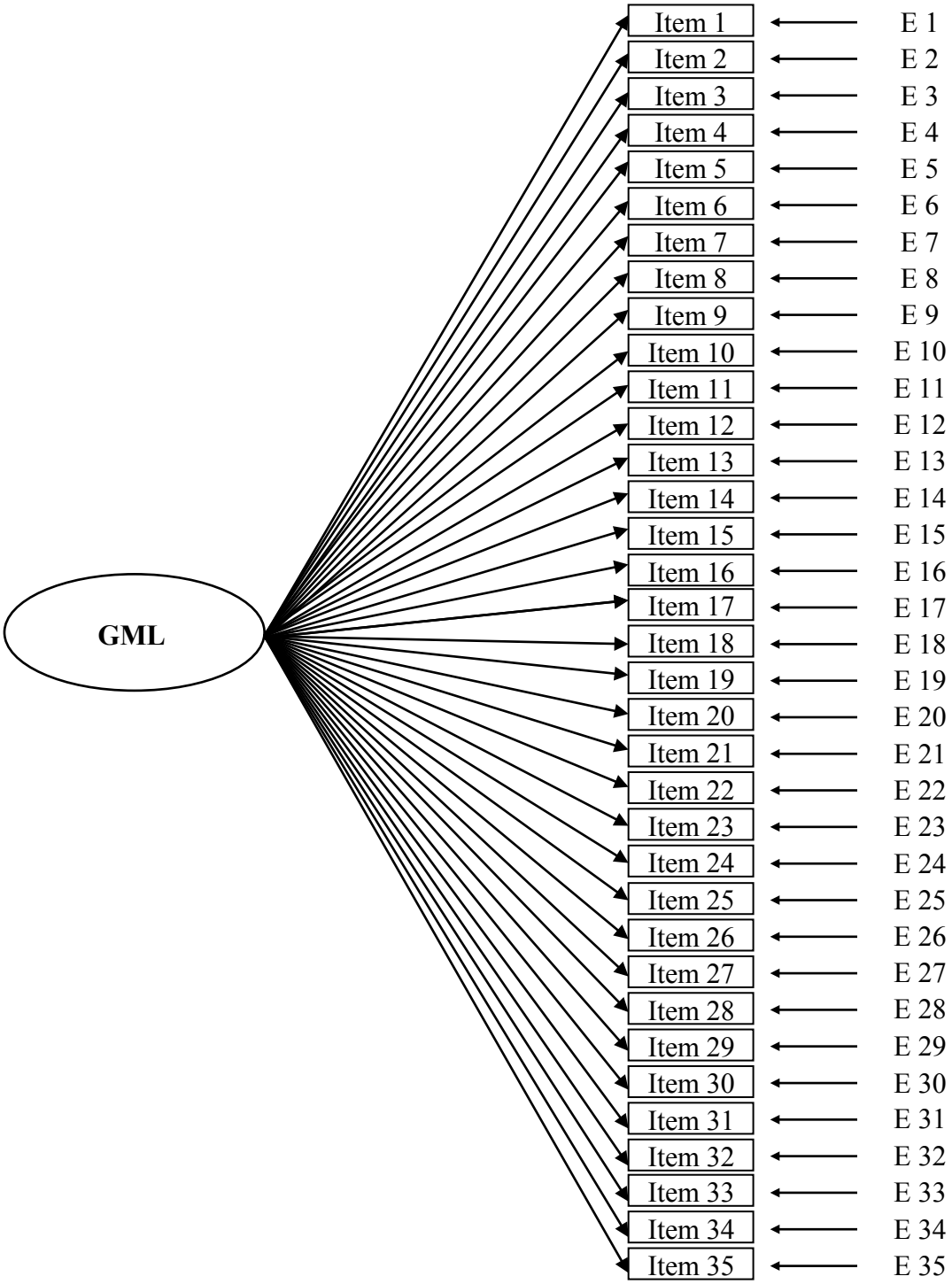


Figure 3.1. Single Factor Model (Model 1)

GML: General Mathematical Literacy, E: Error Term

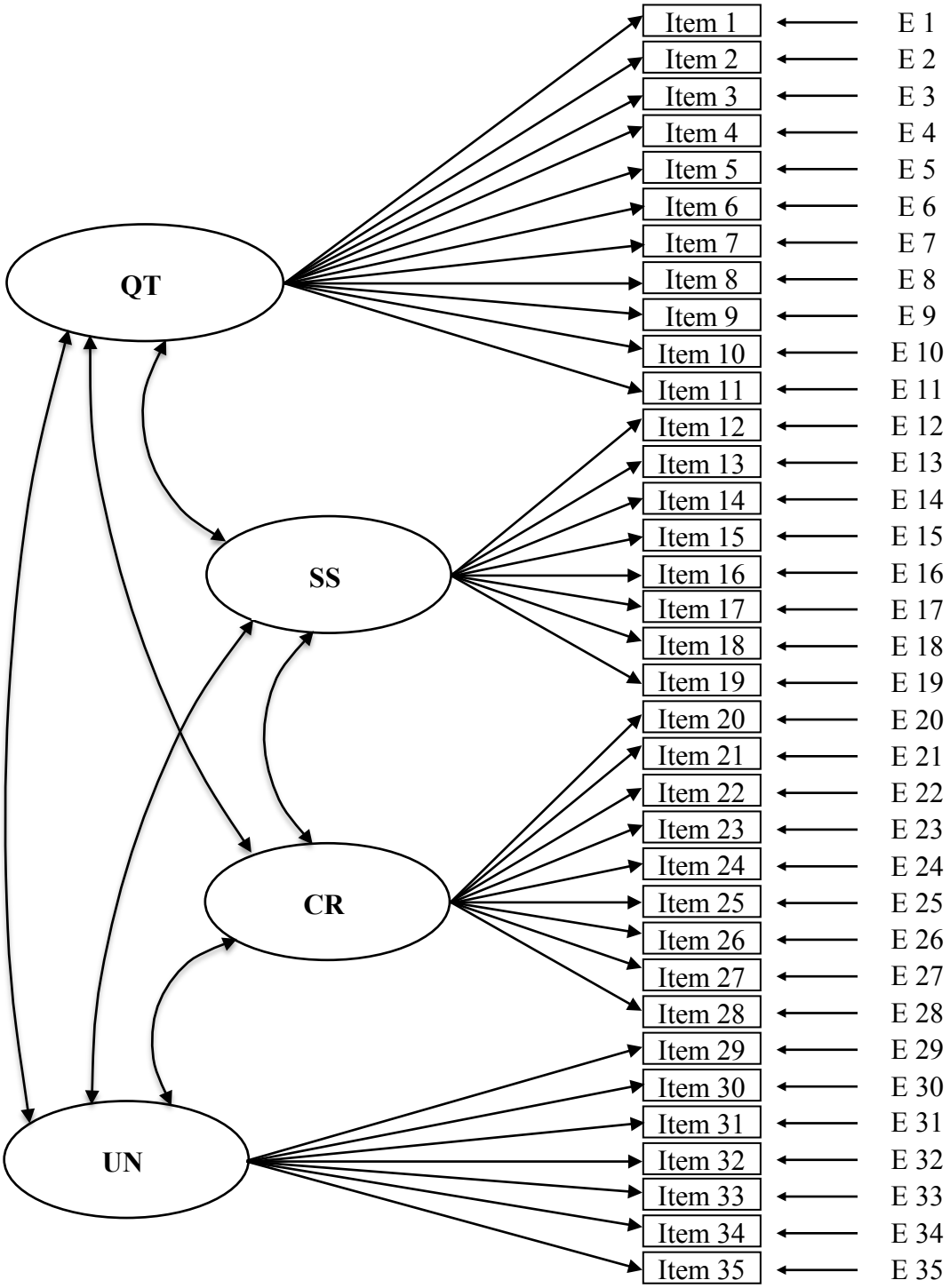


Figure 3.2. Four-Factor Content Model (Model 2)

QT: Quantity, SS: Space & Shape, CR: Change & Relationship, UN: Uncertainty, E: Error Term

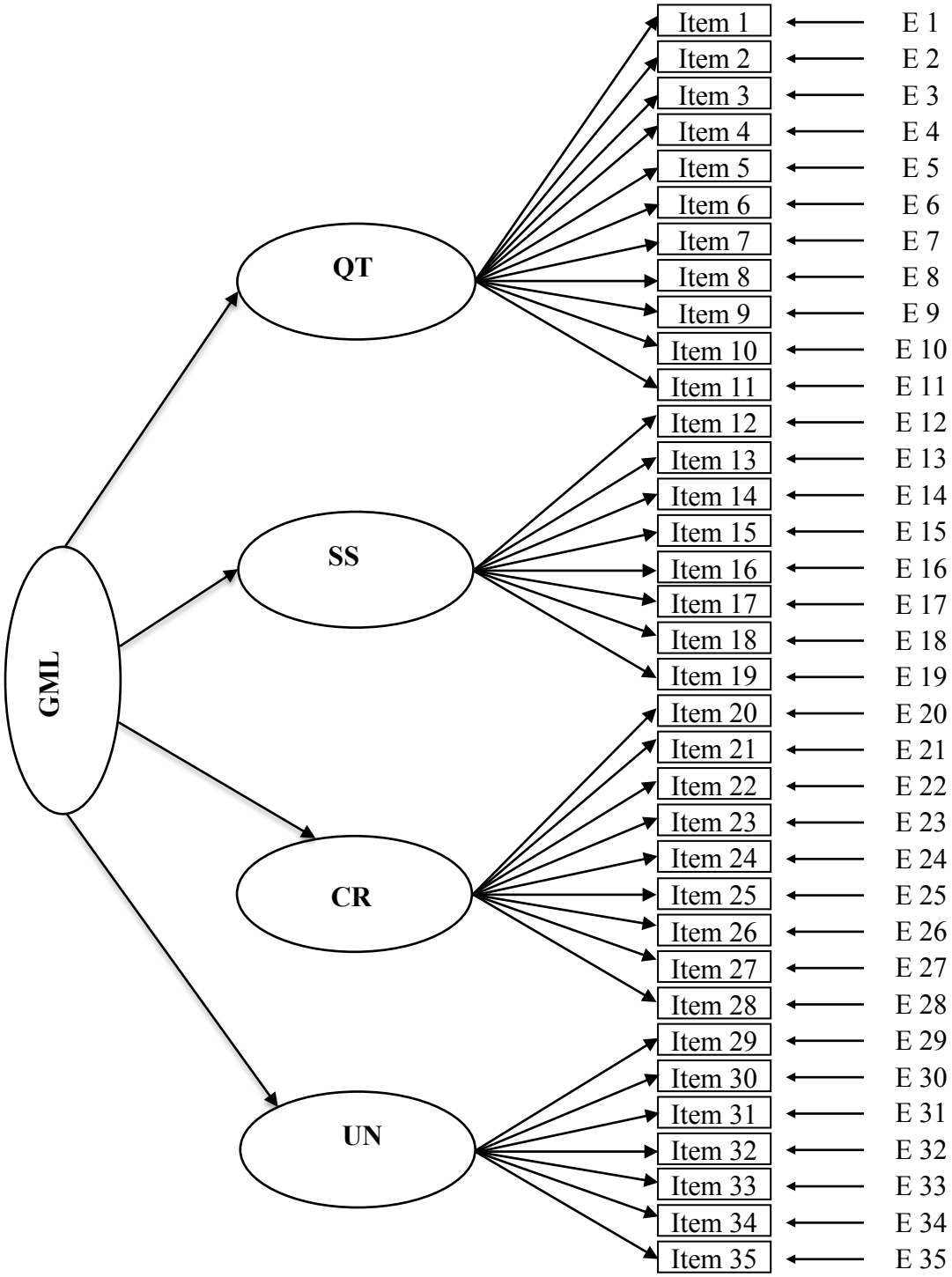


Figure 3.3. 2-Level Content Model (Model 5)

GML: General Mathematical Literacy, QT: Quantity, SS: Space & Shape, CR: Change & Relationship, UN: Uncertainty, E: Error Term.

The four-factor context model (referred to as Model 4 in this dissertation) tests the hypothesis that there are four context factors correlated to each other. These context factors, according to PISA assessment framework, are personal (PER), educational/occupational (EDO), public (PBL), and scientific (SCI). For example, in 2009, four items are expected to load on personal factor, five on educational/occupational, 13 on public, and 13 on scientific. The model specification for 2009 items is similar to Figure 3.2.

Higher order factor (2-level) models (Models 5-7)

Higher order factor models assume that there is a general mathematical literacy ability that accounts for the relationship between sub dimensions. The Model 5 is a second-order hierarchical model that will test the hypothesis that there is a general mathematics literacy (GML) factor (a content-wise one) at the second level that accounts for the relationship between the four content domains of quantity (QT), space and shape (SS), change and relationship (CR), and uncertainty (UN) as level-1 latent variables. Figure 3.3 demonstrates the structure of this model.

The Model 6 tests the hypothesis that a general mathematics competency factor accounts for the relationship between the three competency clusters: reproduction, connections, and reflection. The model specification for 2009 items is similar to Figure 3.3.

Finally, the Model 7 tests the hypothesis that a general mathematics literacy factor in terms of context (or situations) accounts for the relationship between the four different

context sub-dimensions of personal, educational/occupational, public, and scientific. The model specification for 2009 items is similar to Figure 3.3.

Relationship of Models to Research Questions

Goodness of fit indices for models will help answer the first research question: what is the correspondence between the dimensional structure of the PISA mathematics items and PISA's mathematical literacy assessment framework in terms of the content, process, and context dimensions? Individual parameter estimates will demonstrate how much of the variation in responses to each item could be explained by individual models. First, 2003 items will be evaluated in terms of their model-fit. If the single factor model fits 2003 items well enough, then, it means that there is evidence supporting the unidimensional structure of 2003 mathematics items. This would mean, one of the basic assumptions (unidimensionality) for an IRT-based set of items would be supported. If other models also fit well enough to 2003 data, then it means that the response structure of mathematics items is in accordance with its multidimensional assessment framework. However, comparison of the models will reveal which model fits the data best. These results would either support a relaxation in the unidimensionality assumption (Stout, 1990; Tate, 2002) or weakness in the assessment design (NRC, 2001).

Comparing models to each other will reveal the model(s), which represents the structure of the mathematics items the best in terms of different dimensions. The model comparison results, thus, will answer the second research question: what is the best representation for the dimensional structure of the PISA mathematics items for implementation cycles 2003, 2006, and 2009?

Conducting the same analyses for 2006 and 2009 data will bring the longitudinal perspective (Carmona et al., 2011) and will help understand if the trends for dimensional structure change over time. Keeping the student profile the same as much as possible by focusing only on OECD countries participated in all of three cycles, dimensionality of mathematics items are expected to be the same in all of three cycles. This is because PISA mathematics assessment design is supposed to be consistent across cycles. Therefore, it is hypothesized that the dimensionality of mathematics items among different cycles is stable. The longitudinal aspect will explore if this is the case. Evaluating the results for different cycles all together will answer the third research question: how does the dimensional structure of the PISA mathematics items change over time?

Formal Hypotheses for CFA Models

Formally stated, unidimensional model (Model 1) hypothesizes that the dimensional structure of the PISA mathematics items could be explained by a general latent factor called general mathematical literacy (GML); each mathematics item has a nonzero loading onto GML; and the residuals associated with each indicator item variable are uncorrelated.

The 1-Level content model (Model 2) hypothesizes that the dimensional structure of the PISA mathematics items could be explained by four content related mathematical literacy (ML) factors: QT, SS, CR, and UN; each mathematics item has a nonzero loading onto the content factor it was designed to measure, and zero loadings on all other

factors; the four content factors are correlated; and the residuals associated with each indicator item variable are uncorrelated.

The 1-Level process model (Model 3) hypothesizes that the dimensional structure of the PISA mathematics items could be explained by three process related ML factors: REP, REF, and CON; each mathematics item has a nonzero loading onto the process factor it was designed to measure, and zero loadings on all other factors; the four process factors are correlated; and the residuals associated with each indicator item variable are uncorrelated.

The 1-Level context model (Model 4) hypothesizes that the dimensional structure of the PISA mathematics items could be explained by four context related ML factors: PER, EDO, PBL, and SCI; each mathematics item has a nonzero loading onto the context factor it was designed to measure, and zero loadings on all other factors; the four context factors are correlated; and the residuals associated with each indicator item variable are uncorrelated.

The 2-Level content model (Model 5) hypothesizes that the dimensional structure of the PISA mathematics items could be explained by four first-order content related factors (QT, SS, CR, and UN) and one second-order factor (GML); each mathematics item has a nonzero loading onto the first-order content factor it was designed to measure and zero loadings on all other first-order factors; co-variation among the four first-order content factors is explained fully by their regression on the second order factor GML; and the residuals associated with each indicator item variable are uncorrelated.

The 2-Level process model (Model 6) hypothesizes that the dimensional structure of the PISA mathematics items could be explained by three first-order process related factors (REP, CON, and REF) and one second-order factor (GML); each mathematics item has a nonzero loading onto the first-order process factor it was designed to measure and zero loadings on all other first-order factors; co-variation among the three first-order process factors is explained fully by their regression on the second order factor GML; and the residuals associated with each indicator item variable are uncorrelated.

The 2-Level context model (Model 7) hypothesizes that the dimensional structure of the PISA mathematics items could be explained by four first-order context related factors (PER, EDO, PBL, and SCI) and one second-order factor (GML); each mathematics item has a nonzero loading onto the first-order context factor it was designed to measure and zero loadings on all other first-order factors; co-variation among the four first-order context factors is explained fully by their regression on the second order factor GML; and the residuals associated with each indicator item variable are uncorrelated.

Summary

This chapter introduced the instrument developed and used by PISA to assess mathematical literacy all around the world. The second chapter already explained the PISA's assessment framework for mathematical literacy and its dimensional structure. This chapter built on that and proposed seven models to reflect dimensional structure of responses to mathematics items across cycles. These models are: single factor model (Model 1), four-factor content model (Model 2), three-factor process model (Model 3),

four-factor context model (Model 4), higher order content model (Model 5), higher order process model (Model 6), and higher order context model (Model 7). As mentioned above, these models are based on multidimensional definition of mathematical literacy proposed in its by PISA.

CFA is chosen as the methods of inquiry for the dimensionality assessment of PISA mathematics items. A detailed review of methods for dimensionality analysis and criteria for model selection are provided above. To briefly mention again, the prior expectation for multidimensional structure and the goal of comparing different models based on three different dimensions suggest confirmatory item factor analysis using SEM. Seven models are proposed according to dimensions of PISA mathematics items and level of parsimony for confirmatory factor analytic procedures in the literature. Their related hypotheses are formed in reference to PISA assessment framework for mathematical literacy.

Lastly, the dimensionality analyses will be conducted for all of three cycles: 2003, 2006, and 2009. That is, each of seven CFA models will be evaluated for each cycle to obtain the longitudinal picture for response structure of mathematics items. Eventually, the results will address the following research questions:

1. What is the correspondence between the dimensional structure of the PISA mathematics items and PISA's mathematical literacy assessment framework in terms of the content, process, and context dimensions?
2. What is the best representation for the dimensional structure of the PISA mathematics items for implementation cycles 2003, 2006, and 2009?

3. How does the dimensional structure of the PISA mathematics items change over time?

With respect to the first research question, the correspondence between the student responses and the assessment framework is expected to match for each of the three dimensions to provide evidence for construct validity of the PISA assessment. That is, multidimensionality is expected in the response data. The first research question, thus, uncovers how different structural models that are proposed with respect to the PISA mathematical literacy framework would fit the student responses to PISA mathematics items. It is likely that more than one model could fit the response data. Actually, it is hypothesized that all models should fit data well enough because they all have sound rationales on which they are based. However, it is not known (nor is it predictable) whether some items look good in some of the models and not in the others. Thus, the first question investigates the individual item behavior in different models as well overall model-fit.

The second research question explores the models that represent the dimensionality of response data the best. Ekmekci and Carmona (2012) and Schwab (2007) detected unidimensionality in U.S. students' responses to PISA 2003 mathematics and science items. Therefore, it is hypothesized that unidimensional models would have the best fit to the response data. If this is the case, student responses as signs and samples of the construct (Loveinger, 1957), mathematical literacy, which is the central piece of this dissertation, imply that mathematical literacy is a unidimensional construct.

The third research question investigates whether the PISA mathematics assessment has stability in terms of dimensional structure. It is expected that that model comparison results would be stable across different cycles. However, what is unknown is how the variation in the responses to individual items changes over time.

Chapter 4: Results

The first part of the results chapter begins with addressing the issues of sample selection and student weights. Following this section is a summary about statistics, indices, and parameter estimates follow, which are used to evaluate models and their comparisons, and their cut-off values along with what these values mean. This summary also serves as an introduction to cycle-by-cycle presentation of results. For each PISA cycle, the results are presented in the following order: model-fit indices, individual item parameters, and model comparisons. Results for model-fit indices and individual parameter estimates will address the first research question: What is the correspondence between the dimensional structure of the PISA mathematics items psychometrically and PISA's mathematical literacy assessment framework in terms of the content, process, and context dimensions?

Model comparison results in each cycle explore the models that represent the student response data the best in terms of different dimensions, addressing the second research question: What is the best representation for the dimensional structure of the PISA mathematics items for implementation cycles 2003, 2006, and 2009?

Finally, looking across the cycles to see how different models change over time provides the longitudinal aspect and addresses the last research question: How does the dimensional structure of the PISA mathematics items change over time?

A summary of results for each cycle at the end of each section as well as an overall summary at the end of the chapter evaluating overall results and their implications is provided.

Random Sampling and Sampling Weights

This section provides information on sample size selection and issues about random sampling and student weights. These issues are important to address because the accuracy and generalization of the results rely on handling them appropriately.

The total number of respondents for each PISA cycle exceeds 200,000. Including the whole population in the analysis produces a very large sample size that increases the power of the chi-square test for model-fit, resulting in significant values regardless of the model-fit. That is, no matter what the goodness of the model-fit is in reality, the results would imply that the model does not fit the data. So including such a large sample size opens the possibility of making a Type-I error (incorrect rejection of the null hypothesis). Thus, the sample size for this study was reduced through random sampling of the whole population.

The ideal sample size for this study is found to be around 17,000. There are three considerations in reaching this sample size. The first one is the criterion provided by Tate (2002) for dichotomous items: a sample size of at least three to five times the number of correlations between items is required. The sample sizes for this study should be as given in Table 4.1 below according to this criterion.

Secondly, according to PISA, 17,000 is a good sample size for multivariate analysis (OECD, 2009a). In fact, PISA randomly took 500 observations (respondents) from each of 30 OECD countries when the initial item calibrations were performed. Therefore, 15,000 observations allow every school within the same country to contribute to the sample. The sample size for this study then needs to be at least 15,000 to comply

with OECD's criterion. Lastly, sample sizes less than 17,000 produced incomplete matrices for CFA calculations (empty cells in bivariate tables for some item pairs). Therefore, 17,000 is taken as the minimum sample size for each cycle.

Table 4.1. *Sample sizes for each cycle*

	Number of items	**Number of correlations between items**	**Sample size**
2003	84	3,486	10,458 – 17,430
2006	48	1,128	3,384 – 5,640
2009	35	595	1,785 – 2,975

There are two issues, however, that should be addressed with use of random samples from the PISA population. The first one is related to OECD countries that have the greatest number of schools. Canada, Mexico, and Switzerland had more than 500 hundred schools participate in 2003. Similarly, Canada, Italy, Mexico, Spain, and UK had around 500 or more schools participate in the 2006 and 2009 cycles. The issue is whether a random sampling of about 500 students from each country ensures enough diversity. In other words, is the reduced sample an accurate representation of all 15-year-olds within the same country? In the aforementioned countries, it is very likely that students from some schools may not be represented in the sample. However, between-school variance for those countries is relatively very low (OECD, 2009a). Student profile, student performance, and educational characteristics are similar across all schools in these countries. Instead, the variation among students mainly lies at the within-school level. Therefore, having a smaller number of observations than the number of participating

schools is not expected to introduce bias to the random sampling.

The second issue regarding the use of random sampling is PISA's sample design. Students who participated in PISA for a given country might not provide equal representation of the entire student population within the same country if random sampling is used without weights. This is because PISA does not use simple random sampling (SRS) to begin with. Instead, a two-stage sample design is used where schools are first drawn from a list of schools with 15-year-old students. Schools are selected systematically with probabilities that are proportional to a measure of size referred to as probability proportional to size (PPS) sampling. The measure of size is a function of the estimated number of eligible 15-year-old students enrolled. The second-stage sampling units are students within the sampled schools. Once schools are selected for inclusion in the sample, 35 students are selected with equal probability from the list of 15-year-old students within each school having more than 35 students. For lists of fewer than 35, all students on the list were selected. To sum up, PISA does not use a simple random sampling (SRS) procedure. Participating students are not all equally representative of the entire student population within the same country. Therefore, student weights and stratification parameters needs to incorporated into the statistical analyses. For this study, 17,000 respondents were selected through random sampling. Appropriate student weights and stratification variables are incorporated into confirmatory factor analysis (CFA).

Statistical Analyses

This section provides information on what statistics and indices are reported in each section of results by cycle, why these statistics and indices are used, and what they

mean. First, model-fit indices are discussed. Then, statistics individual parameters are explained. Lastly, tests for model comparison are provided. The section for results follows the same order within each cycle. That is, the results for model-fit indices precede that of individual parameters and model comparisons, respectively.

The computer program Mplus 6.12 (Muthén & Muthén, 1998-2011) was used to conduct confirmatory factor analyses for three cycles of PISA: 2003, 2006, and 2009. Three different types of models were evaluated: single-factor models (Model 1), one-level correlated factors models (Models 2-4), and two-level factor models (Models 5-7). The second and third types of models were formed for each of three different dimensions: content, process (competency), and context (situations). There was one model of the first type and three models of the second and the third types for each cycle evaluated. Therefore the total number of models evaluated and compared for each cycle was seven (1+3+3).

The default estimator for analyzing categorical variables is a weighted least squares mean variance (WLSMV) estimator – a robust weighted least squares (WLS) estimator using means and variances to adjust the chi-square test statistic (χ^2). WLSMV estimators are "weighted least square parameter estimates using a diagonal weight matrix with standard errors and mean- and variance-adjusted chi-square test statistic that use a full weight matrix" (Muthén & Muthén, 2012, p.603).

The chi-square test (χ^2) for the model-fit tests the null hypothesis that the model at hand is not different than the baseline (independence) model, in which all observed variables are uncorrelated. A significant chi-square test statistic, χ^2, ($p < 0.05$) results in

the rejection of the null hypothesis. That is, the model does not fit the data. Other fit indices to assess the fit of each model include Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), Root-mean-square Error of Approximation (RMSEA), and Weighted Root-mean-square Residual (WRMR).

CFI and TLI measure the improvement of fit by comparing the hypothesized model with the baseline model (a more restricted one where the observed variables are mutually uncorrelated) (Bentler & Bonett, 1980). Both the CFI and TLI have a range of [0-1]. Closer to 1 means a good fit for the model. Hu and Bentler (1999) recommend a critical minimum value of 0.95 for a good fit.

The Root-mean-square Error of Approximation (RMSEA) is a measure of the residual variances and covariances. An RMSEA value of zero implies a perfect fit. Small values of the RMSEA indicate a good fit. Hu and Bentler (1999) recommend a value of 0.06 or lower for the RMSEA. Mplus also provides a statistic for the probability that RMSEA is less than 0.05 as well as confidence intervals.

Weighted Root-mean-square Residual (WRMR) measures the average differences between the sample and estimated population variances and covariances. Yu (2002) found 1.0 to be an acceptable cut-off for the WRMR for both continuous and dichotomous outcomes. However, the credibility of this index is still controversial. Muthén and Muthén (2012) recommend not relying on WRMR since it is an experimental fit statistic that should not be of concern. Table 4.2 provides a summary of criteria for a good model-fit.

Having discussed model-fit indices, individual parameter estimates and their criteria for good model-fit are presented next (see Table 4.2 for a summary). There are two important types of statistics that apply to all models: factor loadings and R-square values. Factor loadings specify the statistical value for how much connection there is between the observed indicators and their related latent factors. For higher order models level-1 latent variables also have factor loadings onto level-2 latent variables. The cut-off value for a factor loading is 0.400 (Wang & Wang, 2012). R-square values are basically the squared values of factor loadings. The critical value for R-squares is 0.250 (Wang & Wang, 2012), meaning that 25% of the variation in the responses to mathematics items is explained by the model at hand. If it falls below 0.25, then there are other unknown factors causing the variation above and beyond the model.

Correlation coefficients are another type of statistic reported but they only apply to correlated-factors (1-level) models (Models 2-4). Values close to 1 in absolute value imply high correlation between the factors while values close 0 mean low or negligible correlation. Other values are considered as a sign of moderate correlations between the factors (see Table 4.2). There is a hypothesis testing for each of the factor loadings, R-squares, and correlation coefficients. If the hypothesis test results significant then it rejects the null hypothesis that the estimate is no different than 0. That is, estimated factor loadings, R-square values, and correlations are significantly different than 0 and are important elements of the model that is being tested.

Model comparison for confirmatory factor analysis (CFA) models is typically done using a chi-square difference testing. Two models that are nested within each other

could be compared using CFA methods. More restrictive (parsimonious) models (i.e., less parameterized) with more degrees of freedom are nested in less restrictive models with fewer degrees of freedom. It is important note that the unidimensional CFA model (Model 1) in this study is the most restrictive model. The 2-level models (Models 5-7) are less restrictive than the unidimensional model but more restrictive than the 1-level (Models 2-4) models.

Table 4.2. *Critical values for model fit indices and individual parameter estimates*

Index/Parameter	**Possible Values**	**Criterion**
TLI	[0,1]	> 0.95
CFI	[0,1]	> 0.95
RMSEA	[0.1]	< 0.06
WRMR	$[0, \infty)$	< 1.00
Factor Loading	(0,1)	> 0.40
R-Square	(0, 1)	> 0.25
Correlation Coefficient	[-1, 1]	*close to 1 implies high correlation*
Δ*CFI*	[-1, 1]	> -0.01 *points to unrestricted model*

The chi-square statistic for the difference testing is then the difference in chi-square values with difference in degrees of freedom:

$$\Delta\chi^2{}_{\Delta df} = (\chi^2{}_{mr} - \chi^2{}_{lr})_{(df_{mr} - df_{lr})}$$

where ${\chi^2}_{mr}$ and ${\chi^2}_{lr}$ are the chi-square values of the more restrictive and the less restrictive models. Similarly, df_{mr} and df_{lr} are the degrees of freedom for the more restrictive and the less restrictive models.

The chi-square value for WLSMV estimator cannot be used for chi-square difference testing in the regular way. In other words, for binary or categorical data, model comparisons cannot be done with the chi-square difference testing. Instead, the DIFFTEST method, which provides the appropriate adjustment to the chi-square difference test when using WLSMV chi-square, should be used to compare nested models (Muthén & Muthén, 2012). A significant test result means a significant amount of fit is lost with restriction. That is, a less restrictive model is better. A non-significant result implies that there is no significant amount of fit loss with the restriction. Although, the latter result conveys the message that the model-fits are essentially the same for both models that are being compared, one would choose a more restrictive model to move forward because parsimony is preferred with CFA models (Kline, 2010). However, since DIFFTEST is a derivative of the chi-square test, there is chance of Type-I error with the large sample size in this study. Cheung and Rensvold (2002) propose a ΔGFI method to compare models based on the difference of goodness of fit indices (GFI) other than chi-square.

$$\Delta GFI = GFI_{mr} - GFI_{lr}$$

where GFI_{mr} and GFI_{lr} are the values of some selected GFI estimated with respect to the more restrictive and less restrictive model. While their study is about invariance across groups, the use of delta GFIs is extended for nested models in this

dissertation. Among the GFIs Cheung and Rensvold (2002) proposed, only CFI applies to analyses in this study because of the WLSMV estimator selection for CFA analyses. A value of ΔCFI greater than -0.01 indicates that the null hypothesis of invariance between the more restrictive and less restrictive models should not be rejected (see Table 4.2). Stated differently, a ΔCFI less than or equal to -0.01 gives a significant result. A ΔCFI greater than -0.01, then, means models are no different from each other. In this case, a more restrictive model (i.e., the unidimensional model) is preferred over the less restrictive one (e.g., a 2-level model) for the data because it does not lose significant amount of fit (Kline, 2010).

The formal hypotheses are given in chapter 3. However, it is worth it to recall what these are. The first model (single-factor) hypothesizes that PISA mathematics items measure a single construct labeled as general mathematical literacy (GML). The second type of model (Models 2-4) hypothesizes that the PISA mathematics items helps explain mathematics knowledge, competencies, and skills in terms of correlated factors of related dimension (content, process, or context) as the latent constructs. The third type of model (Models 5-7) hypothesizes that the PISA mathematics items measure GML (level-2 factor) by factors of related dimension (the level-1 latent variables).

The remainder of this chapter is devoted to the results that are presented cycle-by-cycle. For each cycle, model-fit indices are presented first, which are followed by individual parameter estimates, and then, model comparisons. The sections for model-fit indices and individual parameter estimates correspond to the first research question (relationship between the dimensional structure of student responses and assessment

framework). The sections for model comparisons relate to the second research question (overall dimensionality of the PISA mathematics items). Lastly, models are evaluated across cycles, addressing the third research question about the longitudinal aspect of dimensionality of PISA mathematics items.

CFA Results: 2003 Cycle

Assessment of Models

Table 4.3 summarizes the model-fit indices for each of the seven models in the 2003 cycle. The chi-square test statistics (χ^2) for each model with 2003 data range from 3859.488 (with *df* = 3396) to 3898.008 (with *df* = 3402), $p < 0.001$ for all models, which rejects the null hypothesis of a good fit. However, as mentioned before, χ^2 statistics are highly sensitive to sample size. Although reduced sample sizes were used in this study to decrease the power, it might be the case that the reduced sample is still large enough to produce significant results. Therefore, the significance of the χ^2 statistics should not be definitive by itself to conclude that models do not fit the data (Wang & Wang, 2012).

In regards to other fit indices, TLI and CFI values for all models are greater than the critical value of 0.95. This implies a good fit for all of the models. Moreover, estimated values of RMSEA for all models are found to be very close to zero. The probability that RMSEA is less than the cut-off value of 0.05 is almost 1. This means that all models fit the data fairly well. Although, the WRMR value is greater than the critical value of 1.0 for all models, the credibility of this index is still controversial. WRMR is an experimental fit statistic that should not be of concern (Muthén & Muthén, 2012). Other fit indices CFI, TLI, and RMSEA are within the interval for a good fit. Overall, it

Table 4.3 *Model fit indices for 2003.*

	Model 1: 1F-GML	**Model 2: 4F-Content**	**Model 3: 3F-Process**	**Model 4: 4F-Context**	**Model 5: L2-Content**	**Model 6: L2-Process**	**Model 7: L2-Context**
Chi-Square Test of Model Fit							
Value	3898.008	3859.488	3892.262	3890.017	3862.815	3894.486	3890.814
Degrees of freedom	3402	3396	3399	3396	3398	3401	3398
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CFI/TLI							
CFI	**0.973**	**0.975**	**0.973**	**0.973**	**0.975**	**0.973**	**0.973**
TLI	**0.972**	**0.974**	**0.972**	**0.972**	**0.974**	**0.972**	**0.972**
RMSEA (Root Mean Square Error of Approximation)							
Estimate	**0.003**	**0.002**	**0.002**	**0.002**	**0.002**	**0.002**	**0.002**
90 Percent C.I.	0.002-0.003	0.002-0.003	0.002-0.003	0.002-0.003	0.002-0.003	0.002-0.003	0.002-0.003
Probability RMSEA <= .05	1.000	1.000	1.000	1.000	1.000	1.000	1.000
WRMR (Weighted Root Mean Square Residual)	1.163 1.161	1.148	1.162	1.161	1.149	1.162	

Values that satisfy the criteria for good model fit are bold-faced.

can be concluded that each of the seven models for PISA 2003 mathematics items fit the data well. Moreover, the values for the model-fit indices are nearly identical for all the models.

Individual Parameter Estimates

There were 84 mathematics items in 2003. Factor loadings and R-square values for these are given in the following sub-sections for each of the seven models as well as correlations between factors where it is applicable. These estimated values relate to the first research question and will help determine how the mathematics framework is reflected in the response data with respect to the three dimensions.

Model 1: Single-factor model (1F-GML)

All of the observed indicators except for items M75, M82, and M83 have a factor loading greater than the cut-off value of 0.400. M75, M82, and M83 have a lower standardized factor loading of 0.383, 0.334, and 0.344, respectively. This normally suggests that these items are a weaker indicator of the latent factor GML (general mathematical literacy). However, their factor loadings are statistically significant ($p <$ 0.001) meaning that they are significantly different from 0. They might be very important theoretically to keep in the model, but might not contribute to good model-fit necessarily. Items M28, M45, M48, and M80 have the greatest factor loadings (higher than 0.800). The factor loadings for the majority of the items lay between 0.500 and 0.700. Estimated R-square values are the square of standardized factor loadings and provide information on how much variance of each observed variable is explained by related latent factors. For example, estimated standardized loading for M01 is 0.512, and then its R-square

value is $(0.512)^2 = 0.262$ meaning that about 26% of the variance in item M01 is explained by GML. M28 has the highest R-square value (0.804) while M82 has the lowest (0.15). A majority of the items have an R-square value greater than 0.250. Only 13 items (M20, M24, M36, M40, M41, M57, M65 M66, M68, M75, M82, M83, and M84) have R-square values below 0.250. However, all R-square values are found to be statistically significant ($p < 0.001$). This further provides additional support for unidimensional model for 2003 response data. Since PISA only made 31 of the 84 items publicly available (see Appendix B for sample items), none of the items whose factor loadings or R-square values that are either at the lower or upper end of the scale are among the ones that have been released by PISA. So, it is not possible to do a finer grain-sized analysis of the other item characteristics like content, language, or context. Although the classifications of these items with respect to the three dimensions of content, process, and context are known (see Appendix A), the items themselves are not available to make further qualitative analyses. To summarize, although some factor loadings and R-square values are lower than their critical values and, thus, might not contribute to good model-fit, all of the individual parameter estimates are statistically significant and should be kept in the model from a theoretical point of view.

Model 2: 1-level (four-factor) content model

As in the single factor model, all of the observed indicators (except M75, M82, and M83) have a factor loading greater than the cut-off value of 0.400. M75 and M83 have a lower standardized factor loading (0.388 and 0.347, respectively) on QT (quantity) content dimension while M82 has a standardized factor loading of 0.335 on UN

(uncertainty) content dimension. This normally suggests that these items are a weaker indicator of their related latent factor. However, their factor loadings are statistically significant ($p < 0.001$). So, they might be very important theoretically to keep in the model, but might not contribute to good model-fit necessarily. On the other hand, factor loadings for the items M07, M27, M28, M45, M48, M63, and M80 are higher than 0.800. The factor loadings for the majority of the items range between 0.500 and 0.700. M28 has the highest R-square value (0.866) while M82 has the lowest (0.112) as in the single-factor model. A majority of the items have an R-square value of 0.250 or greater. Only 11 items (M20, M24, M36, M40, M41, M57, M66, M68, M75, M82, and M83) have R-square values 0.249 or less. This means 25% or more of the variance in 73 of the items is explained by their related factors.

Table 4.4. *Correlations between 2003 content dimensions*

	QT	**SS**	**CR**	**UN**
QT	1			
SS	0.905	1		
CR	0.973	0.911	1	
UN	0.990	0.907	0.984	1

The four latent variables QT, SS, CR, and UN are highly correlated to each other ($p < 0.001$). Among six correlations between four content factors, the lowest one is between SS (space and shape) and QT (quantity): 0.905, which is still very high (see Table 4.4). UN and QT pair has the highest correlation (0.990). It seems that all of four content factors essentially behave as one unifying construct rather than four different latent factors. This provides evidence that supports a unidimensional model much better than a correlated factors content model.

Lastly, none of the items whose factor loadings or R-square values that are either at the lower or upper end of the scale are available for further analysis. However, all factors loadings, correlations, and R-square values are found to be statistically significant. This implies that 4F-Content model accounts very well for the relationships among the 2003 mathematics items. However, despite the good model-fit results for 4F-Content model, high correlations between factors provide strong support towards a unidimensional model, which also showed good model-fit results as described in the previous section.

Model 3: 1-level (three-factor) process model

All of the observed indicators of three-factor process model but three are loaded onto their related factors with a 0.400 loading value or higher. M82 and M83 have a lower standardized factor loading on CON (connections) process dimension (0.335 and 0.345, respectively) while M75 has a standardized factor loading of 0.389 on REP (reproduction) process dimension. Since their factor loadings are statistically significant ($p < 0.001$), they are significantly different from 0. They might be very important theoretically to keep in the model, but might not contribute to good model-fit necessarily. On the other hand, factor loadings for the items M28, M45, M48, and M80 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. R-square values range from 0.112 to 0.810. A majority of the items have an R-square value of 0.250 or greater. Only 12 items (M20, M24, M36, M40, M41, M57, M66, M68, M75, M82, M83, and M84) have R-square values 0.249 or less. This means 25% or more of the variance in 72 of the items is explained by their related factors.

The three latent variables (REP - *reproduction*, CON - *connections*, and REF - *reflection*) are significantly correlated to each other. The correlation coefficients are found to be higher than 0.95. Table 4.5 shows the values of correlation coefficients between all process sub-dimensions, which are very highly correlated to each other with values close to 1. This supports the interpretation that all of the three process factors essentially look like one unifying construct rather than three different latent factors. This provides evidence that supports a unidimensional model much better than a correlated factors process model.

Table 4.5. *Correlations between 2003 process dimensions*

	REP	**CON**	**REF**
REP	1		
CON	0.975	1	
REF	0.966	0.997	1

Lastly, none of the items with high/low factor loadings or R-square values are available for further analysis. Like previous models, this 3F-Process model also accounts very well for the relationships among the 2003 mathematics items. However, high latent factor correlations provide evidence favoring a unidimensional model over this 1-level process model.

Model 4: 1-level (four-factor) context model

Factor loadings for observed indicators in the 1-level context model range in value from 0.342 to 0.899. All but three indicators load onto their related factors more than 0.400. M75 load onto PER (personal) while M82 and M83 load onto SCI (scientific) context latent variables with values of 0.342, 0.352, and 0.382, respectively. Although

these values are less than the cut-off value of 0.400, they are significant ($p < 0.001$). On the other hand, the factor loadings for the items M15, M28, M45, M48, and M80 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. R-square values for observed indicators vary from 0.117 to 0.809. All but 13 indicators have an R-square value greater than 0.250. The amount of variation in student response to mathematics items explained by this model falls below 25% for items M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, and M84. None of these items are available for further analysis. Like the models for other dimensions, all factors loadings, correlations, and R-square values are statistically significant. Therefore, it is concluded that 4F-Context model accounts very well for the relationships among the 2003 mathematics items. Yet, high correlations between factors provide evidence favoring a unidimensional model over this 1-level context model. The four context factors are highly correlated with each other with correlation coefficients ranging from 0.939 to 0.985 (Table 4.6). It seems that all of four context factors essentially behave as one unifying construct rather than four different latent factors. This provides evidence that supports a unidimensional model much better than a correlated factors context model.

Table 4.6. *Correlations between 2003 context dimensions*

	PER	**EDOP**	**PUB**	**SCI**
PER	1			
EDOP	0.984	1		
PUB	0.985	0.972	1	
SCI	0.982	0.939	0.962	1

Model 5: 2-level content model

In this model, all but three observed indicators have a factor loading greater than the cut-off value of 0.400. M75 and M83 have a lower standardized factor loading (0.388 and 0.347, respectively) on QT (quantity) content dimension while M82 has a standardized factor loading of 0.335 on UN (uncertainty) content dimension. On the other hand, factor loadings for the items M07, M27, M28, M45, M48, M63, and M80 are higher than 0.800. The factor loadings for the majority of the items are greater than 0.500. All of the level-1 content factors (QT, SS, CR, and UN) load highly onto the level-2 latent variable GML, ranging from 0.914 to 0.996. The proportions of variance in the level-1 factors explained by the level-2 factor are 0.992, 0.835, 0.970, and 0.990 for QT, SS, CR, and UN, respectively. This provides strong support for a higher order structure of the construct that underlies the mathematics items. That is, level-1 factors QT, SS, CR, and UN are good measures of the level-2 variable, GML (general mathematical literacy). R-square values range from 0.113 to 0.866. A majority of the items have an R-square value of 0.250 or greater. 11 items (M20, M24, M36, M40, M41, M57, M66, M68, M75, M82, and M83) have R-square values 0.249 or less. This means 25% or more of the variance in 73 of the items is explained by their related factors. None of the items whose factor loadings or R-square values that are either at the lower or upper end of the scale are available for further analysis. As in 1-level content factor, all of the factors loadings and R-square values are statistically significant despite some being lower than cut-off values. Thus, evidence also supports this 2-level model.

Model 6: 2-level process model

All but three of the observed indicators have a factor loading of 0.400 or higher. M75 has a lower loading on level-1 factor REP. Likewise, M82 and M83 have lower factor loading on level-1 factor connections (CON). Items M28, M45, M48, and M80 have factor loadings greater than 0.800. The factor loadings for the majority of the items are above 0.500. All of the level-1 content factors (reproduction, connection, and reflection) load highly onto the level-2 latent variable GML, ranging from 0.984 to 0.993. This provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors REP, CON, and REF are good measures of the level-2 variable, GML (general mathematical literacy) in terms of process dimension. The proportions of variation in the level-1 factors REP, CON, and REF explained by the level-2 factor are 0.968, 0.986, and 0.975, respectively. This indicates that the 2-level structure provides a good account for the covariances among the level-1 latent variables. The R-square values for all but 12 items (M20, M24, M36, M40, M41, M57, M66, M68, M75, M82, M83, and M84) are higher than 0.250. The 2-level process model, then, explains 25% or more of the variation in the responses to the majority of the items. All of the factors loadings and R-square values are statistically significant in spite of some low values. None of the items whose factor loadings or R-square values are either at the lower or upper end of the scale are available for further analysis. The conclusion made here is similar to the ones in unidimensional and 3-Factors correlated models: the level-2 model also provides a good structural representation of the PISA 2003 mathematics items in terms of process dimension.

Model 7: 2-level context model

All items load onto their first level context factors with more than 0.400 factor loading values except for three items: M75, M82, and M83. Items M15, M28, M45, and M48, and M80 have high factor loadings greater than 0.800. The factor loadings for the majority of the items are greater than 0.500. Level-1 context factors PER (personal), EDOP (educational/occupational), PUB (public), and SCI (scientific) load highly onto level-2 latent variable GML with values 0.990, 0974, 0.995, and 0.967, respectively. The proportions of variation in the level-1 context factors explained by the level-2 factor are 0.980, 0.949, 0.991 and 0.935, respectively. This indicates that the 2-level structure provides a good account for the covariances among the level-1 latent variables for context dimension. This provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors PER, EDOP, PUB, and SCI are good measures of the level-2 variable, GML (general mathematical literacy) in terms of the context dimension. A majority of the items have R-square values greater than 0.250. The items M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, and M84 have a low R-square value. None of these items, however, are available for further analysis. As in its 1-level model, all of the factors loadings and R-square values are statistically significant. Thus, overall, the 2-level model also provides a good structural representation of the PISA 2003 mathematics items in terms of context dimension. Table 4.7 summarizes results for individual parameter estimates and shows that items, which have critical values for individual parameter estimates related to them, do not differ across the models. In other words, it is almost the same items that land

around the upper/lower boundaries of factor loadings and R-square values with slight differences from one model to the other. None of these items are available for further qualitative analyses. Moreover, correlations among factors are very high across all 1-level models. These results provide evidence for a unidimensional model. Thus, it is concluded that individual parameter estimate results for unidimensional and 1-level models for 2003 response data in mathematics imply that mathematical literacy is a unidimensional construct that unifies all three dimensions (content, process, and context) and their sub-dimensions. Although all of the items could be assigned to a different category in terms of different dimensions, they essentially measure the same overall construct: mathematical literacy. The factors loadings of level-1 latent variables are very high for 2-level models. This, however, provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors are good measures of the level-2 variable, GML (general mathematical literacy) in terms of the three dimensions. These results, considered altogether, do not contradict the multidimensionality of 2003 mathematics items. However, there is more evidence that supports the unidimensionality of PISA mathematics items.

Table 4.7. *Summary of individual parameter estimates for 2003*

Models	Item loadings Low (<0.400)	Item loadings High (>0.800)	Items with low R-square (<0.250)	Correlations b/w L-1 factors	L-1 loadings onto L-2
Model 1: 1F-GML	M75, M82, M83	M28, M45, M48, M80	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	N/A	N/A
Model 2: 1-L Content	M75, M82, M83	M07, M27, M28, M45, M48, M63, M80	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75 M82, M83	>0.905	N/A
Model 3: 1-L Process	M75, M82, M83	M28, M45, M48, M80	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	>0.966	N/A
Model 4: 1-L Context	M75, M82, M83	M15, M28, M45, M48, M80	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	>0.939	N/A
Model 5: 2-L Content	M75, M82, M83	M07, M27, M28, M45, M48, M63, M80	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75 M82, M83	N/A	>0.914
Model 6: 2-L Process	M75, M82, M83	M28, M45, M48, M80	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	N/A	>0.968
Model 7: 2-L Context	M75, M82, M83	M15, M28, M45, M48, M80	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	N/A	>0.967

Model Comparisons

As mentioned above, DIFFTEST method is used to compare nested models (Muthén & Muthén, 2012). However, since DIFFTEST is a derivative of the chi-square test, there is chance of Type-I error with the large sample size in this study. Therefore, in the case of a significant result for a DIFFTEST between any two nested models, the ΔCFI method, an extension of Cheung and Rensvold's (2002) ΔGFI method, is used to compare models in this dissertation. A value of ΔCFI greater than –0.01 indicates that the null hypothesis of invariance between the more restrictive and less restrictive models should not be rejected (see Table 4.2). Stated differently, a ΔCFI less than or equal to -0.01 gives a significant result. A ΔCFI greater than -0.01, then, means models are no different from each other. This same conclusion is arrived as well when DIFFTEST results in a non-significant value. In these cases, a more restrictive model (e.g., Model 1: 1F-GML) is preferred over the less restrictive one (e.g., Model 5: 2L-Content) for the data because it does not lose significant amount of fit (Kline, 2010).

In total, there are seven models tested in each cycle. However, when comparing the models by using either DIFFTEST or ΔCFI methods, all seven cannot be evaluated all together at once. Models have to be compared in pairs and they have to be nested within each other. The unidimensional model is the most restrictive among the all models. It is nested in each and every of the other six models. So, it can be compared to all other six models one by one. 2-level models are nested in 1-level models in terms of each dimension. To be more specific, the 2-level content model is nested only in the 1-level (4F) content model, not in the 1-level process and context models. Basically, nine

possible comparisons could be made in each cycle: unidimensional model compared to every other model (six comparisons in total) and the 2-level model compared to the 1-level model for each of three dimensions (3 comparisons in total).

Table 4.8. *DIFFTEST results for 2003 models*

1F-GML versus	**L2-Content**	**L2-Process**	**L2-Context**
Value	122.009	11.607	42.992
Δ*df*	4	1	4
p-value	0.0000*	0.0000*	0.0000*
	L2-content vs. **4F-Content**	**L2-Process** vs. **3F-Process**	**L2-context** vs. **4F-Context**
Value	8.036	3.935	0.859
Δ*df*	2	2	2
p-value	0.0180*	0.1398	0.6507
1F-GML versus	**4F-Content**	**3F-Porcess**	**4F-Content**
Value	131.131	14.378	14.171
Δ*df*	6	3	6
p-value	0.0000*	0.0024*	0.0278*
* Significant at 0.05			

Table 4.8 summarizes the DIFFTEST results for all the model comparisons in the 2003 cycle. Among all comparisons, only two are found to be non-significant: L2-Process versus L1-Process and L2-Context versus L1-Context. So, 2-level and 1-level models perform about the same for the process and context dimensions. However, the more restrictive models (i.e., the 2-level models when compared to the 1-level models) are preferable because of their parsimony. All other comparisons are found significant,

which, however, might be resulted from a large sample size. Therefore, ΔCFI method is further applied for these comparisons. The results are given in Table 4.9. ΔCFI results show that only two comparisons are significant: 1F-GML versus L2-Content and 1F-GML versus 4F-Content. The L2-Content and L1-Content models are preferred to the 1F-GML (ΔCFI = -0.02). When the 1-level and 2-level models are compared to each other for the content dimension, it is found that model comparison is not significant (ΔCFI = 0). Therefore, the more restrictive model, L2-content is chosen. ΔCFI results also support the non-significant result obtained through DIFFTEST.

Table 4.9. Δ*CFI results for 2003 model comparisons*

	1F-GML vs. L2-Content	**1F-GML vs. L2-Process**	**1F-GML vs. L2-Context**
ΔCFI	-0.02*	0	0
	L2-Content vs. 4F-Content	**L2-Process vs. 3F-Process**	**L2-Context vs. 4F-Context**
ΔCFI	0	0	0
	1F-GML vs. 4F-Content	**1F-GML vs. 3F-Process**	**1F-GML vs. 4F-Context**
ΔCFI	-0.02*	0	0

* Significant at -0.01

To summarize the model comparison results, it would make sense to take them at hand according to each dimension. For the content dimension, the 2-level model is preferred to the 1-level correlated four-factors model, which is preferred to the unidimensional model. Although there are no specific rankings for the process and context dimensions because of no significant differentiation among different models, there is a preference: the more restrictive model is preferred. Therefore, the

unidimensional models are preferred over the 2-level models and the 1-level correlated factors models.

Content: 2-Level Model > 1-Level Model > 1F-GML Model
Process: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model
Context: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model

Figure 4.1. Model Comparisons for 2003

Figure 4.1 illustrates the summary of models comparisons for 2003. The ">" sign refers to a ranking where the model on the left is preferred to the one on the right. The "≥" sign is used for models that are indifferent in terms of their fit performance to the data but the model on the left is always preferred for its parsimony.

Summary of 2003 Results

First, all seven models for the PISA 2003 mathematics items fit the 2003 data well. This implies that the dimensional structure of the 2003 mathematics items do not contradict any of the models proposed by the PISA mathematics framework. Connecting these results to the first research question (response-framework correspondence), overall model-fit results do not indicate any contradiction for the correspondence between dimensional structure of the mathematics items and mathematical literacy framework. However, there is more evidence supporting the unidimensionality of the mathematics items. To better understand how each model corresponds to individual items, parameter estimates were evaluated. First of all, all parameter estimates were found significant for each of the seven models, meaning that all models provide a good account for factor

loadings. In other words, each mathematics item plays an important role in different dimensionality models. This further support the evidence that mathematics framework is reflected in the mathematics items through student response data with respect to the three dimensions.

What is interesting, however, is that the items whose factor loadings and R-square values are around lower and upper boundaries of critical values are almost the same items in all models (see Table 4.7). Unfortunately, none of these items are available for any further analysis. Moreover, the factor loading values for each mathematics item are almost identical across models (see Appendix C for individual factor loadings). It can be concluded that the items psychometrically behave the same in all dimensions. That is, regardless of the dimension (e.g., whether it is an uncertainty (UN), a reflection (REF), or scientific (SCI) question) an item's loading and explained variation by a specific model is the same.

Correlations between the latent variables for 1-level models are very high for all dimensions (see Table 4.4, Table 4.5, and Table 4.6). High correlations provide evidence supporting a unidimensional model rather than either multidimensional correlated-factor models. The level-1 factors' factor loadings onto level-2 latent variable (GML) are very high for all dimensions. This provides a strong support for higher order structure of the construct that underlies the mathematics items. Model comparisons indicate that the unidimensional model is preferable over all models for the process and context dimensions (see Figure 4.1). For the content dimension, the 2-level model is preferred to the1-level correlated four-factors model, which is preferred to the unidimensional model.

QT (quantity), SS (space and shape), CR (change and relationship), and UN (uncertainty) dimensions are a measure of the global construct, GML (general mathematical literacy). Going back the second research question, the unidimensional model is preferable. However, in terms of content structure of mathematics items, the 2-level model better represents the dimensional structure of the PISA 2003 mathematics items. To conclude, unidimensional structure explains the mathematics items best when process and context dimensions are evaluated. However, a rather multidimensional structure represents the mathematics items in terms of the content dimension. The results provide evidence for both unidimensionality and multidimensionality. However, there is additional evidence for the unidimensionality (high correlations between the 1-level models). Thus, overall results conclude that although the multidimensionality cannot be disproven, there is stronger evidence that supports the unidimensionality of the PISA mathematics items.

CFA Results: 2006 Cycle

CFA analyses produced similar results for 2006 cycle as follows.

Assessment of Models

The model-fit indices of models for 2006 are given in Table 4.10. The chi-square statistics (χ^2) for each model range from 1307.484 (with *df = 1074*) to 1347.497 (with *df = 1080*). As in 2003, chi-square tests for all models are significant ($p < 0.001$), which reject the null hypothesis of a good fit. However, the sample size might have inflated chi-square statistics. Therefore, other fit indices are more reliable to make conclusions for this study.

Table 4.10 *Model fit indices for 2006*

	Model 1: 1F-GML	**Model 2: 4F-Content**	**Model 3: 3F-Process**	**Model 4: 4F-Context**	**Model 5: L2-Content**	**Model 6: L2-Process**	**Model 7: L2-Context**
Chi-Square Test of Model Fit							
Value	1347.497	1307.484	1337.503	1330.031	1309.462	1344.309	1331.441
Degrees of freedom	1080	1074	1077	1074	1076	1079	1076
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CFI/TLI							
CFI	**0.980**	**0.982**	**0.980**	**0.981**	**0.982**	**0.980**	**0.981**
TLI	**0.979**	**0.981**	**0.979**	**0.980**	**0.982**	**0.979**	**0.980**
RMSEA (Root Mean Square Error of Approximation)							
Estimate	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**
90 Percent C.I.	0.003-0.004	0.003-0.005	0.003-0.004	0.003-0.004	0.003-0.004	0.003-0.004	0.003-0.004
Probability RMSEA <= .05	1.000	1.000	1.000	1.000	1.000	1.000	1.000
WRMR (Weighted Root Mean Square Residual)	1.208	1.170	1.199	1.192	1.171	1.205	1.193

Values that satisfy the criteria for good model fit are bold-faced.

TLI and CFI values for all models are greater than the critical value of 0.95. Therefore, TLI and CFI values support a good fit for all of the models. Moreover, RMSEA values for all models are found to be very close to zero. The probability that RMSEA is less than the cut-off value of 0.05 is almost 1. This means that all models fit the data fairly well. Although, the WRMR value is greater than the critical value of 1.0 for all models, the overall fit of each model is found to be very good for 2006 mathematics items.

Individual Parameter Estimates

There were 48 mathematics items in 2006. Factor loadings and R-square values for these are given in the following sub-sections for each of the seven models as well as correlations between factors where it is applicable. These estimated values relate to the first research question and will help determine how the mathematics framework is reflected in the mathematics items through the student response data with respect to the three dimensions.

Model 1: Single-factor model

Only one item, M24, among 48 items has a low factor loading (0.373) in the unidimensional model. This item had a moderate loading (around 0.450) in 2003 models. The remaining indicators load onto GML more than the critical value of 0.400. Their factor loadings range from 0.426 to 0.874. About half of the factor loadings are between 0.500 and 0.700. Items M28, M37, M45, and M80 have the greatest factor loadings (higher than 0.800). All factor loadings including that of M24 are significant ($p < 0.001$). This unidimensional model explains 25% or more of the variation in student responses to

a majority of the items (39 of the 48) in 2003. The amount of variance explained falls below the cut-off level 25% for the items M01, M24, M39, M40, M51, M65, M68, M75, and M84. This signals the presence of other factors causing the variation above and beyond the unidimensional model for these items. Although some factor loadings and R-square values are lower than their critical values, all individual parameter estimates are significant. This further provides additional support for the unidimensionality of the PISA 2006 mathematics items. None of the items, whose factor loadings or R-square values that are either at the lower or upper end of the scale, are released. So, it is not possible to do a finer grain-sized analysis of the other item characteristics like content, language, or context.

Table 4.11. *Correlations between 2006 content dimensions*

	QT	**SS**	**CR**	**UN**
QT	1			
SS	0.863	1		
CR	0.936	0.905	1	
UN	0.899	0.846	0.955	1

Model 2: 1-level (four-factor) content model

All item loadings onto their related factors (QT, CR, SS, and UN) are statistically significant. Factor loadings in the 1-level content factor range from 0.406 to 0.917. Factor loadings for the items M15, M27, M28, M37, M45, M48, and M80 are higher than 0.800. The factor loadings for the majority of the items range between 0.500 and 0.700. There are only five items whose variation could not be explained by the four-factor content model more than 25%. These items are M24, M40, M51, M65, and M75. Their R-square values range from 0.157 to 0.239. None of these items are available for further analysis.

The fact that all of factors loadings, correlations, and R-square values are statistically significant indicates that the 4F-Content model accounts for the relationships among the PISA 2006 mathematics items very well, although there are some low factor loadings and R-square values. Four content factors (QT, CR, SS, and UN) correlate with each other significantly high. Correlation coefficients vary from 0.846 to 0.955 as shown in Table 4.11. CR and UN factors correlate the most while the lowest correlation is between SS and UN. It seems that all of four content factors essentially behave as one unifying construct rather than four different latent factors. This provides evidence supporting a unidimensional model, which also showed good model-fit results as described in the previous section, rather than a multidimensional content model.

Model 3: 1-level (three-factor) process model

All of the observed indicators of the three-factor process model load onto their related factors with a 0.400 loading value or higher except indicator M24, which loads onto CON factor with a value of 0.373. The factor loadings for the items M15, M28, M37, M45, and M80 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. R-square estimates are statistically significant with values ranging from 0.139 to 0.799. Only 8 items (M24, M39, M40, M51, M65, M68, M75, and M84) have R-square values 0.249 or less. This means 25% or more of the variance in the majority (40) of the items is explained by their related factors through this model. None of these are available for further analysis. This model explains all factor loadings, correlations, and R-square values at a significant level. Therefore, 3F-process model fits the data for the PISA 2006 mathematics items very well.

Table 4.12. *Correlations between 2006 process dimensions*

	REP	CON	REF
REP	1		
CON	0.906	1	
REF	0.971	0.976	1

Model 4: 1-level (four-factor) context model

The three latent variables (REP, CON, and REF) are significantly correlated to each other. As shown in Table 4.12, the correlation coefficients are found to be higher than 0.90. All of three process dimensions essentially look like one unifying construct rather than three different latent factors. This provides support towards a unidimensional model for 2006 mathematics items.

Factor loadings for the observed indicators in the 1-level context model range from 0.378 (M24 on PUB) to 0.888 (M45 on SCI). M24 is the only one item with a factor loading less than 0.400. The factor loadings for the items M15, M28, M37, M45, M48, and M80 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. R-square values for observed indicators vary from 0.143 (M24) to 0.789 (M45). All but 7 indicators have an R-square value greater than 0.250. The amount of variation explained by this model falls below 25% for items M24, M40, M51, M65, M68, M75, and M84. None of the items whose factor loadings or R-square values are either at the lower or upper end of the scale are available for further analysis. Like the models for other dimensions, all of the factors loadings, correlations, and R-square values are statistically significant. Therefore, it is concluded that the 4F-Context model accounts very well for the relationships among the PISA 2006 mathematics items.

Table 4.13. *Correlations between 2006 context dimensions*

	PER	EDOP	PUB	SCI
PER	1			
EDOP	0.930	1		
PUB	0.946	0.930	1	
SCI	0.969	0.924	0.925	1

Four context factors, PER (personal), EDOP (educational/occupational), PUB (public), and SCI (scientific), are highly correlated with each other as shown in Table 4.13. This indicates that all of four context factors essentially behave as one unifying construct rather than four different latent factors, thus, supporting rather a unidimensional model.

Model 5: 2-level content model

This model produced factor loadings greater than the cut-off value of 0.400 for all items at level-1. Factor loadings for the items M15, M27, M28, M37, M45, M48, and M80 are higher than 0.800. The factor loadings for the majority of the items range between 0.500 and 0.700. Also, all of the level-1 content factors (QT, SS, CR, and UN) load highly onto the level-2 latent variable GML, ranging from 0.906 to 0.998. This provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors QT, SS, CR, and UN are good measures of the level-2 variable, GML (general mathematical literacy). The proportions of variance in the level-1 factors explained by level-2 factor are 0.892, 0.821, 0.996, and 0.903 for QT, SS, CR, and UN, respectively. This indicates that this higher-order structure provides a very good account for the covariances among the level-1 factors. R-square values range from 0.157 to 0.842. All but five of the items (M24, M40, M51, M65, and M75) have an R-

square value of 0.250 or greater. None of these items are available for further analysis. As in 1-level content factor, all of the factors loadings and R-square values are statistically significant despite some lower than cut-off values. Thus, the 2-level content model results reveal that it provides a good account for the PISA 2006 mathematics items.

Model 6: 2-level process model

All observed indicators have a factor loading of 0.400 or higher except M24, which loads onto CON factor with a loading value of 0.374. M45 has the biggest loading onto REF factor (0.882). The factor loadings for the items M15, M28, M37, M45, and M80 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. All of the level-1 content factors (REP, CON, and REF) load very highly onto the level-2 latent variable GML, ranging from 0.979 to 0.992. This provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors REP, CON, and REF are good measures of the level-2 variable, GML (general mathematical literacy) in terms of process dimension. The proportions of variation in the level-1 factors REP, CON, and REF explained by the level-2 factor are 0.958, 0.969, and 0.983, respectively. This indicates that this 2-level structure provides a good account for the covariances among the level-1 latent variables. The R-square values for all but 8 items (M24, M39, M40, M51, M65, M68, M75, and M84) are higher than 0.250. The 2-level process model, then, explains 25% or more of the variation in the responses to a majority of the items. All of the factors loadings and R-square values are statistically significant despite some low estimates. None of the items whose factor loadings or R-square values are either at the lower or upper end of the scale are available

for further analysis. Overall, the 2-level process model provides a good structural representation of the PISA 2006 mathematics items in terms of process dimension.

Model 7: 2-level context model

All items load onto their first level context factors with more than 0.400 factor loading values except item M24. The factor loadings for the items M15, M28, M37, M45, M48, and M80 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. The level-1 context factors EDOP, PUB, SCI, and PER load highly onto the level-2 latent variable GML with values 0.958, 0.960, 0.967, and 0.986, respectively. The proportions of variation in the level-1 context factors explained by level-2 factor are 0.917, 0.923, 0.936, and 0.972, respectively. This indicates that the 2-level structure provides a very good account for the covariances among the level-1 latent variables for the context dimension. This provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors PER, EDOP, PUB, and SCI are good measures of the level-2 variable, GML (general mathematical literacy) in terms of context dimension. All but seven items have R-square values greater than 0.250. The items M24, M40, M51, M65, M68, M75, and M84 have a low R-square value. However, these R-square values are statistically significant. So are all of factor loadings. Therefore, the 2-level context model provides a good structural representation of the PISA 2006 mathematics items in terms of context dimension as well.

Overall, individual parameter estimate results are similar to the ones in 2003. There are some slight differences that will be discussed in longitudinal discussion section at the end of this chapter. As Table 4.14 shows, items that have critical values for

individual parameter estimates related to them do not differ across the models. In other words, it is almost the same items that land around the upper/lower boundaries of factor loadings and R-square values with slight differences from one model to the other. None of these items are available for further qualitative analyses. Moreover, correlations among factors are very high across all 1-level models. This provides evidence for a unidimensional model. Thus, it is concluded that individual parameter estimate results for the PISA 2006 mathematics items imply that mathematical literacy is a unidimensional construct that unifies all three dimensions when 1-level and unidimensional models are evaluated. Although all of the items could be assigned to a different category in terms of different dimensions, they essentially measure the same overall construct: mathematical literacy. The factors loadings of level-1 latent variables are very high for 2-level models. This, however, provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors are good measures of the level-2 variable, GML (general mathematical literacy) in terms of the three dimensions. These results, considered altogether, do not contradict the multidimensionality of the PISA 2006 mathematics items. However, there is stronger evidence that supports the unidimensionality.

Model Comparisons

As mentioned above, DIFFTEST method is used to compare nested models (Muthén & Muthén, 2012). To remedy large sample size effect for significant results, an extension of ΔCFI method (Cheung & Rensvold, 2002) is used. To revisit the criterion, a ΔCFI less than or equal to -0.01 gives a significant result, which provides evidence less

restrictive model. In case of non-significant results for both DIFFTEST test and ΔCFI method, the conclusion is the same: although the fit of models are the same, more restrictive model is preferred over the less restrictive one for the data because it does not lose significant amount of fit.

Table 4.14. *Summary of individual parameter estimates for 2006*

Models	**Item loadings Low (<0.400)**	**Item loadings High (>0.800)**	**Items with low R-square (<0.250)**	**Correlations b/w L-1 factors**	**L-1 loadings onto L-2**
Model 1: 1F-GML	M24	M28, M37, M45, M80	M01, M24, M39, M40, M51, M65, M68, M75, M84	N/A	N/A
Model 2: 1-L Content	-	M15, M27, M28, M37, M45, M48, M80	M24, M40, M51, M65, M75	>0.846	N/A
Model 3: 1-L Process	M24	M15, M28, M37, M45, M80	M24, M39, M40, M51, M65, M68, M75, M84	>0.906	N/A
Model 4: 1-L Context	M24	M15, M28, M37, M45, M48, M80	M24, M40, M51, M65, M68, M75, M84	>0.924	N/A
Model 5: 2-L Content	-	M15, M27, M28, M37, M45, M48, M80	M24, M40, M51, M65, M75	N/A	>0.906
Model 6: 2-L Process	M24	M15, M28, M37, M45, M80	M24, M39, M40, M51, M65, M68, M75, M84	N/A	>0.958
Model 7: 2-L Context	M24	M15, M28, M37, M45, M48, M80	M24, M40, M51, M65, M68, M75, M84	N/A	>0.967

Table 4.15. *DIFFTEST results for 2006 models*

1F-GML versus	**L2-Content**	**L2-Process**	**L2-Context**
Value	105.733	8.150	34.841
Δ*df*	4	1	4
p-value	0.0000*	0.0043*	0.0000*
	L2-content vs. **4F-Content**	**L2-Process** vs. **3F-Process**	**L2-context** vs. **4F-Context**
Value	2.891	15.100	2.214
Δ*df*	2	2	2
p-value	0.2356	0.0005*	0.3306
1F-GML versus	**4F-Content**	**3F-Process**	**4F-Content**
Value	111.713	22.529	34.834
Δ*df*	6	3	6
p-value	0.0000*	0.0001*	0.0000*

* Significant at 0.05

As in the 2003 cycle, there are seven models compared to each other in groups of two. The DIFFTEST results are given in Table 4.15 and ΔCFI results are given Table 4.16. According to the DIFFTEST results, only two comparisons are found to be non-significant: L2-Content versus L1-Content and L2-Context versus L1-Context. So, 2-level and 1-level models perform about the same for the content and context dimensions. However, the more restrictive models (i.e., 2-level models) are preferable because of their parsimony. All other comparisons are found significant, which, however, might be resulted from a large sample size. Therefore, ΔCFI method is further applied for these comparisons. The ΔCFI results show that four comparisons are significant: 1F-GML

versus L2-Content, 1F-GML versus 4F-Content, 1F-GML versus L2-Context, and 1F-GML versus 4F-Context. The L2-Content and L1-Content models are preferred to the 1F-GML (ΔCFI = -0.02). The L2-Context and L1-Context models are preferred to the 1F-GML (ΔCFI = -0.01). When the 1-level and 2-level models are compared to each other for the content and context dimensions, it is found that model comparison is not significant (ΔCFI = 0). Therefore, the more restrictive (2-level) models are better for the content and context dimensions. ΔCFI results also support the non-significant result obtained through DIFFTEST. These results however somewhat different than model comparison results for 2003. The differences will be discussed in the longitudinal section at the end of this chapter.

Table 4.16. Δ*CFI results for 2006 model comparisons*

	1F-GML vs. L2-Content	**1F-GML vs. L2-Process**	**1F-GML vs. L2-Context**
ΔCFI	-0.02*	0	-0.01*
	L2-Content vs. 4F-Content	**L2-Process vs. 3F-Process**	**L2-Context vs. 4F-Context**
ΔCFI	0	0	0
	1F-GML vs. 4F-Content	**1F-GML vs. 3F-Process**	**1F-GML vs. 4F-Context**
ΔCFI	-0.02*	0	-0.01*

* Significant at -0.01

To summarize the model comparison results, it would make sense to take them at hand according to each dimension. For the content and context dimensions, the 2-level model is preferred to the1-level correlated four-factors model, which is preferred to the unidimensional model. Although there is no specific ranking for the process dimension

because of no significant differentiation among different models, there is a preference. The more restrictive model is preferred. The unidimensional model is preferred over the 2-level process and 1-level correlated factors (3F-Process) models.

Content: 2-Level Model > 1-Level Model > 1F-GML Model
Process: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model
Context: 2-Level Model > 1-Level Model > 1F-GML Model

Figure 4.2. Model Comparisons for 2006

Figure 4.2 illustrates the summary of models comparisons for 2006. The ">" sign refers to a ranking where the model on the left is preferred to the one on the right. The "≥" sign is used for models that are indifferent in terms of their fit performance to the data but the model on the left is always preferred for its parsimony.

Summary of 2006 Results

First, all seven models for the PISA 2006 mathematics items fit the data pretty well. This implies that the dimensional structure of the PISA 2006 mathematics items do not contradict any of the models proposed by the PISA mathematics framework. Connecting these results to the first research question (response-framework correspondence), overall model-fit results do not indicate any contradiction for the correspondence between dimensional structure of the mathematics items and mathematical literacy framework. However, there is more evidence supporting the unidimensionality of the mathematics items. To better understand how each model corresponds to individual items, parameter estimates were evaluated. First of all, all

parameter estimates were found significant for each of the seven models, meaning that all models provide a good account for factor loadings. In other words, each mathematics item plays an important role in different dimensionality models. This further support the evidence that mathematics framework is reflected in the PISA 2006 mathematics items through the student response data with respect to the three dimensions.

What is interesting, however, is that the items whose factor loadings and R-square values are around lower and upper boundaries of critical values are almost the same items in all models (see Table 4.14). Unfortunately, none of these items are available for any further analysis. There are slight differences between the 2003 and 2006 results in terms of these critical items, which will be discussed later in the chapter. Moreover, the factor loading values for each mathematics item are almost identical across models (see Appendix C for individual factor loadings). It can be concluded that the items psychometrically behave the same in all dimensions. That is, regardless of the dimension (e.g., whether it is an uncertainty (UN), a reflection (REF), or scientific (SCI) question) an item's loading and explained variation by a specific model is the same.

Correlations between the latent variables for 1-level models are very high for all dimensions (see Table 4.11, Table 4.12, and Table 4.13). High correlations provide evidence supporting a unidimensional model rather than either multidimensional correlated-factor models for each the three dimensions. Correlations coefficients slightly decreased when level-1 models are compared to their counterpart in 2003. This might be caused because of less number of items. It might also be possible that the items there are not included in 2006 (36 in total) are more related to each than others items. The majority

of these items 2003 items that are not included in 2006 are released (see Appendix C). However, since none of the other 2006 items are released, the two sets of items could not be further compared qualitatively.

The level-1 factors' factor loadings onto level-2 latent variable (GML) are very high for all dimensions. This provides a strong support for higher order structure of the construct that underlies the mathematics items.

Model comparisons also indicate that the unidimensional model is preferable over the 1-level and 2-level models for the process dimension. For the content and contexts dimensions, the 2-level model is preferred to the1-level correlated four-factors model, which is preferred to the unidimensional model. QT (quantity), SS (space and shape), CR (change and relationship), and UN (uncertainty) dimensions are a good measure of the global construct, GML (general mathematical literacy) in terms of content. Similarly, PER (personal), PUB (public), EDOP (educational/occupational), and SCI (scientific) dimensions are also good measures of the GML construct in terms of context. Revisiting the answer to the second research question, the unidimensional model is preferable for process dimension. However, in terms of content and context structures of mathematics items, the 2-level models better represent the dimensional structure of the PISA 2006 mathematics items. To conclude, unidimensional structure explains the mathematics items best when process dimension is taken into account. However, a rather multidimensional structure represents the mathematics items in terms of the content and context dimensions. The results provide evidence for both unidimensionality and multidimensionality. However, there is additional evidence for the unidimensionality

(high correlations between the 1-level models). Thus, overall results conclude that although the multidimensionality cannot be disproven, there is stronger evidence that supports the unidimensionality of the PISA mathematics items.

CFA Results: 2009 Cycle

Assessment of Models

The model-fit indices of models for 2009 are given in Table 4.17. The χ^2 statistics for each model range from 711.152 (with *df = 554*) to 743.474 (with *df = 760*). As in both the 2003 and 2006 cycles, chi-square tests for all models are significant ($p < 0.001$). Again, although these results suggest a poor fit, the sample size might have inflated the statistics. Therefore, we should evaluate other fit indices. TLI and CFI values for all models are greater than the critical value of 0.95, as in the 2003 and 2006 cycles. Therefore, TLI and CFI values are in favor of a good fit for all of the models. Moreover, RMSEA values for all models are found to be very close to zero: they range from 0.004 to 0.005. The probability that RMSEA is less than the cut-off value of 0.05 is almost 1. This means that all models fit the data fairly well. Although, WRMR value is greater than the critical value of 1.0 for all models, the overall fit of each model is found to be good for the 2009 cycle as well.

Individual Parameter Estimates

There were 35 mathematics items in 2009. Factor loadings and R-square values for these are given in the following sub-sections for each of the seven models as well as correlations between factors where it is applicable. These estimated values relate to the first research question and will help determine how the mathematics framework is

Table 4.17 *Model fit indices for 2009*

	Model 1: 1F-GML	**Model 2: 4F-Content**	**Model 3: 3F-Process**	**Model 4: 4F-Context**	**Model 5: L2-Content**	**Model 6: L2-Process**	**Model 7: L2-Context**
Chi-Square Test of Model Fit							
Value	743.474	711.152	741.596	729.428	713.741	742.587	731.871
Degrees of freedom	560	554	557	554	556	559	556
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.000
CFI/TLI							
CFI	**0.980**	**0.983**	**0.980**	**0.981**	**0.983**	**0.980**	**0.981**
TLI	**0.979**	**0.982**	**0.979**	**0.980**	**0.982**	**0.979**	**0.980**
RMSEA (Root Mean Square Error of Approximation)							
Estimate	**0.005**	**0.004**	**0.005**	**0.005**	**0.004**	**0.005**	**0.005**
90 Percent C.I.	0.004-0.005	0.003-0.005	0.004-0.006	0.004-0.005	0.003-0.005	0.004-0.005	0.004-0.005
Probability RMSEA <= .05	1.000	1.0000	1.000	1.000	1.0000	1.000	1.000
WRMR (Weighted Root Mean Square Residual)	1.168	1.125	1.167	1.153	1.127	1.167	1.155

Values that satisfy the criteria for good model fit are bold-faced.

reflected in the PISA 2009 mathematics items through the student response data with respect to the three dimensions.

Model 1: Single-factor model

Among the 35 items, only two items, M65 and M75, have a low standardized factor loading: 0.351 and 0.392, respectively. The factor loadings range from 0.411 to 0.913. The items M27 and M28 have the greatest factor loadings (higher than 0.800). The factor loadings for the majority of the items are higher than 0.500. All factor loadings including that of M65 and M75, which have low factor loadings, are significant ($p <$ 0.001). This unidimensional model explains 25% or more of the variation in the responses to the most of the items (26 out of 35) in 2009. The amount of variance explained falls below 25% for the items M01, M20, M36, M40, M51, M65, M66, M75, and M83. This signals the presence of other factors causing the variation above and beyond the unidimensional model for these items. None of the items whose factor loadings or R-square values are either at the lower or upper end of the scale are released for further analysis. However, the variation this unidimensional model explains is found to be statistically significant for all items including those, which have low R-square values. This further provides additional support for the unidimensionality of the PISA 2009 mathematics items.

Model 2: 1-level (four-factor) content model

Factor loadings in 1-level content factor range from 0.367 to 0.945. Only item M75 has a loading smaller than 0.400 onto QT latent variable. All items load onto their related factors (QT, CR, SS, and UN) at a statistically significant level ($p < 0.001$). The

items M27 and M28 have the greatest factor loadings (higher than 0.800). The factor loadings for the majority of the items are higher than 0.500. There are only five items whose variation could not be explained by the four-factor content model more than 25%. These items are M40, M51, M65, M66, and M75. Their R-square values are significant, although they are low, ranging from 0.135 to 0.232. All of the variations explained by this model are significant.

Table 4.18. *Correlations between 2009 content dimensions*

	QT	SS	CR	UN
QT	1			
SS	0.876	1		
CR	0.917	0.898	1	
UN	0.883	0.860	0.962	1

Four content factors correlate with each other significantly high. Correlation coefficients vary from 0.860 to 0.962 as shown in Table 4.18. CR and UN factors correlate the most while the lowest correlation is between SS and UN. This shows that all of the four content factors essentially behave as one unifying construct rather than four different latent factors. This evidence provides support for a unidimensional model, which also showed good model-fit results as described in the previous section, rather than a multidimensional content model.

Model 3: 1-level (three-factor) process model

In this three-factor process model, all of the observed indicators load onto their related factors with a 0.400 or higher loading value except for two indicators: items M65 and M75. Both of these items load onto REP factor with loading values of 0.352 and 0.392, respectively. The factor loadings for the items M27 and M28 are higher than

0.800. The factor loadings for the majority of the items are above 0.500. Only 9 items (M01, M20, M36, M40, M51, M65, M66, M75, and M83) have R-square values 0.249 or less. This means 25% or more of the variance in most of the items (26 out of 35) is explained by their related factors via the three-factor process model. Since all parameter estimates are statistically significant despite some low values, this 3F-Process model also provides a good fit for the PISA 2009 mathematics items through student response data in terms of individual parameters.

Table 4.19. *Correlations between 2009 process dimensions*

	REP	**CON**	**REF**
REP	1		
CON	0.988	1	
REF	0.986	0.981	1

The three latent variables (REP, CON, and REF) are significantly correlated to each other. As shown in Table 4.19, the correlation coefficients are found to be higher than 0.90. All of three process dimensions essentially look like one unifying construct rather than three different latent factors. This provides support towards a unidimensional model for the PISA 2009 mathematics items.

Model 4: 1-level (four-factor) context model

Table 4.20. *Correlations between 2009 context dimensions*

	PER	**EDOP**	**PUB**	**SCI**
PER	1			
EDOP	0.948	1		
PUB	0.937	0.907	1	
SCI	0.947	0.974	0.941	1

Factor loadings for observed indicators in the 1-level context model range from 0.273 (M75 on PER) to 0.920 (M28 on PUB). The items M01, M36, M40, M65, and M75 have a factor loading less than the critical value (0.400). On the other hand, the factor loadings for the items M27 and M28 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. R-square values for the observed indicators vary from 0.074 (M75) to 0.847 (M28). All but 8 indicators have an R-square value greater than 0.250. The amount of variation in the student responses explained by this model falls below 25% for items M01, M20, M36, M40, M51, M65, M66, and M75. However, all of the factor loadings, correlations, and R-square values are statistically significant. Therefore, it is concluded that the 4F-Context model provides a good account for the relationships among the PISA 2009 mathematics items. However, high correlations between factors provide evidence favoring a unidimensional model over this 1-level context model as shown in Table 4.20. Four context factors PER, EDOP, PUB, and SCI are highly correlated with each other. The correlation coefficients range from 0.907 to 0.974. This indicates that all of the four context factors essentially behave as one unifying construct rather than four different latent factors. This provides evidence that supports a unidimensional model much preferred to a correlated factors context model.

Model 5: 2-level content model

In this model, all items load onto their related level-1 factors more than the cut-off value of 0.400 except the item M75, which loads onto QT factor with 0.368. The items M27 and M28 have the greatest factor loadings (higher than 0.800). The factor loadings for the majority of the items are higher than 0.500. Also, all of the level-1 content factors

(QT, SS, CR, and UN) load highly onto the level-2 latent variable GML, ranging from 0.917 to 0.986. This provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors QT, SS, CR, and UN are good measures of the level-2 variable, GML (general mathematical literacy). The proportions of variance in the level-1 factors explained by the level-2 factor are 0.880, 0.841, 0.972, and 0.912 for QT, SS, CR, and UN, respectively. This indicates that the higher-order structure explains the covariances among the level-1 factors very well. R-square values range from 0.135 to 0.894. All but five of the items (M40, M51, M65, M66, and M75) have an R-square value of 0.250 or greater. None of these items are available for further analysis. To sum up, individual estimate results reveal that the 2-level content model provides a good fit for the PISA 2009 mathematics items.

Model 6: 2-level process model

All observed indicators have a factor loading of 0.400 or higher except items M65 and M75, which load onto CON factor with loadings of 0.353 and 0.397, respectively. The factor loadings for the items M27 and M28 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. All of the level-1 content factors (REP, CON, and REF) load highly onto the level-2 latent variable GML, ranging from 0.989 to 0.993. This provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors REP, CON, and REF are good measures of the level-2 variable, GML (general mathematical literacy) in terms of process dimension. The proportions of variation in the level-1 factors REP, CON, and, REF explained by level-2 factor are 0.978, 0.986, and 0.992, respectively. This indicates

that the 2-level structure explains the covariances among the level-1 latent variables very well. The R-square values for all but 9 items are higher than 0.250. That is, 2-level process model explains 25% or more of the variation in the responses to the most of the items. All of the factor loadings and R-square values are statistically significant although some of their estimates are below the acceptable values. Therefore, the 2-level model provides a good structural representation of the PISA 2009 mathematics items through the student response data in terms of process dimension.

Model 7: 2-level context model

Factor loadings for this model range from 0.274 to 0.920. All items load onto their first level context factors with more than 0.400 factor loading values except five items: M01, M36, M40, M65, and M75. On the other hand, the factor loadings for the items M27 and M28 are higher than 0.800. The factor loadings for the majority of the items are above 0.500. The level-1 context factors SCI, PUB, EDOP, and PER load highly onto the level-2 latent variable GML with values 0.939, 0.974, 0.986, and 0.994, respectively. The proportions of variation in the level-1 context factors explained by the level-2 factor are 0.882, 0.949, 0.972, and 0.988, respectively. This indicates that the 2-level structure provides a good account for the covariances among the level-1 latent variables for context dimension. This provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors PER, EDOP, PUB, and SCI are good measures of the level-2 variable, GML (general mathematical literacy) in terms of context dimension. All but eight items have R-square values greater than 0.250. The items M01, M20, M36, M40, M51, M65, M66, and M75

have a low R-square value. However, their R-square values are significant. Therefore, 2-level model, overall, provides a good structural representation of the 2009 mathematics items in terms of context dimension.

Table 4.21 shows items that have critical values for their individual parameter estimates. None of these items are available for further qualitative analyses. Moreover, correlations among factors are very high across all 1-level models. These results provide evidence for a unidimensional model. Thus, it is concluded that individual parameter estimate results for unidimensional and 1-level models for the PISA 2009 mathematics items imply that mathematical literacy is a unidimensional construct that unifies all three dimensions (content, process, and context) and their sub-dimensions. Although all of the items could be assigned to a different category in terms of different dimensions, they essentially measure the same overall construct: mathematical literacy. The factor loadings of level-1 latent variables are very high for 2-level models. This, however, provides a strong support for higher order structure of the construct that underlies the mathematics items. That is, level-1 factors are good measures of the level-2 variable, GML (general mathematical literacy) in terms of the three dimensions. These results, considered altogether, do not contradict the multidimensional structure of the PISA 2009 mathematics items. However, there is more evidence that supports the unidimensionality.

Model Comparisons

As mentioned above, DIFFTEST method is used to compare nested models (Muthén & Muthén, 2012). To remedy large sample size effect for significant results, an extension of ΔCFI method (Cheung & Rensvold, 2002) is used. To revisit the

Table 4.21. *Summary of individual parameter estimates for 2009*

Models	Item loadings Low (<0.400)	Item loadings High (>0.800)	Items with low R-square (<0.250)	Correlations b/w L-1 factors	L-1 loadings onto L-2
Model 1: 1F-GML	M65, M75	M27, M28	M01, M20, M36, M40, M51, M65, M66, M75, M83	N/A	N/A
Model 2: 1-L Content	M75	M27, M28	M40, M51, M65, M66, M75	>0.860	N/A
Model 3: 1-L Process	M65, M75	M27, M28	M01, M20, M36, M40, M51, M65, M66, M75, M83	>0.981	N/A
Model 4: 1-L Context	M01, M36, M40, M65, M75	M27, M28	M01, M20, M36, M40, M51, M65, M66, M75	>0.907	N/A
Model 5: 2-L Content	M75	M27, M28	M24, M40, M51, M65, M75	N/A	>0.841
Model 6: 2-L Process	M65, M75	M27, M28	M01, M20, M36, M40, M51, M65, M66, M75, M83	N/A	>0.989
Model 7: 2-L Context	M01, M36, M40, M65, M75	M27, M28	M01, M20, M36, M40, M51, M65, M66, M75	N/A	>0.882

criterion, a ΔCFI less than or equal to -0.01 gives a significant result, which provides evidence for a less restrictive model. In case of non-significant results for both DIFFTEST test and ΔCFI method, the conclusion is the same: although the fit of models are the same, a more restrictive model is preferred over the less restrictive one for the data because it does not lose significant amount of fit.

Table 4.22. *DIFFTEST results for 2009 models*

1F-GML versus	**L2-Content**	**L2-Process**	**L2-Context**
Value	55.116	1.878	23.658
Δ*df*	4	1	4
p-value	0.0000*	0.1705	0.0001*
	L2-content vs. **4F-Content**	**L2-Process** vs. **3F-Process**	**L2-context** vs. **4F-Context**
Value	4.512	0.851	3.715
Δ*df*	2	2	2
p-value	0.1048	0.6535	0.1561
1F-GML versus	**4F-Content**	**3F-Process**	**4F-Content**
Value	60.436	1.667	26.756
Δ*df*	6	3	6
p-value	0.0000*	0.6442	0.0002*

* Significant at 0.05

As in the 2003 and 2006 cycles, there are seven models compared to each other in groups of two. The DIFFTEST results are given in Table 4.22 and ΔCFI results are given in Table 4.23. According to the DIFFTEST results, five comparisons are found to be non-significant: 1F-GML versus L2-Process, 1F-GML versus L1-Process, L2-Content versus L1-Content, L2-Process versus L1-Process, and L2-Context versus L1-Context. So, the two models in each pair of comparisons perform about the same. However, the more restrictive models are preferable because of their parsimony. Thus, the unidimensional model is preferred over the other multidimensional models of process dimension.

Similarly, 2-level models are preferred over 1-level models for all three dimensions (content, process, and context. The other four comparisons are found to be significant.

Table 4.23. Δ*CFI results for 2009 model comparisons*

	1F-GML vs. L2-Content	**1F-GML vs. L2-Process**	**1F-GML vs. L2-Context**
ΔCFI	-0.03*	0	-0.01*
	L2-Content vs. 4F-Content	**L2-Process vs. 3F-Process**	**L2-Context vs. 4F-Context**
ΔCFI	0	0	0
	1F-GML vs. 4F-Content	**1F-GML vs. 3F-Process**	**1F-GML vs. 4F-Context**
ΔCFI	-0.03*	0	-0.01*

* Significant at -0.01

The ΔCFI results revealed exactly the same conclusions as DIFFTEST results. Four comparisons are significant: 1F-GML versus L2-Content, 1F-GML versus 4F-Content, 1F-GML versus L2-Context, and 1F-GML versus 4F-Context. The L2-Content and L1-Content models are preferred to the 1F-GML (ΔCFI = -0.03). Likewise, the L2-Context and L1-Context models are preferred to the 1F-GML (ΔCFI = -0.01). When the 1-level and 2-level models are compared to each other for the content and context dimensions, it is found that model comparison is not significant (ΔCFI = 0). Therefore, the more restrictive (2-level) models are better for the content and context dimensions. These results however somewhat different than model comparison results for 2003. The differences will be discussed in the longitudinal section at the end of this chapter.

To summarize the model comparison results, it would make sense to take them at hand according to each dimension. For the content and context dimensions, the 2-level

model is preferred to the1-level correlated four-factors model, which is preferred to the unidimensional model. Although there is no specific ranking for the process dimension because of no significant differentiation among different models, there is a preference. The more restrictive model is preferred. The unidimensional model is preferred over the 2-level process and 1-level correlated factors (3F-Process) models.

Figure 4.3 illustrates the summary of models comparisons for 2006. The ">" sign refers to a ranking where the model on the left is preferred to the one on the right. The "≥" sign is used for models that are indifferent in terms of their fit performance to the data but the model on the left is always preferred for its parsimony.

Summary of 2009 Results

To begin with, all seven models for the PISA 2009 mathematics items fit the data pretty well as in the 2003 and 2006 cycles. This implies that the dimensionality of the PISA 2009 mathematics items do not contradict any of the models proposed by the PISA mathematics framework. Connecting these results to the first research question (response-framework correspondence), overall model-fit results do not indicate any contradiction for the correspondence between dimensional structure of the mathematics items and mathematical literacy framework. However, there is more evidence supporting the unidimensionality of the mathematics items. To better understand how each model corresponds to individual items, parameter estimates were evaluated. First of all, all parameter estimates were found significant for each of the seven models, meaning that all models provide a good account for factor loadings. In other words, each mathematics item plays an important role in different dimensionality models. This further supports the

evidence that the mathematics framework is reflected in the PISA 2009 mathematics items through the student response data with respect to the three dimensions.

Content: 2-Level Model > 1-Level Model > 1F-GML Model
Process: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model
Context: 2-Level Model > 1-Level Model > 1F-GML Model

Figure 4.3. Model Comparisons for 2009

What is different than the 2003 and 2006 results is that the items whose factor loadings and R-square values are around lower and upper boundaries of critical values are not as similar across all models as in 2003 and 2006 results (see Table 4.21). Unfortunately, none of these items are available for any further analysis. There are also some differences from the 2003 and 2006 results in terms of these critical items, which will be discussed later in the chapter. Moreover, the factor loading values for each mathematics item are almost identical across models (see Appendix C for individual factor loadings). It can be concluded that the items psychometrically behave the same in all dimensions. That is, regardless of the dimension (e.g., whether it is an uncertainty (UN), a reflection (REF), or scientific (SCI) question) an item's loading and explained variation by a specific model is the same.

Correlations between the latent variables for 1-level models are very high for all dimensions (see Table 4.18, Table 4.19, and Table 4.20). High correlations provide evidence supporting a unidimensional model rather than multidimensional correlated-factor models for each the three dimensions.

The level-1 factors' factor loadings onto level-2 latent variable (GML) are very high for all dimensions. This provides a strong support for a higher order structure of the construct that underlies the mathematics items.

Model comparisons also indicate that the unidimensional model is preferable over the 1-level and 2-level models for the process dimension. For the content and contexts dimensions, the 2-level model is preferred to the 1-level correlated four-factors model, which is preferred to the unidimensional model. QT (quantity), SS (space and shape), CR (change and relationship), and UN (uncertainty) dimensions are a good measure of the global construct, GML (general mathematical literacy) in terms of content. Similarly, PER (personal), PUB (public), EDOP (educational/occupational), and SCI (scientific) dimensions are also good measures of the GML construct in terms of context. Revisiting the answer to the second research question, the unidimensional model is preferable for process dimension. However, in terms of content and context structures of mathematics items, the 2-level models better represent the dimensional structure of PISA mathematics items. To conclude, unidimensional structure explains the mathematics items best when process dimension is taken into account. However, a rather multidimensional structure represents the mathematics items in terms of the content and context dimensions. The results provide evidence for both unidimensionality and multidimensionality. However, there is additional evidence for the unidimensionality (high correlations between the 1-level models). Thus, overall results conclude that although the multidimensionality cannot be disproven, there is stronger evidence that supports the unidimensionality of the PISA mathematics items.

Longitudinal Evaluation of Results

The third research question in this dissertation is about the change in the dimensional structure of the PISA mathematics items over time. This is basically comparing the answers to the first and the second research questions across the three cycles. The results related to the first research question, model-fit indices and individual parameter results, are first discussed across cycles. Then, the model comparison results, which relate to the second research question, are also compared across cycles.

First of all, overall results for model-fit for all seven models were found to be unchanged across cycles. There are some changes in values of fit indices; however, those changes are very small and negligible. For example, TLI/CFI values increase from 0.970's to 0.980's from 2003 to 2006 for all models. These values remain almost the same from 2006 and 2009. Thus, all seven models for the PISA mathematics items fit the data in all the cycles: 2003, 2006, and 2009, which implies that there is evidence supporting both unidimensionality and multidimensionality of the PISA mathematics items in terms of the content, process, and context dimensions. However, there is additional evidence for the unidimensionality (high correlations between the 1-level models). Thus, overall results conclude that although the multidimensionality cannot be disproven, there is stronger evidence that supports the unidimensionality of the PISA mathematics items.

There are some differences across cycles in terms of critical items, whose factor loadings or R-square values are either at the lower or upper end of the scale. For example, the items M75, M82, and M83 have low factor loadings for all seven models in 2003

Table 4.24. *Important parameter estimates across the cycles*

Models	Low item loadings			High item loadings			Low R-square Values			Correlations			L-1 loadings onto L-2		
	2003	2006	2009	2003	2006	2009	2003	2006	2009	2003	2006	2009	2003	2006	2009
Model 1: 1F-GML	M75, M82, M83	M24	M65, M75	M28, M45, M48, M80	M28, M37, M45, M80	M27, M28	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	M01, M24, M39, M40, M51, M65, M68, M75, M84	M01, M20, M36, M40, M51, M65, M66, M75, M83	N/A	N/A	N/A	N/A	N/A	N/A
Model 2: 1L-Content	M75, M82, M83	-	M75	M07, M27, M28, M45, M48, M63, M80	M15, M27, M28, M37, M45, M48, M80	M27, M28	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75 M82, M83	M24, M40, M51, M65, M75	M40, M51, M65, M66, M75	>0.905	>0.846	>0.860	N/A	N/A	N/A
Model 3: 1L-Process	M75, M82, M83	M24	M65, M75	M28, M45, M48, M80	M15, M28, M37, M45, M80	M27, M28	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	M24, M39, M40, M51, M65, M68, M75, M84	M01, M20, M36, M40, M51, M65, M66, M75, M83	>0.966	>0.906	>0.981	N/A	N/A	N/A
Model 4: 1L-Context	M75, M82, M83	M24	M01, M36, M40, M65, M75	M15, M28, M45, M48, M80	M15, M28, M37, M45, M48, M80	M27, M28	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	M24, M40, M51, M65, M68, M75, M84	M01, M20, M36, M40, M51, M65, M66, M75	>0.939	>0.924	>0.907	N/A	N/A	N/A
Model 5: 2L-Content	M75, M82, M83	-	M75	M07, M27, M28, M45, M48, M63, M80	M15, M27, M28, M37, M45, M48, M80	M27, M28	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75 M82, M83	M24, M40, M51, M65, M75	M24, M40, M51, M65, M75	N/A	N/A	N/A	>0.914	>0.906	>0.841
Model 6: 2L-Process	M75, M82, M83	M24	M65, M75	M28, M45, M48, M80	M15, M28, M37, M45, M80	M27, M28	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	M24, M39, M40, M51, M65, M68, M75, M84	M01, M20, M36, M40, M51, M65, M66, M75, M83	N/A	N/A	N/A	>0.968	>0.958	>0.989
Model 7: 2L-Context	M75, M82, M83	M24	M01, M36, M40, M65, M75	M15, M28, M45, M48, M80	M15, M28, M37, M45, M48, M80	M27, M28	M20, M24, M36, M40, M41, M57, M65, M66, M68, M75, M82, M83, M84	M24, M40, M51, M65, M68, M75, M84	M01, M20, M36, M40, M51, M65, M66, M75	N/A	N/A	N/A	>0.967	>0.967	>0.882

while only the item M24 has the lowest factor loading in 2006 and so do the items M65 and M75 in 2009. On the other hand, there are many commonalities, too. For example, the item M28 is among the highest factor loadings in all cycles. Table 4.24 gives a full comparison of interesting items by models and cycles. Identifying interesting items is important because one could determine what a good/poor mathematics item looks like in terms of a measure of the construct, mathematical literacy's different dimensions. Further analyses that are qualitative in nature are needed to make those judgments, but unfortunately, since none of these interesting items were made public by PISA, those analyses could not be included in this dissertation.

Appendix C provides all the factor loadings by each item for different cycles and models. Individual factor loading values for each mathematics item do not change much in value across cycles or across models. This is very interesting because this means that when an item has a high factor loading in one model, then it tends to have high factor loadings in the other models as well. That is, the degree to which that item is explained by the different dimensions of the construct, general mathematical literacy, in all seven models is about the same. Having such stability in the factor loadings indicates the consistency of the PISA mathematics items and the PISA mathematics framework across cycles.

Model comparison results for 2006 and 2009 revealed similar conclusions. For the content and context dimensions, the 2-level model is preferred to the1-level correlated four-factors model, which is preferred to the unidimensional model. Although there is no specific ranking for the process dimension because of no significant differentiation

among different models, there is a preference. The more restrictive model is preferred. The unidimensional model is preferred over the 2-level process and 1-level correlated factors (3F-Process) models. The same conclusions for the content and process dimensions apply in the year 2003. However, for the 2003 context dimension, the results were different. There is no significant differentiation among different models for 2003 context dimension. However, the more restrictive model, 1F-GML, is preferred.

When each dimension is evaluated longitudinally, it is found that the content dimension is stable across cycles. The 2-level model is always better. The results are also consistent for the process dimension across cycles. However, the unidimensional structure is better supported by evidence than the higher order process structures. Lastly, it appears that the context dimension showed some inconsistency in terms of best structural representation of mathematics items. In 2003, the unidimensional model is preferable over the multidimensional context models. In 2006 and 2009, 2-level model is found significantly preferable to the unidimensional model. However, when all of it is taken together, it could be concluded that 2-level model is the best structural representation for the context dimension.

To conclude, model comparison results revealed pretty consistent results across cycles. The 2-level model is found performing better with the mathematics items in terms of the content and the context dimensions. That is, a multidimensional content and context models are preferable to the unidimensional model. However, this is not the case for the process dimension. Multidimensional process models are not preferred to the unidimensional model. These conclusions imply that the dimensional nature of content

and context dimensions in the assessment framework is reflected in the PISA mathematics items through the student responses. QT (quantity), SS (space and shape), CR (change and relationship), and UN (uncertainty) dimensions are a good measure of the global construct, GML (general mathematical literacy) in terms of content. Similarly, PER (personal), PUB (public), EDOP (educational/occupational), and SCI (scientific) dimensions are also good measures of the GML construct in terms of context.

The dimensional structure of the process dimension given in the mathematics framework is not clearly reflected in the mathematics items. Therefore, as provided in the PISA's mathematics assessment framework, REP (reproduction), CON (connections), and REF (reflection) dimensions might not be a good measure for general mathematical literacy. There are two possible explanations for this. The process dimension might not be as fully developed as the other two dimensions. The process dimension comprises of 8 competency clusters: mathematical thinking and reasoning, mathematical argumentation, modeling, problem posing and solving, representation, symbols and formalism, communication, and aids and tools (OECD, 2009a). In order to operationalize these competencies, they are grouped into three process dimensions. It might be the case that these clusters do not differ much so they behave as one construct. Secondly, the items might not be well categorized in terms of these competency clusters. That is, a question might be drawn significantly on more than one dimensions of process. Therefore, this dimension and item categorizations based on it should be revisited.

Summary of Results

The results in this chapter are organized in the following order: model-fit indices, individual item parameters, and model comparisons. The summary of the overall results here follows the same organization. Results for model-fit indices and individual parameter estimates address the first research question: What is the correspondence between the dimensional structure of the PISA mathematics items and PISA's mathematical literacy assessment framework in terms of the content, process, and context dimensions? Model comparison results in each cycle explore the models that best represent the dimensional structure of the PISA mathematics items in terms of different dimensions. These relate to the second research question. Finally, looking across the cycles to see how different models change over time provides the longitudinal aspect and addresses the last research question. A summary of results is given for each cycle at the end of each results sub-section. Although the results could nicely be partitioned into sections in order to draw conclusions for this study, all the evidence gathered has to be considered as a whole. Overall results are summarized next.

All seven models, including the unidimensional and multidimensional, for the PISA mathematics items were found a good fit for all three implementation cycles: 2003, 2006, and 2009. In other words, an analysis of the mathematics items does not contradict any of the models proposed for the dimensionality of PISA mathematics framework. Relating these results to the first research question (response-framework correspondence), overall model-fit results indicate a good reflection of the mathematical literacy framework in the structural representation of the PISA mathematics items

through student responses. This conclusion implies that there is evidence supporting both the unidimensionality and multidimensionality of mathematics framework in terms of the content, process, and context dimensions.

Second, all of the parameter estimates are found significant in each model and each cycle, meaning that all models provide a good account for factor loadings. In other words, each mathematics item plays an important role in different dimensionality models across cycles. This further supports that the mathematics framework is reflected in the PISA mathematics items through the student response data with respect to the three dimensions. There are some differences across cycles in terms of critical items whose factor loadings or R-square values are either at the lower or upper end of the scale. However, further item analysis to study what might cause these differences was not possible because none of these items have been released by PISA. Limited number of items has been released. Sample released items are given in Appendix B.

Model comparison results are very consistent for the content and context dimensions across the cycles. The 2-level model is found to be performing better with the PISA mathematics items in terms of the content and the context dimensions. That is, a multidimensional content and context models are more plausible than the unidimensional model. However, this is not the case for the process dimension, where the unidimensional model is preferred to the multidimensional models.

Chapter 5: Discussion and Conclusions

Mathematical literacy is defined as a multidimensional construct as it is widely documented in the literature (e.g., Kilpatrick, Swafford, & Findell, 2001; OECD, 2003; Ojose, 2011; Steen, 2001). When assessing students' mathematical literacy, whether the focus is in the classroom or large-scale, assessment tasks should operationalize mathematical literacy as a multidimensional construct rather than a single measure (e.g., general mathematics ability). A rigorous assessment design in mathematics requires designers (e.g., mathematics teachers, educators, and researchers) to pay careful attention to the connection among (1) the definition and nature of mathematical literacy as a construct and how people learn it, (2) development of assessment tasks that elicit mathematical literacy, and (3) an interpretation framework, which relate with the construct, of responses (NRC, 2001). This connection between the important components of an assessment design could only be ensured through a well-developed theoretical framework and empirical evidence that shows this theoretical framework is well-reflected in the assessment structure when observing students' responses. Therefore, the multidimensional structure of a mathematical literacy framework should reflect itself in its assessment.

What this connection implies is that an item should be designed to be a measure of mathematical literacy in a way that is defined in the assessment framework, and that this construct should be elicited in students' responses to the item. After the design, each item should statistically be investigated to ensure it measures what it is supposed to measure. Moreover, the test on the collection of items, as a whole, should reflect the

structure of the construct being measured, which should be consistent with the theoretical framework supporting the construct being measured. Measuring what is intended to measure is referred to as the validity of an assessment (AERA, APA, NCME, 1999; Cizek, Rosenberg, & Koons 2008; Loveinger, 1957; Messick, 1989). Dimensionality analyses provide empirical evidence for the correspondence between the theoretical framework and the assessment instrument. This, in turn, also means providing evidence for the construct validity of a set of items given in an assessment. Thus, dimensionality analysis is seen as very important for construct validation of an assessment. It is only possible to analyze the dimensionality of a set of items through the use of student responses to the items. When there is a strong prior expectation about the dimensional structure of an assessment (i.e., when there is a robust framework), confirmatory factor analysis (CFA) is used to analyze the dimensionality of the set of items through student responses (Tate, 2002) and verify for structural consistency.

The psychometric techniques that are used to calibrate items and produce final performance scores for individuals, and average performance scores for groups, comparison purposes might rely on the some basic assumptions about the dimensionality of an assessment instrument. The majority of contemporary tests are based on IRT models that assume unidimensionality (e.g., Rasch models). If the structure of an assessment designed using a Rasch model fails to satisfy this assumption of unidimensionality, then the interpretation of performance results might be inaccurate and misleading, generating damages in the consequential validity of the test (Messick, 1989).

Therefore, the dimensionality analysis is an important aspect of a validity study of an assessment instrument.

PISA provides a very rigorous mathematical literacy framework. In this framework, mathematical literacy is defined as a multidimensional construct comprising of the content, process, and context dimensions (OECD, 2009a). However, whether this multidimensionality is reflected in its mathematics assessment has not been widely studied. Schwab (2007) investigated the scientific literacy and dimensionality of the PISA 2003 science items. She found that student responses to the PISA 2003 science items reflected a unidimensional structure. Thus, the multidimensionality of PISA's scientific literacy framework is not reflected in the PISA 2003 science items. Ekmekci and Carmona (2012) explored the factorial structure of the PISA 2003 mathematics items for students in the United States. They detected unidimensionality in the PISA 2003 mathematics items as well. Both of these studies had a narrower focus, using only a subset of the available data. The first study used the student response data from English speaking countries (Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States). The latter study looked only at the U.S. data for PISA 2003 mathematics items.

This study investigated the dimensionality of the PISA mathematics items for three different cycles (2003, 2006, and 2009) using the student responses from 30 OECD countries. Dimensionality analyses were conducted through CFA methods. Seven CFA models were proposed based on the OECD's mathematical literacy framework. These models were then compared to each other to test the hypotheses about the dimensional

structure of PISA mathematics items for the three cycles. The models were: single factor model (interpreting the general mathematical literacy, GML, as the only latent factor), 1-level (four-factor) content model, 1-level (three-factor) process model, 1-level (four-factor) context model, 2-level (higher order) content model, 2-level (higher order) process model, and 2-level (higher order) context model.

CFA analyses comprised three parts for each cycle: fit of models, individual parameter estimates, and model comparisons. Model-fit indices and individual parameter estimates for each model in each cycle provided evidence to address the first research question (correspondence between the framework and student responses). Model comparisons relate to the second research question (the best representation of dimensional structure). Finally, evaluating these results over time across the cycles and making longitudinal conclusions addressed the third research question. Answers to the research questions and their implications for assessment in mathematics are given below.

Research Questions and Conclusions

The first research question guiding this study is:

> What is the correspondence between the dimensional structure of the PISA mathematics items and PISA's mathematical literacy assessment framework in terms of the content, process, and context dimensions?

Seven models (including the unidimensional and six multidimensional models) were proposed based on the theoretical framework for PISA's mathematical literacy domain. All models for the PISA mathematics items were found to have a good model-fit when analyzing the student responses to the PISA mathematics items for three implementation

cycles: 2003, 2006, and 2009. This shows that the PISA mathematics items do not contradict any of the models proposed for their dimensionality based on the PISA mathematics framework. This conclusion implies that there is evidence supporting both the unidimensionality and multidimensionality of the PISA mathematics items in terms of the content, process, and context dimensions. However, the answer to the first research question becomes clearer after the relationships among the PISA mathematics items are further evaluated through their individual parameter estimates.

All of the parameter estimates are found to be significant for each of the seven models and for three implementation cycles (2003, 2006, and 2009), meaning that all models provide a good account for the relationship among the PISA mathematics items. In other words, each mathematics item plays an important role in different dimensionality models across cycles. Moreover, the factor loadings for level-1 variables were very high in all 2-level models for all cycles. This supports the evidence that the multidimensional nature of the mathematics framework is reflected in the mathematics items with respect to the three dimensions. On the other hand, correlations between the latent variable in all 1-level models were very high, supporting the existence of unidimensionality. This is consistent with CFA model-fit results, in which the unidimensional models fit the response data for the mathematics items well enough.

Although the results could nicely be partitioned into sections in order to draw conclusions for this study, it is important to look at and make sense of all the evidence in its totality. When considered as a whole, the results indicate evidence supporting both unidimensionality and multidimensionality. However, stronger evidence was found to

support the unidimensionality of the PISA mathematics items. This satisfies the unidimensionality assumption for the IRT, contributing to the validation of the use of the Rasch model to produce and scale performance scores for the PISA mathematical literacy assessment. However, it is not supportive of the theoretical framework of the multidimensional nature of mathematical literacy.

Therefore, the connection between the interpretation and cognition components (NRC, 2001) of PISA mathematics assessment is not very strong. The reflection of the multidimensional nature of mathematical literacy in the PISA's framework on the mathematics items is subtle. Thus, the answer to the first research questions is that the OECD's mathematical literacy framework is reflected only minimally in the PISA mathematics assessment.

Although the models for multidimensionality agree with the mathematics items, stronger evidence for the unidimensionality implies *essential unidimensionality* (Stout, 1987). Therefore, weaker indication of multidimensionality might stem from the minor dimensions around the dominant ability (Tate, 2002) of general mathematics literacy (GML). That is, the construct GML is so dominant that other dimensions could only minimally explain the relationship between the mathematics items. The relaxation of the strict unidimensionality (Brandt, 2008; Stout, 1987, 1990; Tate 2002) reconciles the evidence for both unidimensionality and multidimensionality found in the PISA mathematics items.

To conclude, large-scale assessments could be valuable only if they are well designed and appropriately used (NRC, 2001). One of the important questions to be

addressed in a good assessment design is what views of mathematical literacy are these large-scale assessments designed to reflect? The weak connection between PISA's assessment framework and PISA mathematics items implies a discrepancy between the cognition and interpretation components, endangering the construct validity of PISA. Moreover, consequences of this discrepancy might have direct and important impact on educational systems of countries because the ministries of education in participating countries take PISA results seriously. Thus, the ways that could strengthen the connection between the cognition and interpretation components of the PISA assessment design should be looked for.

The second research question guiding this study was:

> What is the best representation for the dimensional structure of the PISA mathematics items for implementation cycles 2003, 2006, and 2009?

The model comparison results provided evidence to address this question. The results for 2006 and 2009 produced identical conclusions for the content and the context dimensions: the 2-level content and context models are preferable to their 1-level counterparts (correlated four-factors model), which are preferable to the unidimensional model. This implies that multidimensionality is preferable to unidimensionality in terms of content and context dimensions in the 2006 and 2009 cycles. The same conclusion was reached for the content dimension in 2003. However, in terms of the context dimension in 2003, multidimensional models and the unidimensional model perform about the same. Thus, there is no specific ranking for the context dimension in 2003, although there is a preference for the more restrictive model, which is the unidimensional model.

Model comparison for the process dimension was consistent across all cycles. No significant difference was found among the unidimensional, 1-level, and 2-level process models. As for the context dimension in 2003, the more restrictive model is preferred. The unidimensional model is preferred over the 2-level process and 1-level correlated factors (3F-Process) models.

The results, then, imply different things for each of the three dimensions. In terms of the content dimension: QT (quantity), SS (space and shape), CR (change and relationship), and UN (uncertainty) dimensions are a good measure of the global construct, GML (general mathematical literacy). This conclusion is consistent across all cycles. In terms of the process dimension, a unidimensional model is preferred although all of the models were found to be equally good with respect to response data for the mathematics items. So, for the process dimensions, there is not a single best representation for the dimensional structure of the PISA mathematics items. This conclusion is also consistent throughout different cycles. There are two inconsistent conclusions, however, for the context dimension. In 2003, a unidimensional model is preferred although all of the models for context dimension were found to be equally good with respect to the student response data for the PISA mathematics items. In 2006 and 2009, PER (personal), PUB (public), EDOP (educational/occupational), and SCI (scientific) dimensions were found to be good measures of the GML construct in terms of the context dimension.

Therefore, results again provide evidence for both unidimensionality and multidimensionality in terms of different dimensions, giving a non-definitive answer to the second research question: the best representation of the dimensional structure of the PISA mathematics items depends on the dimension of mathematical literacy.

The multidimensionality with respect to some dimensions is more evident than it is for other dimension(s). Relating to the essential dimensionality discussion (Tate, 2002), it could be the case that content and context sub-dimensions behave as the minor dimensions around the major construct, GML, explaining the relationships among the PISA mathematics items to some extent. However, process sub-dimensions cannot explain any of the relationships among the mathematics items. A revision of the definition and organization of the process dimension or the classification of the mathematics items might be needed. Reconceptualization of the process dimension of the PISA mathematical literacy framework might help for reaching a better multidimensional process structure in the mathematics items.

The third research question guiding this study is:

How does the dimensional structure of the PISA mathematics items change over time?

One of the findings from the longitudinal analysis is that there were some differences across cycles in terms of the items whose factor loadings or R-square values were either too low or too high. Since none of these items are among the ones that have been released by PISA, it was not possible to do a finer grained-size item analysis by looking at the actual item (see Appendix B for sample released items). If they had been available, qualitative item analysis techniques could have been conduced to further investigate what might explain their variation (by the least amount or the most) through different models.

Secondly, the individual factor loadings for each mathematics item did not change much in value across cycles or across models. This is very interesting because this means that an item, depending on its quality, is explained by a different underlying factor

(different dimensions of mathematical literacy) by the same amount. An item loads by about the same amount onto GML factor (as single dimension) and other factors, say, quantity, (a content dimension), reproduction (a process dimension), and personal (a context dimension). This might be because all different factors behave as one since the items do not distinguish much among the factors (constructs) in terms of their relationships to the constructs as proposed in the CFA models. This seems to further support the unidimensionality of the PISA mathematics items.

Lastly, the model comparison results were very stable across cycles for the content and process dimensions. The 2-level model is found to be performing better with the response data for the PISA mathematics items in terms of the content for all cycles. That is, a multidimensional content structure is preferable to the unidimensional structure. However, for the process dimension, evidence supports a unidimensional structure over multidimensionality. The unidimensional model and the multidimensional process models were found to perform equally well in all cycles. In this case, unidimensional model is preferred for its parsimony.

For the context dimension, the results were not found to be as consistent. In 2003, the unidimensional model and the multidimensional context models performed equally well. In 2006 and 2009, the 2-level context models were found to be performing better.

To summarize, the mathematics items are found to be very stable across cycles in terms of their variation explained by the models. Explanations of the mathematics items by underlying latent constructs within each dimension were also stable. The model comparisons were also consistent across cycles except for the context dimension.

The overall conclusions are trifold. First, the PISA mathematics items are pretty stable, psychometrically speaking, across cycles. Second, the response data for the PISA mathematics items does not contradict their unidimensionality nor does it contradict their multidimensionality. Lastly, the multidimensionality of the content and context dimensions is prominent in all cycles, while the process dimension could not reflect its multidimensionality in any cycle.

Therefore, evidence supports both the IRT assumption of unidimensionality and the expectation of multidimensionality. However, there is stronger evidence for unidimensionality, validating the use of a Rasch model and contradicting the multidimensional nature of mathematical literacy as supported in the theoretical framework. In an attempt to reconcile the evidence for both unidimensionality of multidimensionality of a construct and an assessment instrument in practice, some researchers have addressed the relaxation of the strict dimensionality assumptions (e.g., Tate 2002). "It is universally recognized that the strict assumption of unidimensionality is always violated to some degree, and practitioners are usually willing to accept essentially unidimensional structure for an IRT-based test (Stout, 1987)" (Tate, 2002, p. 159).

Psychometrically speaking, although the findings are mixed in terms of statistical structure of the PISA mathematics items as concluded from evidence for both multidimensionality and unidimensionality, they are stable throughout different cycles, contributing to the construct validity (Loevinger, 1957) of the PISA assessment. These findings appear to be somewhat complicated. However, it is clear that none of the models contradict either the assumption of unidimensionality for IRT based assessments or

multidimensionality expectations provided in the rigorous assessment framework of PISA for mathematics domain. In practical terms, the following two statements hold:

- The unidimensionality assumption of IRT models is not violated in PISA.
- The multidimensional structure of the PISA mathematics framework is not violated.

While this study uses one of the most robust tools, CFA, to analyze the dimensionality (Tate, 2002) of one of the most robust and respected international assessment designs, PISA, in mathematics (OECD, 2009a), the results are, psychometrically speaking, somewhat complicated and ambiguous. By only looking at psychometric measures/methods, it is difficult to determine what the test is measuring. Qualitative analyses looking at the individual mathematics items are definitely needed to make sound judgments about the construct validity of each PISA mathematics item. It is important to determine what each item measures conceptually especially considering the high stakes decisions associated with the interpretation of results, which include national decisions on educational reform for many countries. Having said that, very few items have been released and made available for PISA mathematics domain, all of which had moderate to high factor loadings. Appendix B provides sample items. None of the items whose individual parameter estimates fall below the cut-off values or are among the ones with highest factor loadings have been released.

Implications

The results of this study demonstrated rather a weak relationship between the dimensional structure of the PISA mathematics items and PISA's mathematical literacy

framework. In terms of score reporting, this finding suggests that the common practice of reporting separate dimension scores (i.e., a score for Quantity, another score for Change and Relationship, etc.) does not have strong psychometric support.

Loveinger (1957) suggests that the items which best conform to the structural model should be used in the actual assessment. The conclusions made in this study, then, have the following implications for PISA in the light of this recommendation. First, since they have the least conformity to the structural models produced out of the assessment framework, the PISA mathematics items that have low factor loadings should be re-visited for their content, language, scoring, and other characteristics such as item format. The dimensionality of revised set of items should be re-assessed ordinarily. If the same items still psychometrically perform poor, they should be discarded to increase the correspondence between the assessment framework and the actual test items, in turn increasing the construct validity of the whole assessment and, thus, the accuracy of the interpretations to be made.

In addition, this study, in parallel with the literature, also demonstrated that the dimensionality analysis is important in the sense that having one of most the robust frameworks for mathematical literacy and strong content validation (expert opinions) such as PISA's does not ensure a perfect execution of the intended assessment plan. So, the psychometricians need to be more cautious about the item selection and construct validation as well as other gathering evidence for other types of validity (Loveinger, 1957).

Psychometricians should also be cautious about what Loveinger (1957) calls “the problem of homogeneity,” when developing an assessment test. If the goal of the assessment is to predict a single construct such as mathematical literacy, which is defined in the literature in a multidimensional manner, then the data should not conform to unidimensional model. Likewise, if the data conforms to a unidimensional model then the assessment cannot undertake to predict a construct in a multidimensional manner.

This might have implications for the mathematics education community. Mathematics educators might need to reconsider and conceptualize the mathematical literacy as a construct. Learning, instruction, and assessment are intertwined. If PISA's current design does not allow assessing mathematical literacy in a multidimensional manner, then one alternative would be to consider mathematical literacy in a unidimensional manner. However, the fact that mathematical literacy encompasses several knowledge and skills as clearly (and in a very robust way) demonstrated in the literature from the cognitive sciences and other fields of study disagrees treating mathematical literacy as a single-dimensional construct in learning and instruction. Therefore, it might be the Rasch models preventing from assessing mathematical literacy in a multidimensional and valid way. If this is the case, then that other psychometric techniques or assessment designs are required.

Policy makers, too, need to be aware of the issues related to test construction and validation. The dimensionality analysis is one of the important components to be addressed in test construction and use. Policy makers need to consider these validity

issues when making judgments about the students' mathematical literacy and when making educational policies based on them.

Lastly, although it is clear that the dimensions are useful for organizing mathematics domain and mathematical literacy assessments and therefore have utility independent of the dimensional structure of the assessment, a weaker support for multidimensionality might imply that dimensions of mathematical literacy are so intertwined that teachers, schools, and curriculum materials should emphasize a holistic view of mathematical literacy. That is, rather than focusing on and teaching one dimension (or aspect) of mathematical literacy at a time, it should be handled with strong connections to other dimensions (or aspects). For example, students are typically taught specific mathematics concepts such as average, ratio etc. as stand-alone skills usually out of context. However, making connections with other quantitative reasoning skills and utilizing many skills to solve a real life problem might become more critical in the contemporary education.

This study analyzed PISA's framework on the mathematical literacy (OECD, 2009a), which is a well-developed and comprehensive framework. This framework integrates several research-based perspectives on mathematics literacy, which provide a detailed description of the multidimensional nature of this construct. For example, the literature defines the mathematical literacy in terms of proficiencies or competencies (e.g., Kilpatrick, Swafford, & Findell, 2001) and knowledge and skills (e.g., Ojose, 2011), or according to its connection to real life situations (e.g., Steen, 2001) and content-wise decomposition of mathematical literacy (e.g., Steen, 2001). PISA's framework

seems to bring all of these views together, therefore providing a broader and more detailed view on mathematical literacy. However, some views on mathematical literacy are left out from this framework such as critical education (e.g., Freire, 1970; Ojose, 2011), social and democratic aspects (Frankenstein, 1992; Moses & Cobb, 2001) as well as cultural identity perspective (Jablonka, 2003).

This study provided a dimensionality analysis using data from one of the most widely-recognized assessment designs in the world, PISA, which has a well-articulated and comprehensive framework on mathematical literacy and a robust psychometric design. Yet, the multidimensional nature of mathematical literacy cannot be reflected in the assessment instrument well enough. Psychometric methods currently being used for most large-scale assessments (Rasch models) may be too limiting to provide evidence for the types of constructs the field of mathematics education is interested in and in need of assessing. Therefore, other psychometric methods that can be better coordinated with multidimensional constructs could provide more valid assessments. The field of mathematics education is in high need of new assessment designs that would bring in other views on mathematics literacy -beyond those addressed in PISA, together with more current psychometric models that allow for assessment of multidimensional constructs, and therefore providing a more encompassing perspective and more valid assessments, especially those that are implemented at a large-scale and that have such high stakes decisions based on these results.

Appendix A: 2003 Mathematics Item Descriptions

ITEM	2006	2009	Content Dimension	Process Dimension	Context Dimension
M01	Yes	Yes	SS	REP	PER
M02	Yes	Yes	SS	CON	EDOP
M03	-	-	CR	REP	PER
M04	-	-	CR	CON	PER
M05	-	-	SS	REP	EDOP
M06	-	-	SS	CON	EDOP
M07	-	-	SS	CON	EDOP
M08	-	-	SS	CON	EDOP
M09	-	-	SS	REP	EDOP
M10	-	-	CR	REP	SCI
M11	-	-	CR	REP	SCI
M12	-	-	CR	CON	SCI
M13	Yes	Yes	CR	CON	SCI
M14	Yes	Yes	CR	CON	SCI
M15	Yes	Yes	CR	REF	SCI
M16	Yes	Yes	CR	CON	SCI
M17	-	-	UN	CON	PUB
M18	Yes	Yes	CR	CON	EDOP
M19	-	-	SS	CON	EDOP
M20	Yes	Yes	SS	CON	EDOP
M21	Yes	-	CR	REP	PUB
M22	Yes	-	CR	CON	PUB
M23	Yes	-	CR	REF	PUB
M24	Yes	-	SS	CON	PUB
M25	-	-	CR	CON	PER
M26	-	-	CR	REF	PER
M27	Yes	Yes	SS	CON	PUB
M28	Yes	Yes	SS	CON	PUB
M29	-	-	SS	REF	PUB
M30	Yes	Yes	UN	CON	PUB
M31	Yes	Yes	QT	REP	PUB
M32	Yes	Yes	UN	CON	PUB
M33	-	-	QT	REP	PUB
M34	-	-	QT	REP	PUB
M35	-	-	QT	REF	PUB
M36	Yes	Yes	UN	REF	PER

ITEM	2006	2009	Content Dimension	Process Dimension	Context Dimension
M37	Yes	-	UN	REP	EDOP
M38	Yes	-	UN	REF	EDOP
M39	Yes	-	UN	REF	EDOP
M40	Yes	Yes	UN	REP	PER
M41	-	-	UN	REP	PUB
M42	-	-	UN	CON	PUB
M43	Yes	Yes	QT	REF	PUB
M44	Yes	Yes	CR	REP	SCI
M45	Yes	Yes	CR	REF	SCI
M46	Yes	Yes	SS	REP	PUB
M47	Yes	Yes	SS	REF	SCI
M48	Yes	Yes	SS	CON	PUB
M49	-	-	UN	REP	PER
M50	-	-	UN	REP	EDOP
M51	Yes	Yes	QT	REP	EDOP
M52	-	-	QT	CON	EDOP
M53	Yes	Yes	QT	CON	PUB
M54	Yes	Yes	QT	CON	PUB
M55	-	-	UN	REF	SCI
M56	-	-	UN	REF	SCI
M57	-	-	QT	CON	EDOP
M58	-	-	UN	CON	EDOP
M59	-	-	QT	REP	PER
M60	-	-	QT	REP	PER
M61	-	-	QT	CON	PER
M62	-	-	SS	REP	EDOP
M63	-	-	SS	CON	PER
M64	Yes	Yes	QT	REF	PUB
M65	Yes	Yes	QT	REP	PUB
M66	Yes	Yes	UN	REF	PUB
M67	Yes	Yes	CR	REF	SCI
M68	Yes	-	SS	REF	PER
M69	Yes	Yes	QT	CON	SCI
M70	Yes	Yes	QT	CON	SCI
M71	-	-	UN	CON	PUB
M72	-	-	CR	REP	PUB
M73	-	-	CR	REF	PUB
M74	Yes	-	UN	CON	PUB

ITEM	2006	2009	Content Dimension	Process Dimension	Context Dimension
M75	Yes	Yes	QT	REP	PER
M76	Yes	Yes	UN	CON	EDOP
M77	-	-	QT	REP	EDOP
M78	Yes	-	QT	CON	PER
M79	Yes	-	QT	CON	PER
M80	Yes	-	CR	REF	PER
M81	Yes	Yes	CR	REP	SCI
M82	Yes	Yes	UN	CON	SCI
M83	Yes	Yes	QT	CON	SCI
M84	Yes	-	SS	CON	PER

Released items are bold-faced.

Items that were included in the other cycles are labeled as "Yes" for the corresponding cycle.

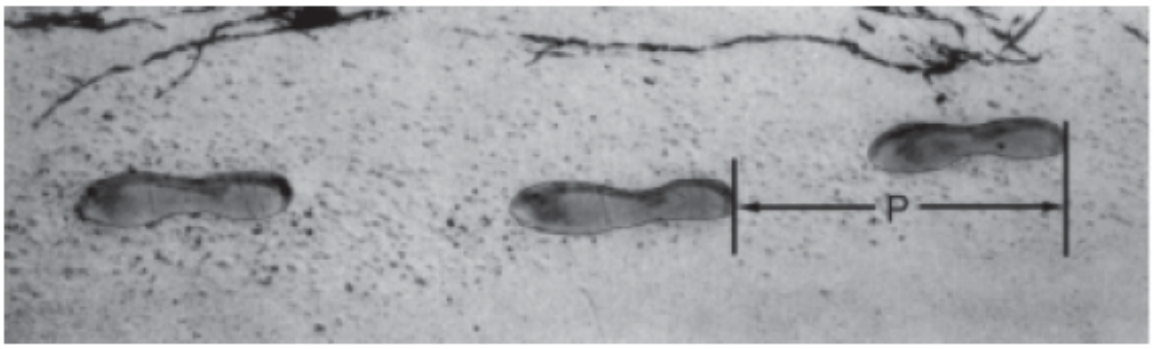
Content Dimensions: QT-quantity, SS-space and shape, CR-change and relationship, UN-uncertainty.

Process Dimensions: REP-reproduction, CON-connections, REF-reflection.

Context Dimensions: PER-personal, EDOP-educational/occupational, PUB-public, SCI-scientific.

Appendix B: Sample Released Items

M03



The picture shows the footprints of a man walking. The pacelength P is the distance between the rear of two consecutive footprints. For men, the formula, $n/P = 140$, gives an approximate relationship between n and P where,

n = number of steps per minute, and

P = pacelength in metres.

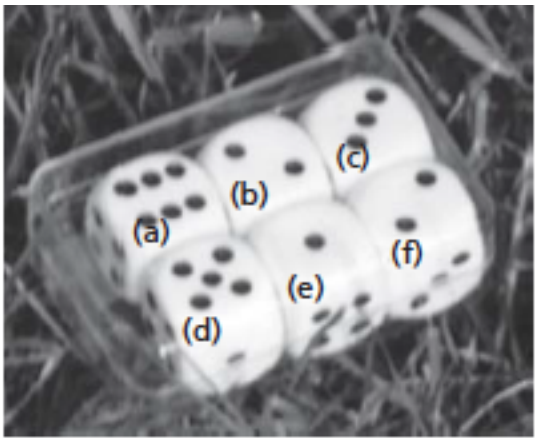
If the formula applies to Heiko's walking and Heiko takes 70 steps per minute, what is Heiko's pacelength? Show your work.

Content: Change and Relationship
Process: Reproduction
Context: Personal

M09

In this photograph you see six dice, labelled (a) to (f). For all dice there is a rule:
The total number of dots on two opposite faces of each die is always seven.

Write in each box the number of dots on the bottom face of the dice corresponding to the photograph.



(a) (b) (c)

(d) (e) (f)

Content: Space and Shape
Process: Reproduction
Context: Educational/Occupational

M25

Mark (from Sydney, Australia) and Hans (from Berlin, Germany) often communicate with each other using "chat" on the Internet. They have to log on to the Internet at the same time to be able to chat.

To find a suitable time to chat, Mark looked up a chart of world times and found the following:

At 7:00 PM in Sydney, what time is it in Berlin?
Answer: ……

Content: Change and Relationship
Process: Connections
Context: Personal

M33

Mei-Ling from Singapore was preparing to go to South Africa for 3 months as an exchange student. She needed to change some Singapore dollars (SGD) into South African rand (ZAR).

During these 3 months the exchange rate had changed from 4.2 to 4.0 ZAR per SGD. Was it in Mei-Ling's favour that the exchange rate now was 4.0 ZAR instead of 4.2 ZAR, when she changed her South African rand back to Singapore dollars? Give an explanation to support your answer.

Content: Quantity
Process: Reflection
Context: Public

Appendix C: Factor Loadings for PISA Mathematics Items

2003 ITEMS	Model1: 1F-GML	Model 2: 1L-CNT		Model 3: 1L-PRO		Model 4: 1L-CXT		Model5: 2L-CNT	Model6: 2L-PRO	Model7: 2L-CXT
M01	0.512	0.537	(SS)	0.520	(REP)	0.511	(PER)	0.537	0.516	0.511
M02	0.586	0.613	(SS)	0.588	(CON)	0.595	(EDOP)	0.613	0.590	0.595
M03	0.677	0.687	(CR)	0.687	(REP)	0.675	(PER)	0.686	0.684	0.675
M04	0.775	0.785	(CR)	0.779	(CON)	0.773	(PER)	0.784	0.780	0.773
M05	0.609	0.638	(SS)	0.619	(REP)	0.617	(EDOP)	0.637	0.616	0.617
M06	0.589	0.617	(SS)	0.592	(CON)	0.597	(EDOP)	0.617	0.593	0.597
M07	0.780	0.812	(SS)	0.783	(CON)	0.790	(EDOP)	0.813	0.785	0.791
M08	0.695	0.727	(SS)	0.698	(CON)	0.704	(EDOP)	0.727	0.699	0.704
M09	0.733	0.765	(SS)	0.744	(REP)	0.745	(EDOP)	0.765	0.741	0.745
M10	0.703	0.716	(CR)	0.713	(REP)	0.722	(SCI)	0.716	0.710	0.722
M11	0.543	0.552	(CR)	0.550	(REP)	0.557	(SCI)	0.552	0.548	0.556
M12	0.619	0.631	(CR)	0.622	(CON)	0.635	(SCI)	0.631	0.623	0.635
M13	0.506	0.513	(CR)	0.507	(CON)	0.517	(SCI)	0.512	0.508	0.517
M14	0.673	0.681	(CR)	0.675	(CON)	0.687	(SCI)	0.681	0.676	0.687
M15	0.784	0.796	(CR)	0.786	(REF)	0.802	(SCI)	0.795	0.791	0.802
M16	0.532	0.540	(CR)	0.534	(CON)	0.544	(SCI)	0.539	0.535	0.544
M17	0.614	0.618	(UN)	0.616	(CON)	0.615	(PUB)	0.617	0.617	0.613
M18	0.624	0.634	(CR)	0.627	(CON)	0.633	(EDOP)	0.634	0.627	0.633
M19	0.541	0.573	(SS)	0.543	(CON)	0.550	(EDOP)	0.573	0.544	0.550
M20	0.442	0.467	(SS)	0.444	(CON)	0.450	(EDOP)	0.467	0.444	0.449
M21	0.608	0.617	(CR)	0.616	(REP)	0.611	(PUB)	0.617	0.614	0.611
M22	0.579	0.588	(CR)	0.581	(CON)	0.582	(PUB)	0.588	0.582	0.583
M23	0.711	0.721	(CR)	0.712	(REF)	0.715	(PUB)	0.721	0.718	0.715
M24	0.444	0.468	(SS)	0.446	(CON)	0.446	(PUB)	0.467	0.447	0.446
M25	0.507	0.516	(CR)	0.510	(CON)	0.506	(PER)	0.516	0.510	0.506
M26	0.608	0.617	(CR)	0.609	(REF)	0.607	(PER)	0.617	0.613	0.606
M27	0.781	0.811	(SS)	0.785	(CON)	0.783	(PUB)	0.811	0.785	0.783
M28	0.897	0.930	(SS)	0.900	(CON)	0.899	(PUB)	0.930	0.901	0.899
M29	0.661	0.688	(SS)	0.663	(REF)	0.663	(PUB)	0.688	0.667	0.663
M30	0.520	0.523	(UN)	0.522	(CON)	0.522	(PUB)	0.523	0.523	0.522
M31	0.651	0.658	(QT)	0.660	(REP)	0.652	(PUB)	0.658	0.657	0.652
M32	0.531	0.534	(UN)	0.534	(CON)	0.532	(PUB)	0.534	0.534	0.532
M33	0.741	0.749	(QT)	0.750	(REP)	0.743	(PUB)	0.748	0.748	0.743
M34	0.788	0.799	(QT)	0.799	(REP)	0.790	(PUB)	0.798	0.796	0.790
M35	0.630	0.638	(QT)	0.631	(REF)	0.631	(PUB)	0.638	0.635	0.631
M36	0.468	0.469	(UN)	0.469	(REF)	0.467	(PER)	0.470	0.472	0.467

2003 ITEMS	Model1: 1F-GML	Model 2: 1L-CNT		Model 3: 1L-PRO		Model 4: 1L-CXT		Model5: 2L-CNT	Model6: 2L-PRO	Model7: 2L-CXT
M37	0.781	0.788	(UN)	0.794	(REP)	0.791	(EDOP)	0.787	0.790	0.791
M38	0.536	0.540	(UN)	0.537	(REF)	0.542	(EDOP)	0.539	0.540	0.542
M39	0.597	0.602	(UN)	0.598	(REF)	0.604	(EDOP)	0.601	0.602	0.604
M40	0.441	0.444	(UN)	0.447	(REP)	0.440	(PER)	0.444	0.446	0.440
M41	0.452	0.455	(UN)	0.460	(REP)	0.455	(PUB)	0.455	0.458	0.455
M42	0.653	0.655	(UN)	0.656	(CON)	0.656	(PUB)	0.656	0.657	0.656
M43	0.714	0.720	(QT)	0.715	(REF)	0.718	(PUB)	0.720	0.720	0.718
M44	0.680	0.691	(CR)	0.689	(REP)	0.697	(SCI)	0.692	0.686	0.697
M45	0.801	0.814	(CR)	0.803	(REF)	0.821	(SCI)	0.815	0.808	0.822
M46	0.665	0.703	(SS)	0.677	(REP)	0.668	(PUB)	0.702	0.673	0.668
M47	0.562	0.591	(SS)	0.563	(REF)	0.576	(SCI)	0.591	0.567	0.576
M48	0.823	0.865	(SS)	0.827	(CON)	0.826	(PUB)	0.866	0.828	0.826
M49	0.685	0.690	(UN)	0.697	(REP)	0.683	(PER)	0.689	0.694	0.683
M50	0.519	0.520	(UN)	0.528	(REP)	0.528	(EDOP)	0.521	0.526	0.528
M51	0.504	0.508	(QT)	0.511	(REP)	0.511	(EDOP)	0.508	0.509	0.512
M52	0.726	0.733	(QT)	0.729	(CON)	0.741	(EDOP)	0.732	0.730	0.740
M53	0.660	0.665	(QT)	0.662	(CON)	0.663	(PUB)	0.665	0.663	0.663
M54	0.528	0.532	(QT)	0.530	(CON)	0.531	(PUB)	0.532	0.531	0.531
M55	0.560	0.562	(UN)	0.561	(REF)	0.574	(SCI)	0.562	0.565	0.574
M56	0.527	0.528	(UN)	0.527	(REF)	0.538	(SCI)	0.529	0.531	0.538
M57	0.450	0.453	(QT)	0.452	(CON)	0.457	(EDOP)	0.454	0.453	0.457
M58	0.596	0.598	(UN)	0.598	(CON)	0.606	(EDOP)	0.598	0.599	0.606
M59	0.628	0.635	(QT)	0.636	(REP)	0.626	(PER)	0.635	0.634	0.626
M60	0.687	0.695	(QT)	0.697	(REP)	0.685	(PER)	0.695	0.694	0.685
M61	0.545	0.552	(QT)	0.548	(CON)	0.544	(PER)	0.552	0.548	0.544
M62	0.579	0.607	(SS)	0.587	(REP)	0.588	(EDOP)	0.607	0.585	0.588
M63	0.793	0.831	(SS)	0.798	(CON)	0.791	(PER)	0.832	0.799	0.791
M64	0.559	0.567	(QT)	0.561	(REF)	0.563	(PUB)	0.567	0.564	0.563
M65	0.498	0.502	(QT)	0.506	(REP)	0.499	(PUB)	0.502	0.504	0.499
M66	0.480	0.482	(UN)	0.481	(REF)	0.481	(PUB)	0.482	0.483	0.481
M67	0.597	0.606	(CR)	0.598	(REF)	0.615	(SCI)	0.607	0.601	0.614
M68	0.426	0.446	(SS)	0.427	(REF)	0.424	(PER)	0.445	0.430	0.424
M69	0.523	0.528	(QT)	0.525	(CON)	0.536	(SCI)	0.528	0.525	0.536
M70	0.531	0.536	(QT)	0.532	(CON)	0.544	(SCI)	0.536	0.533	0.544
M71	0.547	0.550	(UN)	0.549	(CON)	0.549	(PUB)	0.550	0.550	0.549
M72	0.697	0.708	(CR)	0.708	(REP)	0.701	(PUB)	0.709	0.705	0.701
M73	0.784	0.798	(CR)	0.785	(REF)	0.788	(PUB)	0.799	0.790	0.789
M74	0.506	0.508	(UN)	0.509	(CON)	0.507	(PUB)	0.509	0.509	0.507

2003 ITEMS	Model1: 1F-GML	Model 2: 1L-CNT		Model 3: 1L-PRO		Model 4: 1L-CXT		Model5: 2L-CNT	Model6: 2L-PRO	Model7: 2L-CXT
M75	0.383	0.388	(QT)	0.389	(REP)	0.382	(PER)	0.388	0.387	0.382
M76	0.706	0.709	(UN)	0.708	(CON)	0.718	(EDOP)	0.709	0.709	0.718
M77	0.636	0.641	(QT)	0.645	(REP)	0.646	(EDOP)	0.642	0.643	0.646
M78	0.545	0.550	(QT)	0.547	(CON)	0.543	(PER)	0.551	0.548	0.543
M79	0.585	0.592	(QT)	0.587	(CON)	0.583	(PER)	0.592	0.588	0.583
M80	0.806	0.821	(CR)	0.807	(REF)	0.804	(PER)	0.822	0.813	0.804
M81	0.564	0.571	(CR)	0.575	(REP)	0.578	(SCI)	0.571	0.572	0.578
M82	0.334	0.335	(UN)	0.335	(CON)	0.342	(SCI)	0.335	0.335	0.342
M83	0.344	0.347	(QT)	0.345	(CON)	0.352	(SCI)	0.347	0.345	0.352
M84	0.483	0.508	(SS)	0.485	(CON)	0.482	(PER)	0.508	0.486	0.482

Released items are bold-faced.

2006 ITEMS	Model1: 1F-GML	Model 2: 1L-CNT		Model 3: 1L-PRO		Model 4: 1L-CXT		Model5: 2L-CNT	Model6: 2L-PRO	Model7: 2L-CXT
M01	0.496	0.529	(SS)	0.511	(REP)	0.507	(PER)	0.529	0.504	0.509
M02	0.586	0.627	(SS)	0.586	(CON)	0.609	(EDOP)	0.626	0.591	0.609
M13	0.553	0.561	(CR)	0.553	(CON)	0.564	(SCI)	0.560	0.556	0.563
M14	0.693	0.702	(CR)	0.693	(CON)	0.709	(SCI)	0.701	0.697	0.709
M15	0.795	0.806	(CR)	0.809	(REF)	0.815	(SCI)	0.805	0.801	0.815
M16	0.564	0.571	(CR)	0.564	(CON)	0.577	(SCI)	0.571	0.568	0.577
M18	0.616	0.627	(CR)	0.617	(CON)	0.646	(EDOP)	0.626	0.618	0.648
M20	0.530	0.562	(SS)	0.531	(CON)	0.556	(EDOP)	0.561	0.534	0.556
M21	0.685	0.693	(CR)	0.705	(REP)	0.695	(PUB)	0.693	0.697	0.698
M22	0.611	0.617	(CR)	0.611	(CON)	0.624	(PUB)	0.617	0.616	0.625
M23	0.718	0.726	(CR)	0.731	(REF)	0.734	(PUB)	0.726	0.724	0.734
M24	0.373	0.396	(SS)	0.373	(CON)	0.378	(PUB)	0.396	0.374	0.378
M27	0.766	0.802	(SS)	0.766	(CON)	0.776	(PUB)	0.803	0.768	0.775
M28	0.864	0.917	(SS)	0.864	(CON)	0.875	(PUB)	0.917	0.867	0.875
M30	0.594	0.623	(UN)	0.595	(CON)	0.607	(PUB)	0.622	0.598	0.607
M31	0.649	0.675	(QT)	0.672	(REP)	0.664	(PUB)	0.675	0.662	0.665
M32	0.540	0.565	(UN)	0.541	(CON)	0.552	(PUB)	0.565	0.544	0.552
M36	0.542	0.567	(UN)	0.554	(REF)	0.553	(PER)	0.566	0.547	0.555
M37	0.842	0.876	(UN)	0.877	(REP)	0.880	(EDOP)	0.877	0.860	0.878
M38	0.510	0.528	(UN)	0.519	(REF)	0.529	(EDOP)	0.528	0.514	0.528
M39	0.487	0.505	(UN)	0.496	(REF)	0.506	(EDOP)	0.505	0.491	0.506
M40	0.426	0.450	(UN)	0.440	(REP)	0.436	(PER)	0.449	0.436	0.437
M43	0.734	0.765	(QT)	0.747	(REF)	0.752	(PUB)	0.765	0.740	0.752
M44	0.749	0.758	(CR)	0.772	(REP)	0.765	(SCI)	0.758	0.762	0.766
M45	0.874	0.882	(CR)	0.894	(REF)	0.888	(SCI)	0.882	0.882	0.889
M46	0.614	0.651	(SS)	0.629	(REP)	0.627	(PUB)	0.650	0.624	0.627
M47	0.688	0.737	(SS)	0.700	(REF)	0.700	(SCI)	0.737	0.694	0.700
M48	0.797	0.848	(SS)	0.798	(CON)	0.816	(PUB)	0.848	0.803	0.815
M51	0.471	0.488	(QT)	0.487	(REP)	0.489	(EDOP)	0.488	0.480	0.489
M53	0.729	0.759	(QT)	0.729	(CON)	0.740	(PUB)	0.760	0.732	0.740
M54	0.608	0.631	(QT)	0.608	(CON)	0.617	(PUB)	0.631	0.611	0.617
M64	0.600	0.628	(QT)	0.614	(REF)	0.613	(PUB)	0.628	0.606	0.613
M65	0.438	0.453	(QT)	0.452	(REP)	0.445	(PUB)	0.453	0.448	0.444
M66	0.529	0.558	(UN)	0.539	(REF)	0.536	(PUB)	0.557	0.535	0.536
M67	0.687	0.699	(CR)	0.701	(REF)	0.709	(SCI)	0.698	0.695	0.709
M68	0.470	0.504	(SS)	0.478	(REF)	0.478	(PER)	0.503	0.474	0.478
M69	0.523	0.541	(QT)	0.523	(CON)	0.539	(SCI)	0.541	0.524	0.539

2006 ITEMS	Model1: 1F-GML	Model 2: 1L-CNT		Model 3: 1L-PRO		Model 4: 1L-CXT		Model5: 2L-CNT	Model6: 2L-PRO	Model7: 2L-CXT
M70	0.638	0.663	(QT)	0.639	(CON)	0.659	(SCI)	0.663	0.641	0.659
M74	0.565	0.585	(UN)	0.565	(CON)	0.577	(PUB)	0.585	0.569	0.577
M75	0.460	0.480	(QT)	0.471	(REP)	0.468	(PER)	0.479	0.466	0.469
M76	0.735	0.767	(UN)	0.735	(CON)	0.762	(EDOP)	0.767	0.740	0.763
M78	0.637	0.666	(QT)	0.637	(CON)	0.647	(PER)	0.665	0.641	0.647
M79	0.718	0.752	(QT)	0.718	(CON)	0.731	(PER)	0.751	0.722	0.730
M80	0.855	0.862	(CR)	0.872	(REF)	0.872	(PER)	0.863	0.862	0.871
M81	0.730	0.739	(CR)	0.751	(REP)	0.744	(SCI)	0.739	0.743	0.744
M82	0.577	0.604	(UN)	0.578	(CON)	0.590	(SCI)	0.603	0.582	0.590
M83	0.582	0.609	(QT)	0.583	(CON)	0.595	(SCI)	0.609	0.586	0.595
M84	0.484	0.520	(SS)	0.485	(CON)	0.492	(PER)	0.519	0.487	0.492

2009 ITEMS	Model1: 1F-GML	Model 2: 1L-CNT		Model 3: 1L-PRO		Model 4: 1L-CXT		Model5: 2L-CNT	Model6: 2L-PRO	Model7: 2L-CXT
M01	0.488	0.518	(SS)	0.490	(REP)	0.379	(PER)	0.517	0.243	0.142
M02	0.623	0.662	(SS)	0.624	(CON)	0.622	(EDOP)	0.661	0.389	0.388
M13	0.695	0.704	(CR)	0.697	(CON)	0.706	(SCI)	0.705	0.485	0.498
M14	0.676	0.686	(CR)	0.678	(CON)	0.687	(SCI)	0.686	0.459	0.473
M15	0.782	0.798	(CR)	0.790	(REF)	0.798	(SCI)	0.798	0.619	0.638
M16	0.532	0.540	(CR)	0.533	(CON)	0.541	(SCI)	0.540	0.284	0.293
M18	0.633	0.652	(CR)	0.634	(CON)	0.629	(EDOP)	0.651	0.401	0.395
M20	0.491	0.516	(SS)	0.493	(CON)	0.490	(EDOP)	0.516	0.243	0.240
M27	0.844	0.867	(SS)	0.845	(CON)	0.847	(PUB)	0.869	0.714	0.718
M28	0.913	0.945	(SS)	0.915	(CON)	0.920	(PUB)	0.945	0.836	0.847
M30	0.578	0.606	(UN)	0.580	(CON)	0.586	(PUB)	0.606	0.336	0.343
M31	0.655	0.687	(QT)	0.659	(REP)	0.663	(PUB)	0.686	0.438	0.440
M32	0.526	0.549	(UN)	0.527	(CON)	0.532	(PUB)	0.549	0.278	0.283
M36	0.482	0.504	(UN)	0.487	(REF)	0.372	(PER)	0.504	0.235	0.138
M40	0.411	0.438	(UN)	0.413	(REP)	0.313	(PER)	0.436	0.173	0.098
M43	0.709	0.745	(QT)	0.716	(REF)	0.718	(PUB)	0.743	0.508	0.516
M44	0.682	0.696	(CR)	0.686	(REP)	0.696	(SCI)	0.697	0.474	0.484
M45	0.743	0.756	(CR)	0.751	(REF)	0.756	(SCI)	0.756	0.559	0.571
M46	0.628	0.660	(SS)	0.631	(REP)	0.635	(PUB)	0.659	0.401	0.403
M47	0.615	0.652	(SS)	0.620	(REF)	0.625	(SCI)	0.651	0.382	0.391
M48	0.737	0.776	(SS)	0.739	(CON)	0.745	(PUB)	0.775	0.545	0.555
M51	0.460	0.482	(QT)	0.463	(REP)	0.460	(EDOP)	0.481	0.216	0.211
M53	0.632	0.662	(QT)	0.633	(CON)	0.637	(PUB)	0.663	0.400	0.406
M54	0.542	0.566	(QT)	0.542	(CON)	0.546	(PUB)	0.566	0.294	0.298
M64	0.531	0.560	(QT)	0.536	(REF)	0.537	(PUB)	0.559	0.285	0.288
M65	0.392	0.406	(QT)	0.395	(REP)	0.394	(PUB)	0.407	0.157	0.155
M66	0.443	0.470	(UN)	0.448	(REF)	0.446	(PUB)	0.468	0.199	0.199
M67	0.648	0.667	(CR)	0.656	(REF)	0.666	(SCI)	0.666	0.425	0.443
M69	0.501	0.524	(QT)	0.501	(CON)	0.513	(SCI)	0.524	0.251	0.263
M70	0.648	0.681	(QT)	0.649	(CON)	0.667	(SCI)	0.683	0.421	0.445
M75	0.351	0.367	(QT)	0.352	(REP)	0.273	(PER)	0.368	0.125	0.075
M76	0.741	0.776	(UN)	0.743	(CON)	0.740	(EDOP)	0.778	0.552	0.547
M81	0.723	0.736	(CR)	0.727	(REP)	0.736	(SCI)	0.737	0.532	0.541
M82	0.528	0.552	(UN)	0.529	(CON)	0.539	(SCI)	0.551	0.280	0.290
M83	0.497	0.524	(QT)	0.498	(CON)	0.505	(SCI)	0.523	0.248	0.255

References

Abedi, J. (1997). *Dimensionality of NAEP subscale scores in mathematics*: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411-423.

Anderson, J. O., Lin, H.-S., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using Large-Scale Assessment Datasets For Research In Science And Mathematics Education: Programme For International Student Assessment (Pisa). *International Journal of Science and Mathematics Education, 5*, 591-614.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin, 88*(3), 588-606.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261-280.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. *IERI monograph series: Issues and methodologies in large-scale assessments, 1*, 51-69.

Burg, S. S. (2007). *An Investigation of Dimensionality Across Grade Levels and Effects on Vertical Linking for Elementary Grade Mathematics Achievement Tests.* Unpublished Dissertation, University of North Carolina Chapel Hill, School of Education.

Byrne, B. M. (2011). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* New York: Taylor & Francis.

Carmona, G., Krause, G., Monroy, M., Lima, C., Ávila, A., & Ekmekci, A. (2011). *A Longitudinal Study to Investigate Changes in Students' Mathematics Scores in Texas.* Paper presented at the 2011 Annual Meeting of American Educational Research Association (AERA), New Orleans, LA.

de Champlain, A. F. (1992). *Assessing test dimensionality using two approximate chi-square statistics.* Unpublished Dissertation, University of Ottawa, Canada.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9*(2), 233-255.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-Scale Assessments that Support Learning: What will it take? *Theory Into Practice, 42*(1), 75-83.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507): American Council on Education Washington, DC.

Deng, N., Wells, C., & Hambleton, R. (2008). *A Confirmatory Factor Analytic Study Examining the Dimensionality of Educational Achievement Tests.* Paper presented at the 39th Annual Conference of Northeastern Educational Research Association (NERA), Rocky Hill, Connecticut.

Ekmekci, A., & Carmona, G. (2012). Mathematical Literacy Assessment Design: A Multivariate Analysis of PISA 2003 Mathematics Items in the U.S. In Van Zoest, L. R., Lo, J.-J., & Kratky, J. L. (Eds.), *Proceedings of the 34th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (p. 390). Kalamazoo, MI: Western Michigan University.

Frankenstein, M. (1992). Critical mathematics education: An application of Paulo Freire's epistemology. In K. Weiler & C. Mitchell (Eds.), *What schools can do: Critical pedagogy and practice*. Albany, NY: State University of New York Press.

Freire, P. (1970). *Pedagogy of the Oppressed*. New York: Continuum.

Griffo, V. B. (2011). *Examining NAEP: The Effect of Item Format on Struggling 4th Graders' Reading Comprehension.* Unpublished Dissertation, University of California, Berkeley.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*(3), 287-302.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

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

Jablonka, E. (2003). Mathematical Literacy. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Second International Handbook of Mathematics Education* (pp. 77-104). Berlin: Springer.

Jablonka, E., & Gellert, U. (2007). Mathematisation – Demathematisation. In U. Gellert & E. Jablonka (Eds.), *Mathematisation and demathematisation: Social, philosophical and educational ramifications* (pp. 1-18). Rotterdam, The Netherlands: Sense Publishers.

Joreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality & Quantity, 24*(4), 387-404.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin; Psychological Bulletin, 112*(3), 527.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding It Up: Helping Children Learn Mathematics*. Washington, DC.: National Academy Press.

Kilpatrick, J. (2001). Understanding Mathematical Literacy: The Contribution of Research. *Educational Studies in Mathematics, 47*(1), 101-116.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling* (3 ed.). New York: Guilford Press.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing, 9*(2), 122-133.

Lesh, R., & Carmona, G. (2003). Piagetian Conceptual Systems and Models for Mathematizing Everyday Experiences. In R. Lesh & H. M. Doerr (Eds.), *Beyond Constructivism, Models and Modeling Perspectives on Mathematics Problem Solving, Learning, and Teaching* (pp. 71-96). Manweh, NJ: Lawrence Erlbaum Associates.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Loevinger, J. (1957). Objective Tests As Instruments of Psychological Theory: Monograph Supplement 9. *Psychological reports, 3*(3), 635-694.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education, Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational and Behavioral Statistics, 11*(1), 3-31.

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

Moses, R. P., & Cobb, C. E. (2001). *Radical Equations: Civil Rights from Mississippi to the Algebra Project*. Boston, MA: Beacon.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Amsterdam, the Netherlands: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-243). Newbury Park, CA: Sage.

Muthén, L.K. and Muthén, B.O. (1998-2011). *Mplus, Version 6.12 (Computer Software).* Los Angeles, CA: Muthén & Muthén

Muthén, L.K. and Muthén, B.O. (2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

National Research Council. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: The National Academies Press.

Ojose, B. (2011). Mathematics Literacy: Are We Able To Put The Mathematics We Learn Into Everyday Use? *Journal of Mathematics Education, 4*(1), 89-100.

Organisation for Economic Co-operation and Development (OECD). (2003). *The PISA 2003 Assessment Framework - Mathematics, Reading, Science, and Problem Solving Knowledge and Skills*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2009a). *PISA 2009 Assessment Framework - Key Competencies in Reading, Mathematics and Science*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2009b). *Learning Mathematics for Life: A Perspective from PISA*. Available from http://www.oecd.org//edu/preschoolandschool/programmeforinternationalstudentassessmentpisa/learningmathematicsforlifeaperspectivefrompisa.htm

Roussos, L., Stout, W., & Marden, J. L. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1-30.

Rubio, D. M., Berg-Weger, M., & Tebb, S. S. (2001). Using Structural Equation Modeling to Test for Multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(4), 613-626.

Schwab, C. J. (2007). *What Can We Learn from PISA? Investigating PISA's Approach to Scientific Literacy.* Unpublished Dissertation, University of California, Berkeley, U.S.

Skovsmose, O. (1994). *Towards a Philosophy of Critical Mathematics Education*. Dordrecht, The Netherlands: Kluwer.

Somerville, J. T. (2012). *Detection of Differential Item Functioning in the Generalized Full-Information Item Bifactor Analysis Model.* Unpublished Dissertation, University of California, Los Angeles, United States -- California.

Stacey, K. (2010). Mathematical and Scientific Literacy Around The World. *Journal of Science and Mathematics Education in Southeast Asia, 33*(1), 1-16.

Steen, L. A. (2001). *Mathematics and democracy: The case for quantitative literacy*. Princeton, NJ: National Council on Education and the Disciplines.

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*(2), 193-214.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589-617.

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

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

Tate, R. (2002). Test Dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*. Mahnah, New Jersey, London: L. Erlbaum.

Tate, R. (2003). A Comparison of Selected Empirical Methods for Assessing the Structure of Responses to Test Items. *Applied Psychological Measurement, 23*(3), 159-203.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. UK: Wiley.

Wei, H. (2008). *Multidimensionality in the NAEP Science Assessment: Substantive perspectives, psychometric models, and task design.* Unpublished Dissertation, University of Maryland, College Park, United States.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Yildirim, H. H., & Berberoglu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing, 9*(2), 108-121.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* Unpublished Dissertation, University of California, Los Angeles, United States -- California.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 34, 213-249.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24*(4), 293-308.