

University of Denver

Digital Commons @ DU

---

Electronic Theses and Dissertations

Graduate Studies

---

1-1-2016

## A Comparison of Latent Class Analysis and the Mixture Rasch Model: A Cross-Cultural Comparison of 8th Grade Mathematics Achievement in the Fourth International Mathematics and Science Study (TIMSS-2011)

Turker Toker  
*University of Denver*

Follow this and additional works at: <https://digitalcommons.du.edu/etd>



Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Elementary Education Commons](#), and the [Statistics and Probability Commons](#)

---

### Recommended Citation

Toker, Turker, "A Comparison of Latent Class Analysis and the Mixture Rasch Model: A Cross-Cultural Comparison of 8th Grade Mathematics Achievement in the Fourth International Mathematics and Science Study (TIMSS-2011)" (2016). *Electronic Theses and Dissertations*. 1172.  
<https://digitalcommons.du.edu/etd/1172>

This Dissertation is brought to you for free and open access by the Graduate Studies at Digital Commons @ DU. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ DU. For more information, please contact [jennifer.cox@du.edu](mailto:jennifer.cox@du.edu), [dig-commons@du.edu](mailto:dig-commons@du.edu).

---

# A Comparison of Latent Class Analysis and the Mixture Rasch Model: A Cross-Cultural Comparison of 8th Grade Mathematics Achievement in the Fourth International Mathematics and Science Study (TIMSS-2011)

## Abstract

This study provides a comparison of the results of latent class analysis (LCA) and mixture Rasch model (MRM) analysis using data from the Trends in International Mathematics and Science Study - 2011 (TIMSS-2011) with a focus on the 8th-grade mathematics section. The research study focuses on the comparison of LCA with Mplus version 7.31 and MRM with WinMira 2011 to determine if results obtained differ when the assumed psychometric model differs. Also, a log-linear analysis was conducted to understand the interactions between latent classes identified by LCA and MRM. The data set used in the study was from four diverse countries (Turkey, USA, Finland, and Singapore) participating in TIMSS-2011. There are instructional differences and historical performance differences for each country, which was why they were selected. Analyses yielded class results associated mostly with nation of the participants, which was, in turn, associated with performance level.

Although the two approaches and the outcomes in terms of class designations overlapped, assumptions about the nature of the data and the information derived from each analysis differed. The literature review summarized the theory and application of latent class analysis and the mixture Rasch model in identifying latent classes in the social sciences. The results suggest that TIMSS-2011 8th-grade mathematics data yield different subgroups based on ability levels of students.

The findings of this paper do not reveal unequivocally whether a model based on primarily qualitative differences (LCA), that is, different strategies, instructional differences, curriculum etc. or a model including additional factors of quantitative differences within strategies (MRM) should be used with this particular dataset. Both of the tests provided similar results with more or less similar interpretations. Both techniques fit the data similarly, a result found in prior research. Nonetheless, for tests similar to TIMSS exams, item difficulty parameters can be useful for educational researchers giving potential priority to use of MRM.

## Document Type

Dissertation

## Degree Name

Ph.D.

## Department

Quantitative Research Methods

## First Advisor

Kathy E. Green, Ph.D.

## Second Advisor

Duan Zhang

## Third Advisor

Antonio Olmos

---

**Keywords**

Latent class analysis, Mathematics education, Mixture Rasch model, TIMSS-2011, Turkish educational system, Validity

**Subject Categories**

Educational Assessment, Evaluation, and Research | Elementary Education | Statistics and Probability

**Publication Statement**

Copyright is held by the author. User is responsible for all copyright compliance.

A COMPARISON OF LATENT CLASS ANALYSIS AND THE MIXTURE RASCH  
MODEL: A CROSS-CULTURAL COMPARISON OF 8TH GRADE MATHEMATICS  
ACHIEVEMENT IN THE FOURTH INTERNATIONAL MATHEMATICS AND  
SCIENCE STUDY (TIMSS-2011)

---

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

---

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Turker Toker

August 2016

Advisor: Dr. Kathy E. Green

©Copyright by Turker Toker 2016

All Rights Reserved

Author: Turker Toker

Title: A COMPARISON OF LATENT CLASS ANALYSIS AND THE MIXTURE RASCH MODEL: CROSS-CULTURAL COMPARISON OF 8TH GRADE MATHEMATICS ACHIEVEMENT IN THE FOURTH INTERNATIONAL MATHEMATICS AND SCIENCE STUDY (TIMSS-2011)

Advisor: Dr. Kathy E. Green

Degree Date: August 2016

### **Abstract**

This study provides a comparison of the results of latent class analysis (LCA) and mixture Rasch model (MRM) analysis using data from the Trends in International Mathematics and Science Study – 2011 (TIMSS-2011) with a focus on the 8th-grade mathematics section. The research study focuses on the comparison of LCA with Mplus version 7.31 and MRM with WinMira 2011 to determine if results obtained differ when the assumed psychometric model differs. Also, a log-linear analysis was conducted to understand the interactions between latent classes identified by LCA and MRM. The data set used in the study was from four diverse countries (Turkey, USA, Finland, and Singapore) participating in TIMSS-2011. There are instructional differences and historical performance differences for each country, which was why they were selected. Analyses yielded class results associated mostly with nation of the participants, which was, in turn, associated with performance level.

Although the two approaches and the outcomes in terms of class designations overlapped, assumptions about the nature of the data and the information derived from each analysis differed. The literature review summarized the theory and application of latent class analysis and the mixture Rasch model in identifying latent classes in the

social sciences. The results suggest that TIMSS-2011 8th-grade mathematics data yield different subgroups based on ability levels of students.

The findings of this paper do not reveal unequivocally whether a model based on primarily qualitative differences (LCA), that is, different strategies, instructional differences, curriculum etc. or a model including additional factors of quantitative differences within strategies (MRM) should be used with this particular dataset. Both of the tests provided similar results with more or less similar interpretations. Both techniques fit the data similarly, a result found in prior research. Nonetheless, for tests similar to TIMSS exams, item difficulty parameters can be useful for educational researchers giving potential priority to use of MRM.

Keywords: Latent Class Analysis, Mixture Rasch Model, TIMSS-2011, Mathematics Education, Turkey, USA, Finland, Singapore

## **Acknowledgements**

After several years at DU, today is the day to show my deepest gratitude to those who have supported and helped me so much throughout this period. I give my sincerest gratitude to those who have guided and sustained me through this advanced education opportunity. I could not have completed this journey without the love, understanding, patience, and inspiration of my wonderful wife, Pinar Toker. She gave me two beautiful daughters, Erem Leyl and Hazel Hacer, whose love have always encouraged me.

I am also very grateful to my advisor, Kathy Green, whose encouragement, supervision and support from the beginning of my journey to the concluding level enabled me to develop an understanding of the subject. I also would like to thank my committee members Duan Zhang, Antonio Olmos and Bin Ramke whose comments were always helpful. Furthermore, I also would like to thank my elementary school teacher Aysegul Ozat whose love, guidance and support took me from the suburbs of Adana to a PhD in the US.

Additionally, I am forever indebted to my mother, Emine Kumru, who instilled in me the values of family, faith, and dedication. She was always there for me with her wise counsel. Furthermore, I would like to thank to colors of Turkey from Edirne to Kars, whose hard work have emotionally and financially supported my education. I would never have this opportunity without the support provided by Ministry of National Education officials. Lastly, I would like to thank those who have helped me become who I am today and without them I could have never achieve such a success.



## Table of Contents

Chapter One: Introduction and Review of the Literature .....	1
Purpose of the Study.....	3
Research Questions .....	4
Introduction .....	5
Latent Class Analysis of Item Responses.....	7
An example of Latent Class Analysis .....	17
The Mixture Rasch Model.....	18
An example of the Mixture Rasch Model .....	24
Latent Class Analysis vs. the Mixture Rasch Model.....	25
Log-linear Analysis .....	28
TIMSS-2011 .....	30
Education Systems of Compared Countries .....	35
Turkey .....	35
United States of America .....	38
Finland.....	43
Singapore.....	45
Definition of Terms .....	50
Chapter Two: Method.....	52
Participants.....	52
Instrument .....	53
Procedure .....	57
Analysis.....	58
Research Question One .....	58
Research Question Two.....	59
Research Question Three.....	60
Research Question Four .....	60
Chapter Three: Results.....	62
Research Question One.....	62
Number of Latent Classes .....	62
Booklet One.....	63
Classification Quality .....	64
Definition of Latent Classes .....	66
Booklet Four.....	68
Classification Quality .....	69
Definition of Latent Classes .....	70
Booklet Six .....	72
Classification Quality .....	72
Definition of Latent Classes .....	73
Research Question Two .....	75

Number of Latent Classes .....	75
Booklet One .....	76
Booklet Four .....	78
Booklet Six.....	81
Research Question Three .....	85
Booklet One .....	85
Booklet Four .....	86
Booklet Six.....	87
Research Question Four .....	89
Booklet One .....	89
Associations .....	92
Booklet Four .....	94
Associations.....	95
Booklet Six.....	98
Associations.....	100
Chapter Four: Discussion.....	102
Summary of the Study .....	102
Important Findings.....	104
Research question one .....	104
Research question two .....	105
Research question three .....	108
Research question four .....	110
Implications for Research .....	112
Implications for Turkish Educators .....	114
Limitations of the Study.....	116
Recommendations for Further Research .....	117
References .....	119
Appendix A	
University of Denver Institutional Review Board Application .....	131
Appendix B	
University of Denver Institutional Review Board Approval Letter.....	134
Appendix C	
LCA 2 Class Model Specification for Booklet One .....	136
Appendix D	
LCA model for Booklet Four.....	137
Appendix E	
LCA 3 Class Model Specification for Booklet Four .....	138

Appendix F	
LCA Model for Booklet Six .....	139
Appendix G	
LCA 4 Class Model Specification for Booklet Six.....	140
Appendix H	
Turkish Translation of Acknowledgements (Türkçe Teşekkür Sayfası) .....	141

## List of Tables

Table 1	Gender and Age of TIMSS-2011 Subjects .....	52
Table 2	LCA Model Fit Indices for Booklet One .....	65
Table 3	Final Latent Class Size and Percentage for Booklet One .....	65
Table 4	Average Latent Class Probabilities for Most Likely Latent Class Membership for Booklet One .....	66
Table 5	Three-Class Latent Class Membership for Booklet One .....	67
Table 6	LCA Model Fit Indices for Booklet Four .....	69
Table 7	Final Latent Class Size and Percentage for Booklet Four .....	69
Table 8	Average Latent Class Probabilities for Most Likely Latent Class Membership for Booklet Four .....	70
Table 9	Three-Class Latent Class Membership for Booklet Four .....	71
Table 10	LCA Model Fit Indices for Booklet Six.....	72
Table 11	Final Latent Class Size and Percentage for Booklet Six.....	73
Table 12	Average Latent Class Probabilities for Most Likely Latent Class Membership for Booklet Six .....	73
Table 13	Two-Class Latent Class Membership for Booklet Six.....	74
Table 14	p-values of Model Fit Indices for MRM .....	76
Table 15	Item Fit Assessed by the Q-index for All Classes of Booklet One .....	77
Table 16	Item Parameters of Booklet One by Classes .....	78
Table 17	Item Fit Assessed by The Q-Index for All Classes of Booklet Four.....	80
Table 18	Item Parameters of Booklet Four by Classes .....	81

Table 19	Item Fit Assessed by the Q-Index for All Classes of Booklet Six.....	83
Table 20	Item Parameters of Booklet Six by Classes .....	84
Table 21	Item Parameter Comparisons of Booklet One LCA and MRM by Class .	85
Table 22	LCA and MRM 3 Class Model BIC Fit Indices for Booklet One .....	85
Table 23	LCA and MRM 3 Class Model Class Sizes for Booklet One .....	86
Table 24	Item parameter comparisons of Booklet Four LCA and MRM by Class .	86
Table 25	LCA and MRM 3 vs. 2 Class Model BIC Fit Indices for Booklet Four....	87
Table 26	LCA and MRM 3 vs. 2 Class Model Class Sizes for Booklet Four .....	87
Table 27	Item parameter comparisons of Booklet Six LCA and MRM by Class ...	88
Table 28	LCA and MRM 2 Class Model BIC Fit Indices for Booklet Six.....	88
Table 29	LCA and MRM 2 Class Model Class Sizes for Booklet Six .....	89
Table 30	K-Way and Higher-Order Effects for Booklet One.....	90
Table 31	Partial Associations for Booklet One .....	91
Table 32	Goodness-of-Fit Tests for 2-way Interaction Model for Booklet One .....	91
Table 33	Crosstabulation of Nation vs. LCA Class Membership for Booklet One..	92
Table 34	Crosstabulation of Nation vs. MRM Class Membership for Booklet One	93
Table 35	Crosstabulation of LCA Class Membership vs. MRM Class Membership for Booklet One .....	93
Table 36	K-Way and Higher-Order Effects for Booklet Four.....	94
Table 37	Partial Associations for Booklet Four .....	95
Table 38	Goodness-of-Fit Tests for 2-way Interaction Model for Booklet Four.....	95
Table 39	Crosstabulation of Nation vs. LCA Class Membership for Booklet Four.	96

Table 40	Crosstabulation of Nation vs. Gender for Booklet Four .....	96
Table 41	Crosstabulation of Nation vs.MRM Class Membership for Booklet Four	97
Table 42	Crosstabulation of LCA Class Membership vs. MRM Class Membership for Booklet Four .....	98
Table 43	K-Way and Higher-Order Effects for Booklet Six .....	98
Table 44	Partial Associations for Booklet Six.....	99
Table 45	Goodness-of-Fit Tests for 2-way Interaction Model for Booklet Six .....	99
Table 46	Crosstabulation of Nation vs. LCA Class Membership for Booklet Six .	100
Table 47	Crosstabulation of Nation vs.MRM Class Membership for Booklet Six	101
Table 48	Crosstabulation of LCA Class Membership vs. MRM Class Membership for Booklet Six .....	101

## List of Figures

Figure 1	The Latent Class Model .....	8
Figure 2	A General Latent Variable Modeling Framework .....	10
Figure 3	Turkish Educational System .....	37
Figure 4	US Educational System .....	42
Figure 5	Finnish Educational System .....	44
Figure 6	Singapore Educational System .....	46
Figure 7	Sample Item in Reasoning Cognitive Domain (TIMSS-2011 Report, 2012) ..	54
Figure 8	Sample Item in Applying Cognitive Domain (TIMSS -2011 Report, 2012) ...	55
Figure 9	Sample Item in Knowing Cognitive Domain (TIMSS -2011 Report, 2012) ...	56
Figure 10	LCA Model for Booklet One .....	64
Figure 11	Conditional Probability Profiles of Endorsing “Correct Answer” for 3-Class LCA Model for Booklet One .....	68
Figure 12	Conditional Probability Profiles of Endorsing “Correct Answer” for 3-Class LCA Model for Booklet Four .....	71
Figure 13	Conditional Probability Profiles of Endorsing “Correct Answer” for 2-Class LCA Model For Booklet Six .....	75
Figure 14	Class Specific Item Parameter Profiles for Booklet One .....	78
Figure 15	Class Specific Item Parameter Profiles for Booklet Four .....	81
Figure 16	Class Specific Item Parameter Profiles for Booklet Six .....	84

## **Chapter One**

*(To the Great Nation of Turks)*

### **Introduction and Review of the Literature**

The purpose of this study was to compare of the results of latent class analysis and mixture Rasch model analysis for a major international assessment in mathematics. Latent class analysis and mixture Rasch model analysis are two approaches to identification of latent classes in data. The purpose of the two approaches and likely the outcomes overlap but assumptions about the nature of the data and the information derived from each approach differ. The existence of multiple latent classes in test data speaks to the validity of test scores, particularly with the mixture Rasch model. If multiple latent classes are found in test data, distinct groups of participants exist for whom the construct varies, making cross-country comparisons suspect. The use of an international mathematics assessment for four diverse participating countries (Turkey, USA, Finland, and Singapore) is reviewed in this study, with a brief summary in the discussion of assessment implications for education in Turkey. Four countries with diverse educational systems were selected with the idea that variation in item response patterns might be found based on diversity in instructional systems. Since participant nations attribute great importance to these assessments, it is important to analyze the latent class structure for this test. A test with support for validity of cross-national



comparisons would ideally yield a single latent class and comparable results with both analytic techniques.

The main reason for the selection of the countries was to capture as much variance as possible so that the possible latent classes could be explored by both analyses. The four nations selected are distinctly different from each other and exploring different latent classes in the data is critical for participating countries. International test results are being used widely by researchers as a reference to compare nations to each other or to see the progress of a nation over the time. It is expected that the nations which are participating in the test can be compared using TIMSS results, but the presence of multiple latent classes calls that comparability into question.

International test results are used to guide modifications and development in educational systems for entire nations. Test results are interpreted in comparison with results from other participant nations. It is critical, then, that the test used to assess student performance for a nation has support for validity that makes results comparable cross-nationally. The intent of the present study was to assess whether results of analysis of test data with two current analytic methods, both of which identify latent classes in the data, yield similar results. If distinct latent classes are, indeed, identified, there is a suggestion that the construct measured may not be invariant across those classes. And, if latent class is associated with national origin, the validity of cross-national comparisons is called into question. Latent class analysis (LCA) is a subgroup of structural equation modeling which is used to find categorical groups or subtypes of cases, in the present

case based on responses to test items (McCutcheon, 1987). Mixture Rasch models, which combine Rasch models with latent class analysis, have been used to identify latent classes who might use different problem-solving techniques or who use different skills in response to test items. Both analytic approaches result in identification of latent classes but each approach makes different assumptions about the nature of data and uses different estimation procedures. One main difference between these two analyses is that LCA uses raw response data whereas the mixture Rasch model uses item parameters from Rasch analysis to estimate latent classes within a dataset. Additionally, LCA assumes items are locally independent given class while MRM assumes that items are locally independent given class and ability within the sub-population. Both analytic methods are used primarily in methodology research rather than as a general tool employed by psychometricians (e.g., Dallas & Wilse, 2013) and, to the researcher's knowledge, results of analysis using the two methods have not been directly compared.

This study's primary purpose was to compare results of the two analyses and secondarily to provide evidence addressing the validity of an international mathematics test for making cross-national comparisons.

### **Purpose of the Study**

The purpose of this study was to conduct a comparison of the results of latent class analysis (LCA) and mixture Rasch model (MRM) analysis using data from the Trends in International Mathematics and Science Study – 2011 (TIMSS-2011) with a focus on the 8th-grade mathematics section. The research study focuses on the

comparison of LCA with Mplus version 7.31 and the MRM with WinMira 2011 to determine if results obtained differ when the assumed psychometric model differs. This comparison was conducted in the context of an examination of cross-cultural differences between the four nations' (listed above) educational systems. After a brief introduction, the statistical procedures that are the focus of this paper, LCA and the MRM, are reviewed. Then, literature examining TIMSS-2011 is reviewed, with a focus on four participating countries' educational systems which are briefly described, and a particular emphasis on the author's home country of Turkey.

### **Research Questions**

The following research questions were addressed by this study using LCA and the MRM with the TIMSS-2011 8th-grade mathematics data.

1. Does analysis of TIMSS-2011 8th-grade mathematics data from four proposed nations yield multiple latent classes using LCA techniques?
2. Does analysis of TIMSS-2011 8th-grade mathematics data from four proposed nations yield multiple latent classes using the MRM techniques?
3. Do LCA and the MRM analysis results differ in terms of:
  - a. Item fit parameters for TIMSS-2011 8th-grade mathematics?
  - b. Item class parameters for TIMSS-2011 8th-grade mathematics?
4. Are there associations between LCA and the MRM latent classes, nation, and gender for TIMSS-2011 8th-grade mathematics?

## **Introduction**

Large scale international assessment has become an important tool for countries to see how they perform compared to their rivals and neighbors and also to assess the progress made by their own education systems. Cross-cultural comparisons of results also help academicians to set international standards in education. Results from TIMSS-2011 showed a gap in mathematics achievement between those in top performing countries and Turkish 8th-graders. Although Turkey ranked 24<sup>th</sup> out of 56 participating countries, the country ranked 10<sup>th</sup> within advanced level students' results (Yücel, Karadağ, & Turan, 2013). Unfortunately, results show that variation in performance is very high within the population of Turkish students. As a result of this, equity in the Turkish educational system should be examined.

Turkey has begun to benefit from international assessments starting with TIMSS-1999. In 2002, the Turkish Educational System initiated its biggest steps in education reform since the early stages of the young Republic. There have been numerous developments over the last decade. The underlying purpose of these changes is to take the country to rank within the top ten big economies in the world by 2023, the 100<sup>th</sup> anniversary of the Republic of Turkey. The political party in office currently has made extensive changes to the system. However, there is no local or national tool to measure if those ongoing efforts had a positive or negative impact. The standardized tests for transfers within the school system are not designed to see if the changes are effective. In other words, TIMSS and the Programme for International Student Assessment (PISA) studies are the only way to examine the effects of recent changes.

Assessment practices affect grades, placement, advancement, instruction, curriculum, policy, and also funding (Toker & Green, 2014). The quality of the assessment used for any of these purposes is important. For example, analyses show that Finnish mathematics education practices are likely to explain the TIMSS achievement by Finnish students. The data created by international large scale assessment results are becoming increasingly useful for those who are the key players in an educational system such as academicians, administrators, policy makers, teachers, and also parents.

Although it can easily be said that international test data might be useful for handling some policy questions, and most likely these data are the only way to test the impact of differences in educational systems that vary across countries, barriers to drawing causal inferences based on such data exist (Schneider, 2009). There is no persuasive evidence that questions in different languages are valid and understood equally by all students or that the process used to respond to questions is the same (Holliday, 1999). Nonetheless, the popularity of international assessments is rising and their utility in making policy recommendations without considering such potential limitations is as well. And, rankings of countries on international assessments do not reflect where a country stands as far as world politics, army forces, and economic growth are concerned. The mean score of these standardized tests summarizes the performance of students overall and so shows that some standards differ greatly among countries and economies in ways that cannot simply be accounted for by the countries' different stages of economic development. Research shows that a country's wealth and spending on education affects educational success; but GDP per capita accounts for only 6% of the

differences between countries' average student performance. The remaining 94% reflects the fact that two countries of similar economic levels can show very different educational results (OECD, 2010). For example the total expenditure on the education of a 15-year-old Finland student represents the international average and is lower than it is in the US. However, the difference between these two nations in mean performance scores on the PISA science scale is about 50 points in favor of Finland (OECD, 2009a).

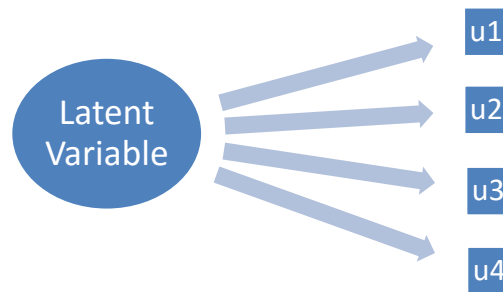
In this study, results of two statistical techniques for latent class estimation based on students' responses were compared. This study evaluated and compared the performance of LCA and MRM methods. Both techniques were used in terms of questionnaire validation to see if TIMSS-2011 data yielded different sub-groups within the selected nations. It is believed that comparison of different techniques, which have similar purposes and outcomes, might contribute advice and cautions for future studies where researchers have similar data.

The next section of this paper focuses on the statistical procedures LCA and MRM which are used in this paper.

### **Latent Class Analysis of Item Responses**

The first model discussed in this study is latent class analysis. LCA was first introduced in 1950-1959 by Lazarsfeld. He used the procedure mainly for clustering based on categorical observed variables. After 1950, the technique was studied widely by other statisticians. In 1974, Goodman developed an algorithm for obtaining maximum likelihood estimates of the model parameters so that the model could be applicable in

practice. He also studied polytomous manifest variables and multiple latent variables. Further, he completed an important work on the issue of model identification. In the same time period, Haberman (1979) presented the relationship between LC models and log-linear models for frequency tables with unknown cell counts. Some other important studies have also been conducted since then, such as development of models containing (continuous) covariates, ordinal variables, several latent variables, and repeated measures. Hagenaars (1990) proposed a general framework for categorical data analysis with discrete latent variables. This study was extended by Vermunt (1997b) and the resulting LC model with a latent variable and four observed variables ( $u_1, u_2, u_3, u_4$ ) is pictured in Figure 1.



*Figure 1.* The Latent Class Model (Vermunt & Magidson, 2004).

According to Collins and Lanza (2010), LC is a latent variable model used in the social, behavioral, and health sciences to determine if individuals can be divided into subgroups or latent classes based on an unobserved construct. The statistical procedure is related to confirmatory factor analysis (CFA: Harrington, 2008) and item response theory (IRT) (Lazarsfeld, 1950; Lord, 1952; Rasch, 1960/1980) when analyzing cross-sectional

data. LCA and confirmatory factor analysis have similar underlying ideas. However, in CFA the latent variable (i.e., factor) is continuous and has a normal distribution with indicators treated as continuous, while in LCA the latent variable (i.e., latent class variable) is categorical and has a multinomial distribution with indicators treated as categorical (Collins & Lanza, 2010). LCA is also conceptually similar to a one parameter IRT as a generalization of discrete response models (Samuelsen & Dayton, 2010); however, the latent variable is categorical in LCA whereas it is continuous in IRT (Collins & Lanza, 2010).

LCA is a statistical technique whose purpose is to identify class membership among subjects using categorical observed variables. Latent variables are not directly observed variables but are rather indicated by observed variables which are directly measured (see Figure 1). One of the main differences between LCA and other latent analyses is that LCA is person-oriented since it is focused on finding groups based on individuals' response patterns (Collins & Lanza, 2010). Based on this difference and the nature of the data, LCA was selected for use in this study

“to arrive at an array of latent classes that represents the response patterns in the data, and to provide a sense of the prevalence of each latent class and the amount of error associated with each variable in measuring these latent classes” (Collins & Lanza, 2010, p. 27).

Consider next the special case of the general modeling framework shown in Figure 2. The framework is characterized by using categorical latent variables, denoted by the circle  $c$  in Figure 2. Although the figure provides a general framework for all LCA related analysis, the categorical version of the analysis will be used in this study.



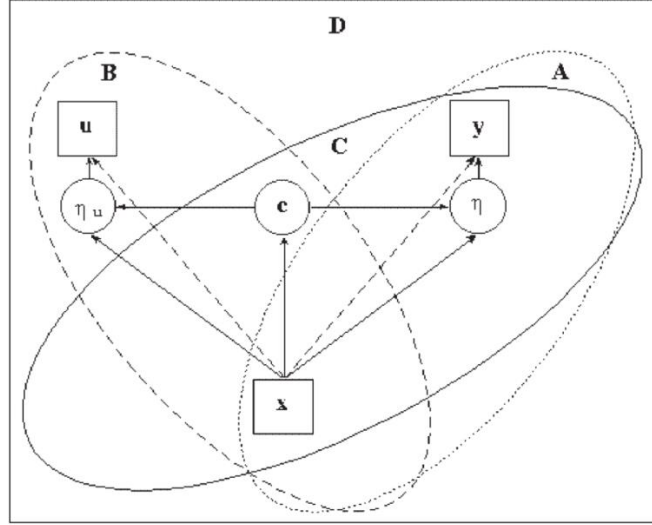


Figure 2. A general latent variable modeling framework (Muthén, 2001).

First, a general modeling framework of LCA as used in *Mplus* (Muthén & Muthén, 1998-2012) is shown to provide the basic mathematical model. This is followed by a discussion of latent class analysis as applied to this study.

According to Muthén (2001) (see Figure 2),  $c$  denotes a categorical latent variable with  $K$  classes,  $c_i = (c_{i1}, c_{i2}, \dots, c_{iK})'$ , where  $c_{ik} = 1$  if individual  $i$  belongs to class  $k$  and zero otherwise. The framework has two levels:  $c$  related to  $x$  and  $u$  related to  $c$  and  $x$ ;  $c$  is related to  $x$  by a multinomial logistic regression using the  $K - 1$ -dimensional parameter vector of logit intercepts  $\alpha_c$  and the  $(K - 1) \times q$  parameter matrix of logit slopes  $\Gamma_c$ , where for  $k = 1, 2, \dots, K$

$$P(c_{ik} = 1 | \mathbf{x}_i) = \frac{e^{\alpha_{c_k} + \Gamma'_{c_k} \mathbf{x}_i}}{\sum_{j=1}^K e^{\alpha_{c_j} + \Gamma'_{c_j} \mathbf{x}_i}}, \quad (1)$$

where the last class is a reference class with coefficients standardized to zero,  $\alpha_{ck} = 0$ ,  $\gamma_{ck} = 0$ . For  $\mathbf{u}$ , conditional independence is assumed given  $\mathbf{c}_i$  and  $\mathbf{x}_i$ ,

$$P(u_{i1}, u_{i2}, \dots, u_{ir} | \mathbf{c}_i, \mathbf{x}_i) = P(u_{i1} | \mathbf{c}_i, \mathbf{x}_i) P(u_{i2} | \mathbf{c}_i, \mathbf{x}_i) \dots P(u_{ir} | \mathbf{c}_i, \mathbf{x}_i). \quad (2)$$

The categorical variable  $u_{ij}$  ( $j = 1, 2, \dots, r$ ) with  $S_j$  ordered categories go an ordered polytomous logistic regression (proportional odds model), where for categories  $s = 0, 1, 2, \dots, S_j -$

1 and,  $\tau_{j,k,0} = -\infty$ ,  $\tau_{j,k,S_j} = \infty$ ,

$$u_{ij} = s, \text{ if } \tau_{j,k,s} < u_{ij}^* \leq \tau_{j,k,s+1}, \quad (3)$$

$$P(u_{ij} = s | \mathbf{c}_i, \mathbf{x}_i) = F_{s+1}(u_{ij}^*) - F_s(u_{ij}^*), \quad (4)$$

$$F_s(u^*) = \frac{1}{1 + e^{-(\tau_s - u^*)}}, \quad (5)$$

where for

$$\mathbf{u}_i^* = (u_{i1}^*, u_{i2}^*, \dots, u_{ir}^*)', \quad \boldsymbol{\eta}_{ui} = (\eta_{u_{1i}}, \eta_{u_{2i}}, \dots, \eta_{u_{ri}})' \quad (6)$$

and conditional on class  $k$ ,

$$\mathbf{u}_i^* = \boldsymbol{\Lambda}_{uk} \boldsymbol{\eta}_{ui} + \mathbf{K}_{uk} \mathbf{x}_i, \quad (7)$$

$$\boldsymbol{\eta}_{ui} = \boldsymbol{\alpha}_{uk} + \boldsymbol{\Gamma}_{uk} \mathbf{x}_i, \quad (8)$$

where  $\boldsymbol{\Lambda}_{uk}$  is an  $r \times f$  logit parameter matrix differing across the  $K$  classes,  $\mathbf{K}_{uk}$  is an  $r \times q$  logit parameter matrix differing across the  $K$  classes,  $\boldsymbol{\alpha}_{uk}$  is an  $f \times 1$  vector logit parameter vector differing across the  $K$  classes, and  $\boldsymbol{\Gamma}_{uk}$  is an  $f \times q$  logit parameter matrix differing

across the  $K$  classes. The thresholds may be stacked in the  $\sum_{j=1}^r (S_j - 1) \times 1$  vectors  $\tau_k$  differing across the  $K$  classes.

It is important to emphasize that (7) does not include intercept terms given the existence of  $\tau$  parameters. Furthermore,  $\tau$  parameters have opposite signs than  $u^*$  in (7) because of their interpretation as cutpoints or thresholds that a latent continuous response variable  $u^*$  goes beyond or falls below (Agresti, 1990). For example, with a binary  $u$  scored 0/1 (5) leads to

$$P(u = 1 | \mathbf{c}, \mathbf{x}) = 1 - \frac{1}{1 + e^{-(\tau - u^*)}}. \quad (9)$$

For example, the higher the  $\tau$  the higher  $u$  needs to be to exceed it, and the lower the probability of  $u = 1$ .

A latent categorical variable is used to identify unobserved heterogeneity in latent class analysis. In this case, the specific goal is to find groups (latent classes) of individuals who are similar in response patterns. It is presumed that an adequate number of latent classes for the categorical latent variable results in conditional independence among the observed variables (Collins & Lanza, 2010). Since the latent class variable is the only source of dependence among the outcome variables, the latent class analysis is similar to factor analysis with uncorrelated residuals (Collins & Lanza, 2010; Samuelsen & Dayton, 2010; Wang & Wang, 2012).

Muthén (2001) explains that latent class analysis typically uses categorical variables  $\mathbf{u}$  of the latent class variable  $c$ . The variables of  $\mathbf{u}$  are binary, ordered

polytomous, or unordered polytomous. As a result of the conditional independence specification, the joint probability of all  $u$ 's is

$$P(u_1, u_2, \dots, u_r) = \sum_{k=1}^K P(c = k) P(u_1|c = k) P(u_2|c = k) \dots P(u_r|c = k). \quad (10)$$

Above model has two types of parameters. The distribution of the categorical latent variable is shown by  $P(c = k)$  expressed in terms of the logit parameters  $\alpha_{ck}$  in (1). The conditional  $u$  probabilities are indicated via logit parameters in line with (9) where for a binary  $u$   $\text{logit} = -\tau_k$  for class  $k$ , i.e. the  $u^*$  part of (7) is not needed. Almost identical to factor analysis, the conditional  $u$  probabilities present an interpretation of the latent classes such that some results represented by the different  $u$ 's are more or less likely in some latent classes than others. The latent class counterpart of factor scores yields posterior probabilities for each participant belonging to all classes as computed by Bayes' formula;

$$P(c = k|u_1, u_2, \dots, u_r) = \frac{P(c = k) P(u_1|c = k) P(u_2|c = k) \dots P(u_r|c = k)}{P(u_1, u_2, \dots, u_r)}. \quad (11)$$

According to Samuelsen and Dayton (2010), the primary assumptions of LCA are as follows:

- Number of classes specified by the model is correct.
- There is only one latent class for one respondent
- In one latent class all respondents are homogenous.

Based on these assumptions one fundamental concept of LCA is local independence where latent class membership is known when observed responses are independent.

There are four main steps to estimate a simple LCA model:

- 1- find the optimal number of classes,
- 2- assess the quality of the classification of latent class membership,
- 3- define the latent classes
- 4- predict latent class membership (Wang & Wang, 2012)

To determine the optimal number of classes a series of LCA  $k$ -class models are compared to  $k-1$  class models iteratively. Since  $\chi^2$  statistics are inappropriate in the presence of too many zero indicator cells in the contingency table, it is not appropriate to use  $\chi^2$  to determine model fit (Wang & Wang, 2012). According to Muthén (2004), a  $k-1$  class model is a unique version of the  $k$ -class model with one latent class probability value set to zero. As a result, the difference in the log-likelihood between two of the models does not follow a  $\chi^2$  distribution.

There are different fit indices used in LCA model fit comparisons such as the following information criterion indices: AIC (Akaike, 1973, 1983), consistent AIC (CAIC; Bozdogan, 1987), BIC (Schwarz, 1978), \Lo-Mendell-Rubin likelihood ratio test (LMR LR: Lo, Mendell, & Rubin, 2001), and an adjusted version of LMR LR. There is also a bootstrap likelihood ratio test (BLRT) developed by McLachlan and Peel (1987, 2000). The model log-likelihood based and penalty terms related to model complexity type fit indices are commonly used to compare different LCA models. *Mplus* provides

three different types of information criterion indices such as AIC, BIC, and ABIC.

Smaller values of these indices shows better model fit. A model with the lowest BIC or AIC is preferred. According to Lin and Dayton (1997), if the model is more complex AIC provides better model fit information than other indices.

Once the possible optimal number of classes is fit, cases are loaded into latent classes. Based on the response pattern of an individual, the probability of latent class membership is measured via posterior class-membership probability (Wang & Wang, 2012). The determination of latent class membership is not definite yet it is based on the most likely latent class assessed via the highest estimated posterior class-membership probability. A probability close to 1.0 shows a very low chance of misclassification of that individual. For example, if there is a 4-class LCA model, and for an individual estimated posterior class-membership probability scores for Classes 1, 2, 3, and 4 are as follows, 0.07, 0.09, 0.10, and 0.74, respectively, the individual will be assigned to class 4. The probability of the case being assigned to the correct class is 0.74 and probability of false classification will be  $(0.07 + 0.09 + 0.10) = 0.26$ . Since, in practice, it is almost impossible to have a posterior probability score of 1.0, a general guide according to Nagin (2005) is 0.70 or greater for assignment to a class.

*Mplus* provides another criterion calculated by Kamakura and Wedel (2000) called REN which is based on Celeux and Soromenho's (1996) work called entropy (EN). This criterion is based on an entropy term calculated on the basis of the posterior probabilities for every sample unit and mixture component. The entropy criterion

introduced assesses the ability of a mixture model to provide well-separated classes and is derived from a relation underscoring the differences between the maximum likelihood approach and the classification maximum likelihood approach to the mixture problem (Celeux & Soromenho, 1996). The values of REN range from 0.0 to 1.0 where a higher value shows a better classification. Although there are no clear cut-off points, a value of 0.80 is high, 0.60 is medium, and 0.40 is considered low entropy (Clark, 2010). After all individuals are classified into latent classes it is important to note the size of each class. To have a meaningful classification, sizes of each class should not be too small or too big. Also it is important to have theoretically meaningful and interpretable classes (Wang & Wang, 2012).

Just as in factor analysis, it is important to define classes in a way that makes sense. Once a set of latent classes are decided upon, the researcher needs to ensure that each latent class is meaningful and interpretable. The main goal of an LCA analysis is to explain heterogeneity in the data set. This explanation is based on the patterns decided by the statistical analysis. As a result of meaningful and interpretable latent class determination, the identified model will make sense to the researcher's audience. Also, even if the model is identified and meets all requirements of a mathematical analysis, if one cannot supply a theoretically interpretable latent class, the estimated model will not be useful (Wang & Wang, 2012). At this point, TIMSS data demographics such as gender and country of origin can be used to interpret the results.

The final step of the LCA is the class membership prediction. For this purpose, during the analysis covariates can be readily included. This gives an advantage to LCA over traditional cluster analysis. It is possible to run the analysis and include covariates at the same time (Muthén, 2004). On the other hand, a well-known problem in LCA modeling is that the model might provide incorrect parameter estimates due to difficulties in converging on the global maximum likelihood, but rather provide incorrect parameter estimates based on local maxima (Wang & Wang, 2012). One practical solution to this problem is to estimate the model with different sets of random starting values.

### **An example of LCA**

Higginbotham (2013) studied the latent factor structure of the November 2011 version of the Air Force Academy's Character Mosaic Virtues (CMV) questionnaire. He used the item responses from a CMV nine factor *post hoc* modified model as the input data for the LCA. There were 27 items in the model. Items were designed to measure character *virtues* based on the following nine theoretical constructs: *courage*, *accountability*, *humility*, *duty*, *care for others*, *self-control*, *respect for human dignity*, *attention to detail*, and *excellence*. The item responses were coded to dichotomous responses with "very much like me" and "like me" recoded as "like me" with a value of 1, while the item responses "neutral," "unlike me," and "very much unlike me" were recoded as "unlike me" with a value of 0.

He defined three latent classes based on the estimated posterior probabilities. Most cadets (n = 101) were assigned to class 1, 95 cadets were assigned to class 2, and 57



cadets were assigned to class three. Classes were defined as follows: strong identification with virtues, moderate identification with virtues, and weak identification with virtues (Higginbotham, 2013).

### **The Mixture Rasch Model**

The mixture Rasch model was first introduced by Rost in 1990. The model was proposed to bind the Rasch model with latent class analysis. It assumes that the Rasch model holds for all participants within a latent class, but it allows for different sets of item parameters between the latent classes (Rost, 1990). Since it assumes latent classes for which separate Rasch models hold, the model is applied to validate responses to an exam or questionnaire. Also, Rost states that if a model with two or more latent classes identified fits better than a model with one latent class, the measurement invariance assumption is violated and a single Rasch model is not a fit. When there are several latent classes in the data, separate Rasch models with separate sets of item difficulties are required. These different sets are considered latent clusters in the sense that they are not determined by covariates (Frick, Strobl, & Zeileis, 2015).

According to Rost (1990), item parameters might differ as a result of poor construction of items, use of different type of solving strategies by participants belonging to different latent classes, or different cognitive processing styles of participants across subpopulations. The mixture Rasch model is a unidimensional model, though the supposed dimension changes across the classes. Item difficulty estimates should remain constant for different clusters of people in a unidimensional Rasch model. The MRM, on

the other hand, can account for data when difficulty patterns of items consistently differ in subclasses of the population. This gives the MRM an advantage over a unidimensional Rasch model where the MRM allows item parameters to differ across subclasses of the population, when the unidimensional Rasch model does not fit for the entire population (Rost, 1990; Rost & von Davier, 1995). Since the Rasch model has some strict item and homogeneity assumptions, the MRM becomes useful when some population and item homogeneity assumptions are relaxed. Mixture Rasch models can detect participant heterogeneity and the related item structures, the size of latent classes, and the latent score distribution (Baghaei & Carstensen, 2013).

However, the MRM is still a Rasch model since each subset of population can be broken down separately with a unidimensional RM (Rost, Carstensen, & von Davier 1997). According to Rost (1990), rather than rejecting an entire data set for fit purposes, the MRM study can easily be applied in such situations and study different cognitive processes for latent classes of the population. The probability of a correct response to an item relies on more than one person ability dimension in multidimensional Rasch and IRT models. However, in the MRM the probability of a correct response to an item relies on one person ability dimension and also a categorical variable, called the latent class to which the participant belongs. One disadvantage of latent class models is the requirement of consistent response probabilities for all individuals in a latent class. Research has shown that for every cognitive structure or response strategy, multiple latent classes are needed in order to account for individual differences in ability. Rost (1990) suggests that

a generalized latent class model allowing for ability differences within classes should be used in such cases.

Rost (1990) explains the proposed model via a series of following mathematical formulas. Let  $p_{vig}$  indicates person  $v$  answering “yes” or correctly answering item  $i$  and this person belongs to latent class  $g$ . One can say that subjects’ response probabilities can be shown by the dichotomous Rasch model

$$p_{vig} = \frac{\exp(\tau_{vg} + \sigma_{ig})}{[1 + \exp(\tau_{vg} + \sigma_{ig})]}, \quad (1)$$

where  $\tau_{vg}$  is the participant’s ability and  $\sigma_{ig}$  is the item easiness parameter. Within each latent class  $g$  an indeterminacy constraint  $\sum_i \sigma_{ig} = 0$  must hold. Furthermore, if the researcher thinks latent classes are mutually exclusive and exhaustive, structure of the latent class is as follows:

$$p_{vi} = \sum_g \pi_g p_{vig} = \sum_g \pi_g \frac{\exp(\tau_{vg} + \sigma_{ig})}{[1 + \exp(\tau_{vg} + \sigma_{ig})]} \quad (2)$$

where  $p_{vi}$  is the unconditional response probability and  $\pi_g$  is the class size parameter or “mixing proportion with constraints between 0 and 1 and  $\sum_g \pi_g = 1$  .

Since none of these equations yet define the entire model because they do not specify how to deal with the person parameter,  $\tau_{vg}$  , it is important to control the person parameters using a Rasch-like model structure. To get the likelihood function, it is important to obtain the pattern of probabilities  $p(x)$  which is  $x = (x_1, x_2, \dots, x_i, \dots, x_k)$  where  $x_i = 0$  or  $1$ . The formula for the pattern probability can be written as follows:

$$p(x) = \sum_g \pi_g p(x | g) \quad (3)$$

where  $p(x | g)$  is the product of response probabilities defined by Equation 1 over all items. In the Rasch model the number of correct item responses is used to estimate  $\tau$ . So all participants with the same score  $r$  have the same  $\tau$  score. As a result of this, the pattern probability  $p(x | g)$  can be rewritten with the score  $r$  associated with a given pattern as follows

$$p(x | g) = p(x | g, r) \cdot p(r | g). \quad (4)$$

This factorization is quite important and useful since only the first factor depends on the item parameters  $\sigma_{ig}$ ,

$$p(x | g, r) = \exp(\sum_i x_i \sigma_{ig}) / \Phi_r[\exp(\sigma)] \quad (5)$$

In this formula,  $\Phi$  values are the symmetric functions of order  $r$  of the delogarithmized item parameter values. Moreover, only the second factor depends on the ability distribution in class  $g$ . The MRM is also a “distribution free” model just like the simple Rasch model.

A combination of all these elements defines the likelihood function of the model as follows;

$$L = \prod_x \{ \sum_g \pi_g \pi_{rg} \exp(\sum_i x_i \sigma_{ig}) / \Phi_r[\exp(\sigma)] \}^{n(x)}, \quad (6)$$

where  $n(x)$  denotes the observed number of response patterns  $x$ , and the score probabilities  $\pi_{rg} = p(r | g)$  have been rewritten by using Greek letters for renaming the model parameters.

Therefore the number of independent model parameters is constructed as follows:

- a.  $h - 1$  class size parameters  $\pi_g$ , where  $h$  is the number of classes,

- b.  $(k - 1)h$  class-specific item parameters, where  $k$  is the number of items measured and must be lowered by 1 because of the norming constraint, and
- c.  $2 + h(k - 2)$  class-specific score probabilities, because one parameter in each class depends on the sample size and the class size, and the two parameters for the 0 and 1 vectors are class independent

As can be noted, the Rasch model is a one-class solution of the proposed model in Equation 17. Also the same Equation is a special case of simple latent class analysis.

The Rasch model is a useful tool to generate item difficulty estimates. However it can be quite difficult to meet some assumptions of the Rasch model. It is possible that some items might behave differently for subgroups or participants' responses might depend on the latent class to which they belong (von Davier & Yamamoto, 2007). Basically the MRM is a Rasch model with an added latent class structure. It is assumed that item parameters depend on the particular latent class. This latent class structure is useful when item difficulty differs for different sub-groups and also if different participants use different strategies to answer items. Because of the potential for different item parameters, a Q-index is calculated for item fit for each class. The Q index shows the relationship between items and each latent class. The Q index is calculated based on the log-likelihood of the observed item-response pattern. "The fit of an item  $i$  is measured with the conditional probability of its observed item response vector" (von Davier, 2001b, p. 76). The Q index values are between 0 and 1, where 0 represents perfect fit and 1 represents perfect misfit or negative discrimination. A Q index of .50

shows no relation of the item to the trait or participants' random response behavior. The standardized form of the Q index (ZQ) with zero mean and variance of unity (which can be assumed to be asymptotically normal) is also provided by WINMIRA 2011 (von Davier, 2001b). The familiar  $\pm 1.96$  standard error boundary of a 95% confidence interval can be used for the interpretation of a standardized Q index. In this paper, TIMSS-2011 items were studied to explore if items differed in terms of difficulty when there were data from multiple participating nations.

In the MRM, there are different sets of item parameters estimated for each class. It is presumed that one participant only belongs to one latent class where the class membership is unknown. Since class membership is unknown, the probabilities of being in each class are estimated. However, one of the main critiques of the MRM is the difficulty of interpreting the qualitative meaning of the class membership (Embretson, 2006). The MRM can be used for different goals, e. g.,

- a. to test fit of a Rasch model via comparing a one-class solution to two- or multi-class solution,
- b. to identify a Rasch-scalable subpopulation,
- c. to analyze rating data when different subsets have different response sets,
- d. for profile analysis purposes when a set of items have ordinal responses,
- e. to measure a latent ability when different participants apply different solution strategies for answering questions (von Davier, 2001).

To estimate parameters in the MRM an iterative algorithm called estimation-maximization (EM algorithm) or iterative proportional fitting is used since latent classes are not known before the analysis is done. The EM algorithm works in two steps:

- 1- Within each (E)stimation-step, for each subpopulation, the expected frequencies of the sufficient statistics for the model parameters are calculated via computation of posterior probabilities given the current parameter estimates.
- 2- Within each (M)aximization-step, by using the sufficient statistics from the previous E-step, maximum likelihood estimates in each subpopulation are calculated (Rost, 1990).

### **An example of MRM analysis**

Baghaei and Cartensen (2013) applied MRM analysis with a reading comprehension test. Results showed that a two-class solution fit better than a one-class solution. Class sizes were 50.5% and 49.4% respectively. Participants in Class 1 showed better performance in short text items whereas participants in Class 2 showed better performance in long text items. The latent classes showed a difference with respect to reading competence, where Class 2 had a significantly higher reading mean. Item fit assessed by a  $Q$  index which showed that the items fit well within the classes, other than one item which did not have good fit in Class 2. The authors suggested that texts with different lengths have different cognitive demands which in turn have an impact on the internal validity of the test in terms of its fit to the Rasch model.

## **Latent Class Analysis and the Mixture Rasch Model**

Since both techniques are used in educational sciences, it is important to summarize their similarities and differences. Rasch models assume that participants who have the same ability have similar item solution techniques, skills, and psychological procedures used for solution (Fischer & Molenaar, 2012). However, studies in cognitive psychology and standardized testing have suggested that participants at the same ability level might use totally different techniques and strategies and take different paths to arrive at a solution (Sigott, 2004; Sternberg, 1985). If so, the test construct may change for different participants depending on the paths they take for solving the items, which is a threat to construct validity. LCA and the MRM are statistical models used to examine this threat.

Analysis of examinee responses to test items typically rests on the assumption that item parameters are homogeneous across examinees; that is, the items are assumed to behave in the same way for all examinees. In a conventional Rasch analysis, a single difficulty parameter is estimated for each item, and all item difficulty estimates are located on a single dimension along with a single ability parameter for each examinee. However, when examinees systematically differ in the ways they understand or solve items, this assumption may no longer hold. Differences in item solution processes, for example, can give rise to differences in item position parameters and hence to different latent classes.



The fundamental concept underlying LCA is straightforward: some of the parameters of a statistical model differ across unobserved subgroups. These subgroups, which are posited to be nations in this case, are the categories of a categorical latent variable (Vermunt & Magidson, 2004). The mixture Rasch model, on the other hand, is based on the Rasch model (Rasch, 1960), and was introduced by Rost (1990). It is a mixture of a latent trait approach and a latent class approach to model qualitative and quantitative ability differences. The model assesses a set of items as a whole. Therefore, it is the set of item parameters for all items that is tested for differences between latent classes rather than each item parameter being tested individually (Frick, Strobl, & Zeileis, 2015).

LCA estimates relationships between indicator variables due to class membership only. Also it calculates class membership probabilities instead of fixed class memberships. For example, if there are four suspected classes in a data set the probability of a participant being in each class might be as follows: 0.76, 0.14, 0.08, and 0.02. Since LCA does not provide fixed class memberships for each case, another step takes place within the model selection process called “quality of the classification of latent class membership” (Wang & Wang, 2012). A criterion value from Nagin’s (2005) study is used to determine the quality (.70 and higher). Finally, LCA requires each latent class to be defined in a meaningful manner so variance within the population can be described. As a result of this, latent class interpretation is a very important step of LCA.

However, in the MRM, because each class of participants shows a different pattern of response, there are different parameter estimates for each class. The class-related differences in item parameter estimates (the relative difficulty of items) provides differences in how the construct being examined is understood by that class's respondents. Unlike LCA, the class assignment method the MRM uses is a fixed assignment procedure called modal class. One important point is that LCA's path for class membership divides the sample into different groups. Final class membership probabilities provide percentages rather than fixed class membership. At first, one might emphasize that LCA's procedure can provide statistical optimization. However while gaining statistical optimization, classification interpretability and usability can be lost. Also, in the case of a follow up study with same participants, 72% of one case cannot be invited to a focus group while 28% of the same case stays in another group (Dallas & Wilse, 2013).

The solution the mixture Rasch model provides on this matter is using item difficulty parameters. Since the main product of each class is item difficulty parameters, interpretation of classes is derived from differences in item difficulties. Therefore, there is no need to evaluate the quality of the classification of latent class membership, and to define the latent classes for modeling purposes in the MRM.

Several studies have examined international test data using LCA or MRM. Choi and colleagues (2015) used a mixture three-parameter logistic model to explore possible latent classes within the TIMSS-2007 mathematics dataset using internet access as their

covariate. Two latent classes were found, mainly formed around the test performance dimension. In another study, data from the 2006 PIRLS assessment (International Association for the Evaluation of Educational Achievement, 2008) were used to explore differential item functioning using possible latent classes. PIRLS is also an international exam similar to TIMSS and PISA. The latent class approach yielded three latent classes, showing proof of heterogeneity in students' response patterns (Oliveri, Ercikan, Zumbo & Lawless, 2014). Additionally, Zhang, Orrill and Campbell (2015) also studied the PISA-2009 dataset using responses for students in China where they explored two distinct latent classes via MRM analysis for both the mathematics and science sections of the exam.

A simulation study along with analysis of real data was conducted where researchers compared results of latent transition analysis (LTA) which is similar to LCA but with the inclusion of longitudinal data used to see changes in the latent classes over time, with a combination of LTA-MRM techniques (Cho et al., 2010). MRM analysis provided more useful results in both simulation and real data applications. Additionally, the study with the real dataset showed that the MRM-related technique detected the intervention effect clearly. To summarize, past research with LCA and MRM analyses have typically found multiple latent classes in international test data.

### **Log-linear Analysis**

In this paper, another statistical method, called log-linear analysis, is also used to compare the significance level of interaction between LCA, MRM and two of the

demographics such as nation and gender. Log-linear analysis is a method used widely in educational statistics to measure the associations between more than two categorical variables (Knoke & Burke, 1980). In the past, contingency tables--two-way tables formed by cross classifying categorical variables--were typically analyzed by using chi-square tests of association. If more than two variables were analyzed, the chi-squares for two-way tables were computed and then computed again for multiple sub-tables formed from them in order to examine if associations and/or interactions were taking place among the variables. Goodman and Kruskal (1979) analyzed cross-classified data with multiple categorical variables and changed the field dramatically with the publication of a series of papers on log-linear models.

Log-linear analysis is a more complex application of two-way contingency tables. The conditional relationship between two or more categorical variables is examined by taking the natural logarithm of the cell frequencies within a contingency table. Although log-linear models can be used to analyze the relationship between two categorical variables (two-way contingency tables), a more common version of the analysis called multi-way contingency tables involving three or more variables was used to examine the relationships between expected latent class memberships for the MRM and LCA analysis as well as nation and gender (Gupta & Kapoor, 2000). The variables analyzed by the model were all treated as “response variables” which means, there were no distinctions made between independent and dependent variables. Hence, log-linear models only demonstrate associations between variables.

## **TIMSS-2011**

As the name indicates, The Trends in International Mathematics and Science Study is an international mathematics and science study in which numerous countries participate. TIMSS exams are administered at two grade levels. TIMSS-2011 was conducted with 4th-graders and 8th-graders. Nations that participated have different characteristics. Some are large. Some are small. Some are rich. Some are poor. They vary in religious, ethnic, language, economic, and cultural traditions. They have different educational goals and different expectations from their curricula, and the meaning of achievement varies among these participating countries.

Although a common reason to participate in such a large scale assessment is to compare results with those of neighbors or competitors, each of the participating nations has unique reasons as well. Among those reasons are to see what the effects of applications of educational policies and practices of countries whose students regularly achieve success in mathematics and science are; also to create a benchmark of data within a nation so future assessment results can be used to measure progress .

There is no magic bullet for creating a better educational system, which means there is no clear path to be found by trying to simply copy neighbors which are ranked at higher positions in a large scale international assessment (Atkin & Black, 1997). In addition, same research shows that the educational systems which are admired are also not satisfied with their existing programs. It is important to consider that it can be risky to

alter the complete educational system based on the relationship between students' test results and other parts of the countries' educational system.

The main goal of TIMSS is to create an international benchmark where participating countries can use their own data to improve mathematics and science education (Robitaille & Robeck, 1996). TIMSS-2011 is the fifth in a four-year-cycle of assessments (previously administered in 1995, 1999, 2003, and 2007). The study is conducted in four-year-cycles to be able to assess progress in student achievement. TIMSS measures the mathematics and science proficiency of children in two main populations: 4th-grade and 8th-grade students. Since TIMSS is applied with 4<sup>th</sup> and 8th-graders, in four years, 4th-grade students will be 8th-graders. This four-year cycle has the advantage of being able to compare countries' educational progress. TIMSS was designed to align with mathematics and science curricula in the participating countries. TIMSS results assess the mastery level to which students have learned mathematics subjects and aptitudes likely to have been taught. TIMSS tests put an emphasis on questions and tasks that offer insight into the analytical, problem-solving, and inquiry skills and capabilities of students. Moreover, organizers requested students, teachers, and school principals in each participating nation to complete surveys with respect to the context for learning mathematics and science in addition to achievement testing, so answers might provide logical explanations for interpreting the achievement score results and to track changes in instructional practices (Shen, 2000).

At the beginning of every TIMSS cycle, an expert group comprising curriculum experts in mathematics and science from participating countries builds the framework for the coming test. Although this framework should be confirmed by all member countries as being representative of their country's curricula, since there are numerous countries participating and it is difficult to overlap all of those curricula, there is always some content that is not covered in the curriculum of every participating country. This problem is solved at the analysis stage of the test by removing the items to which a participant country objects. Using the appropriate technique, this item exclusion rarely has any positive or negative effect on a country's score (Toker & Green, 2012). Four countries were selected for comparison in this study. Turkey is the focus of the work as this is the researcher's home country. The USA was selected since the researcher is currently studying within the American education system. Finland and Singapore were selected because they are two top performing nations whose educational systems differ widely. It is anticipated that comparison of these four nations' test results and education systems can provide useful information to all parties of education such as policy makers, leaders, teachers, parents.

Starting in 1999, Turkey participated in three TIMSS studies including 2007 and 2011 (Yücel et al., 2013). There were 38 countries that took part in TIMSS-1999. This was the first time Turkey participated. The international mean mathematics score was 500 with a standard deviation of 100. Turkey's mean score was 429. Turkey ranked 31st in the study. In this study, the overall international mean mathematics score for 8th-graders' was 487 (TIMSS-R Turkey Report, 2003). TIMSS-2007 was the second study

in which Turkey decided to participate. There were 48 countries at 8th-grade level and Turkey ranked 31st in the study with a mean score of 432. The international mean mathematics score and standard deviation were a mean mathematics score of 500 and a standard deviation of 100 in TIMSS-1999. Turkey also participated in TIMSS-2011. The mean mathematics score was again 500 with a standard deviation of 100. Out of 56 countries participating in the 8th-grade mathematics test, Turkey ranked 24<sup>th</sup> with a score of 452.

By comparison, the United States placed 19<sup>th</sup> in 1999 with a score of 502, 9<sup>th</sup> in 2007 with a score of 508, and again 9<sup>th</sup> in 2011 with a score of 509. Finland administered the test to a random sample of 7<sup>th</sup> graders in both 1999 and 2011. From 1999 to 2011 the mathematics score of Finland dropped from 520 to 482 yet their 8th-graders ranked 8<sup>th</sup> with a score of 514 in 2011; Singapore has been within the top three best performing education systems in all TIMSS tests. The peak of Singapore scores was the 1995 participation of the country with a score of 643 and a ranking of 1<sup>st</sup> place. In 1999, Singapore was placed 1<sup>st</sup> again but their score decreased to 604. Gaining one point in 2003 placed Singapore at the top of the list once again. In 2007 Singapore went down to 3<sup>rd</sup> place with a score of 593. Finally, in 2011 they were ranked 2<sup>nd</sup> with a score of 611 (TIMSS 2011 International Results in Mathematics, 2012).

TIMSS exams are widely recognized as high quality measurement tools. The quality of the TIMSS exam is supported by academicians and professionals in the field of educational measurement and evaluation. Having the opportunity for international



comparison is one of the main benefits of the test. In other words, countries that are tested can see how well or how badly they do globally. There is also another advantage of the test. Since the test is administered in four year cycles and administered with 4<sup>th</sup> and 8<sup>th</sup>-graders, if a country participates in both 4<sup>th</sup> and 8<sup>th</sup>-grade there is an opportunity to see the development of the country within its' own educational system. It is highly recommended to participate in all four-year cycles to see the development of a country in education basics. Turkey has participated in three tests.

TIMSS tests are high-profile international tests for mathematics and science achievement. It is important to apply high quality standards and advanced measurement techniques to address reliability and validity concerns. Since the test has major effects on both countries' education systems and political decisions, it is important to ensure that the results are not impacted by outside factors. Organizers of TIMSS apply strict procedures to ensure the test is reliable. Reliability is a large part of assuring the quality of measurement. But since reliability is not enough by itself to support the worth of a test, it is also important to assess the validity of the test.

The validity of test items is studied by organizers with the collaboration of participating countries. There is agreement on the part of educators in mathematics and science for assessment of both 4<sup>th</sup>- and 8<sup>th</sup>-grade students. This agreement means that the test items included in the tests measure agreed-upon elements of mathematics and science (TIMSS 2011 International Results in Mathematics, 2012). To achieve the goal of validity, organizers of TIMSS follow a strict procedure where they include experts from

participating countries for curriculum coverage, translation, scoring, and etc. purposes. Every step of the test is also controlled by the organizers in order to prepare and administer a valid and reliable assessment.

However, reliability and validity procedures are not the main concerns of the current study. For comprehensive information related to the reliability and validity of the TIMSS exams, see the TIMSS technical reports (TIMSS 2011 International Results in Mathematics, 2012). The next section of this study briefly reviews the educational systems of Turkey, USA, Finland, and Singapore. The intent of this brief review is to provide background information about why data from these countries were selected for use in this study as the education systems differ substantially.

### **Education Systems of Compared Countries**

#### *Turkey*

Turkey has a population of over 77 million (Turkish Statistical Institute, 2015). It is expected to reach 84 million by 2023, 93 million by 2050, and 89 million by 2075 (10<sup>th</sup> Development Plan, 2013). According to data from Turkish Statistical Institute, during the 2006-2007 school year, the pre-school enrollment rate was 24%, the primary school (including middle school) enrollment rate was 96.3%, and the secondary school enrollment rate was 86.6%. Applying gender-gap closing projects like “Girls, Let’s Go to School” increased the enrollment rate, especially for secondary education in the 2012-2013 school year. The pre-school enrollment rate went up by 20%, reaching 44% of the

4-5 year old pupil population. The enrollment rate for primary education was 97.6%. For secondary education, the number went up by 10.2% and reached 96.8%. Higher education was also affected by the country's educational progress.

Pre-school education is one of the main concerns of the Ministry of National Education (MONE). The Turkish National Education System is organized by laws on education and training, development plans, government programs, and recommendations of the National Education Councils. Recently the main focus was to increase the enrollment rate of pre-school education.

The National Education System is structured on the National Education Basic Act No. 1739, which has two main parts, named "formal education" and "non-formal education." Formal education is the standard education given within a school for individuals in a certain age group (excluding higher education where there are no age restrictions) and at the same level, under certain curricula developed in accordance with the goal stated in Act No. 1739. Formal education includes pre-primary, primary school, lower secondary school, upper secondary, and higher education institutions (Buyruk, 2015). Figure 3 provides a schematic of the general organization of the system.

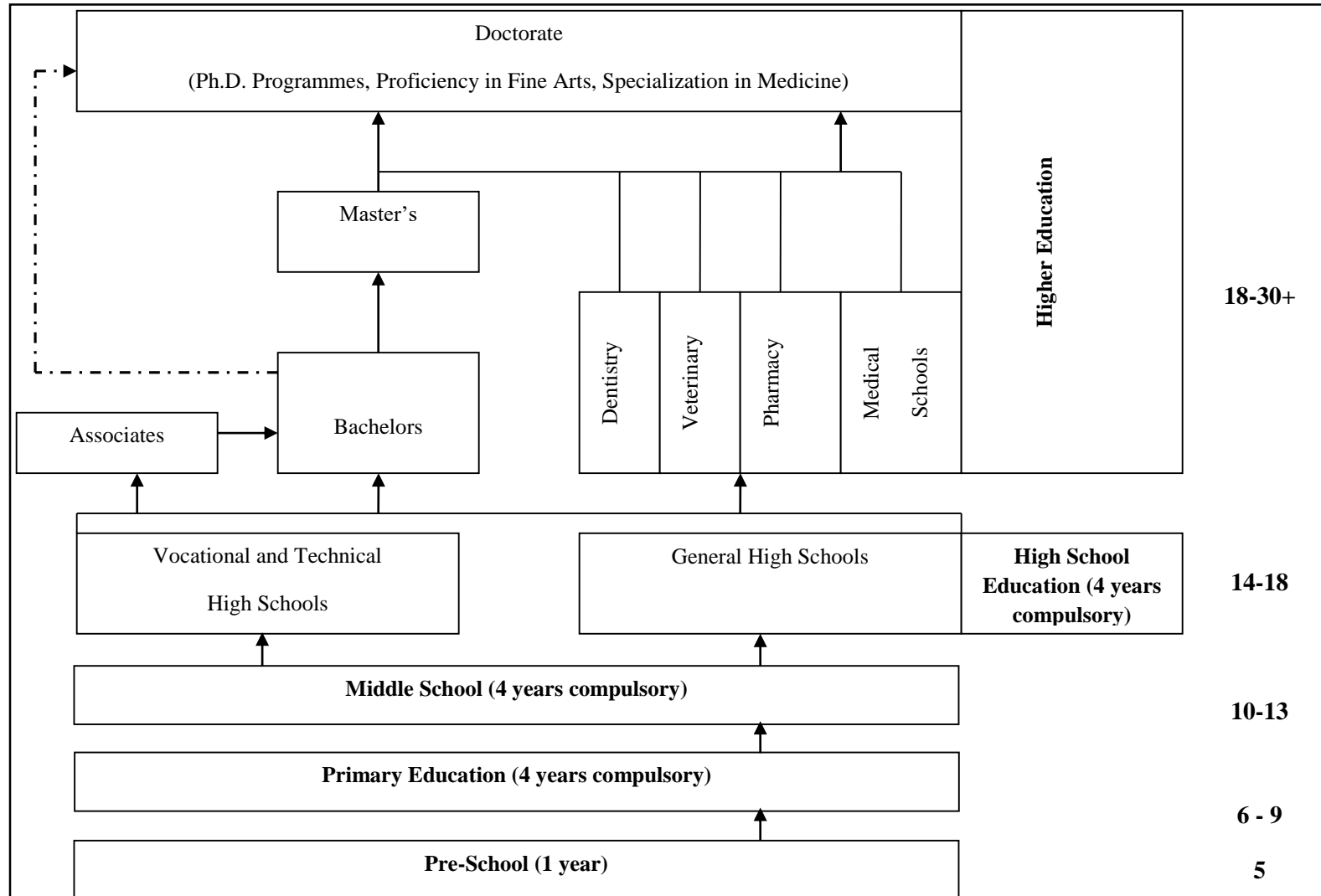


Figure 3. Turkish Educational System (Age shown on the right) (MONE, 2014).

The primary objective of the Turkish Education System is to ensure that every child masters the basic knowledge, skills, behaviors, and habits to become a good citizen, is raised in line with national moral concepts, and is prepared for life and becomes a happy citizen with a job parallel to his/her interests and skills.

#### *United States of America*

According to US Census Bureau (2015), the United States of America has a population of some 321 million. The American education system is one of the largest education systems in the world with a total of 57.4 million students being educated in public, charter, and private school systems.

The American education system begins with daycare. Since the majority of the population is working it is very common for most kids to attend early childhood education starting with daycare around the age newborn to 3 years old. At the age of 3-4, families have the option to attend pre-schools. When children are 5 years old they also can go to kindergarten. The school system is structured as primary and secondary school for a combined total of 12 years. U.S. educators frequently use the term K-12 education to refer to all primary and secondary education, from kindergarten prior to the first year (or 1st grade) of formal schooling, through secondary graduation (12th Grade). Although there are small differences in school systems throughout the country (sometimes even within the states), the following pattern is usually used in the community:

• *Elementary school (K-5), middle school (6-8), high school (9-12 U.S. children enter formal schooling around age 5.*

Elementary students are typically in one classroom with the same teacher most of the day. Recently schools have more art and music studios which prevents children being in the same class with the same children for most of the day. Some schools are designed to offer different options to children where they can attend classes based on their needs or interests. After completing elementary school, students proceed to junior high school (also called middle school in some districts), where they usually move from class to class each period, with a new teacher and a new mixture of students in every class. This is one of the unique characteristics of the western education system. Students can select from a wide range of academic classes and elective classes during elementary years. This gives them the option to focus on more specialized areas at early levels of their education.

During both Elementary and Middle School (or Junior High), children generally stay in the classroom an average of 6.5 to 7 hours per day (Institute of Education, 2015). Families might select before and after school programs which are generally made available through the schools. However, these programs are not free for most families. Financial assistance is available from schools budget, state or federal funded programs for some families. Most of the time, the family will have to pay for the cost of the after school program. Also if the program is located somewhere else, transportation is provided by student's school.

In High School, students are called freshman in their first year, sophomore in their second year, junior in their third year, and in their last and fourth year senior. Subjects have more variety than elementary school. Students generally sit in the classroom around 7.5 hours per day and must earn a certain number of credits in order to be awarded a High School Diploma – there is no final examination like in many other countries (Institute of Education, 2015).

The main requirement to enroll in postsecondary education is a high school diploma or equivalent. Some high schools offer college level courses where pupils transfer them to college level after successfully completing the requirements of the class. During their high school years, students are given "grades" for all their courses, and these are recorded. At the end of 12th Grade, the student's grades are averaged out to a "GPA" or Grade Point Average, which will often be used as a selection criterion when they apply to college or university along with some other documentation such as purpose statement, reference letters, and financial documents. Students in 12th Grade also take "SAT's", Scholastic Aptitude Tests, or "ACTs", American College Tests. These are the second principal tests used as criteria for admission to college or university; although these are still large scale assessments, they are not exams in the same way as are their European or Far Asian equivalents (French baccalauréat, German Abitur, English "A" levels, Turkey's University Entrance Exam), and are generally less demanding.

Although this is the general framework for transition to college level education, many students choose to attend 2 years of community college education where it is easier to be admitted and less expensive. Later they transfer their credits to 4-year colleges and complete their education (Institute of Education, 2015).



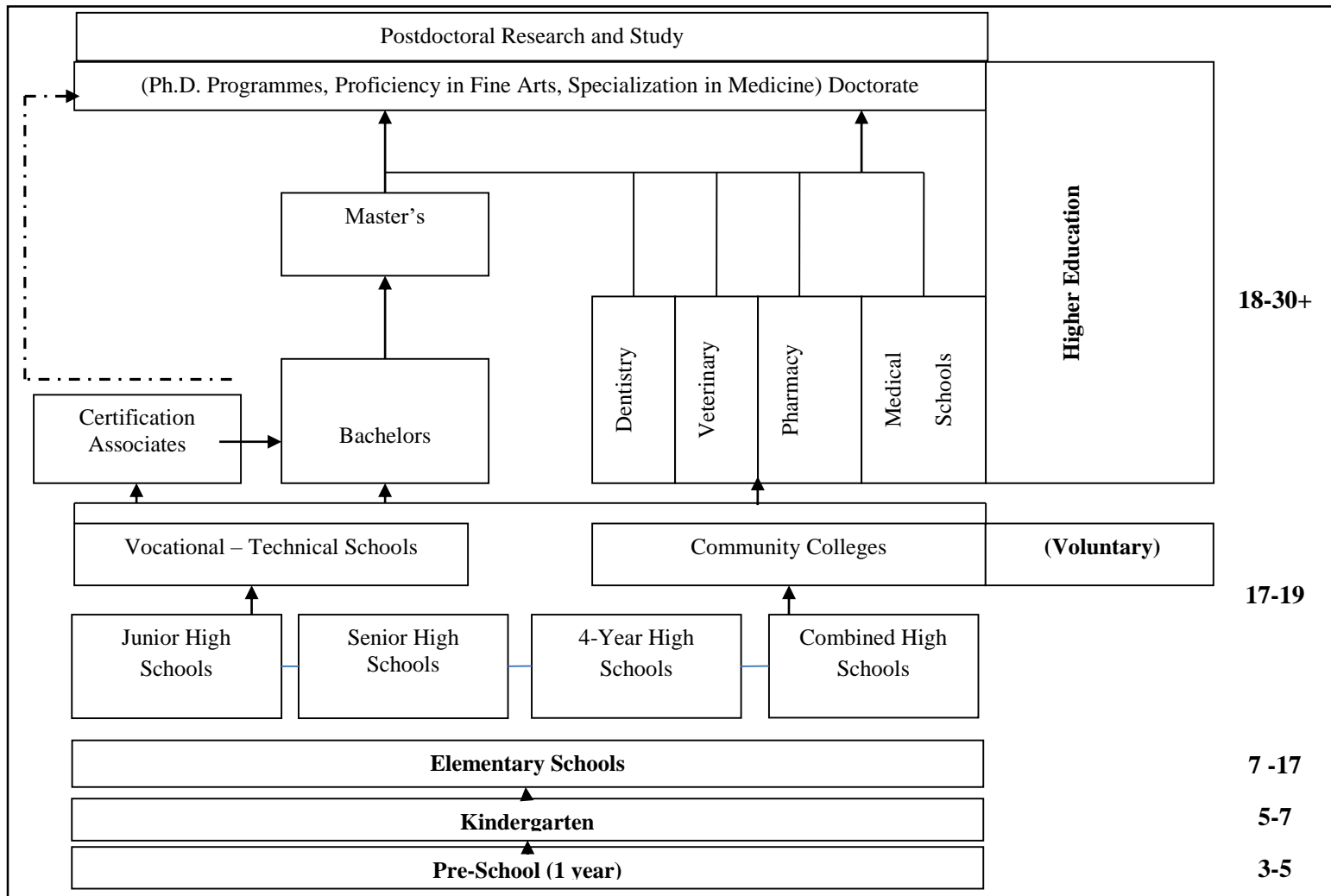


Figure 4. US Educational System (Age shown on the right) (US Department of Education, 2015).

## *Finland*

Academic life in Finland is different from most countries participating in TIMSS in the pace at which pupils enter academic life (Figure 5). Finnish students start school the year they turn seven. There is almost no or very little stress on academic education in a child's life before they start school (Kupiainen, Hautamäki, & Karjalainen, 2009). Every citizen has the right to attend early childhood education before the age of 6 but enrollment rates are very low (Kammerman, 2000). There is one year of preschool or kindergarten attendance for children to ensure school readiness.

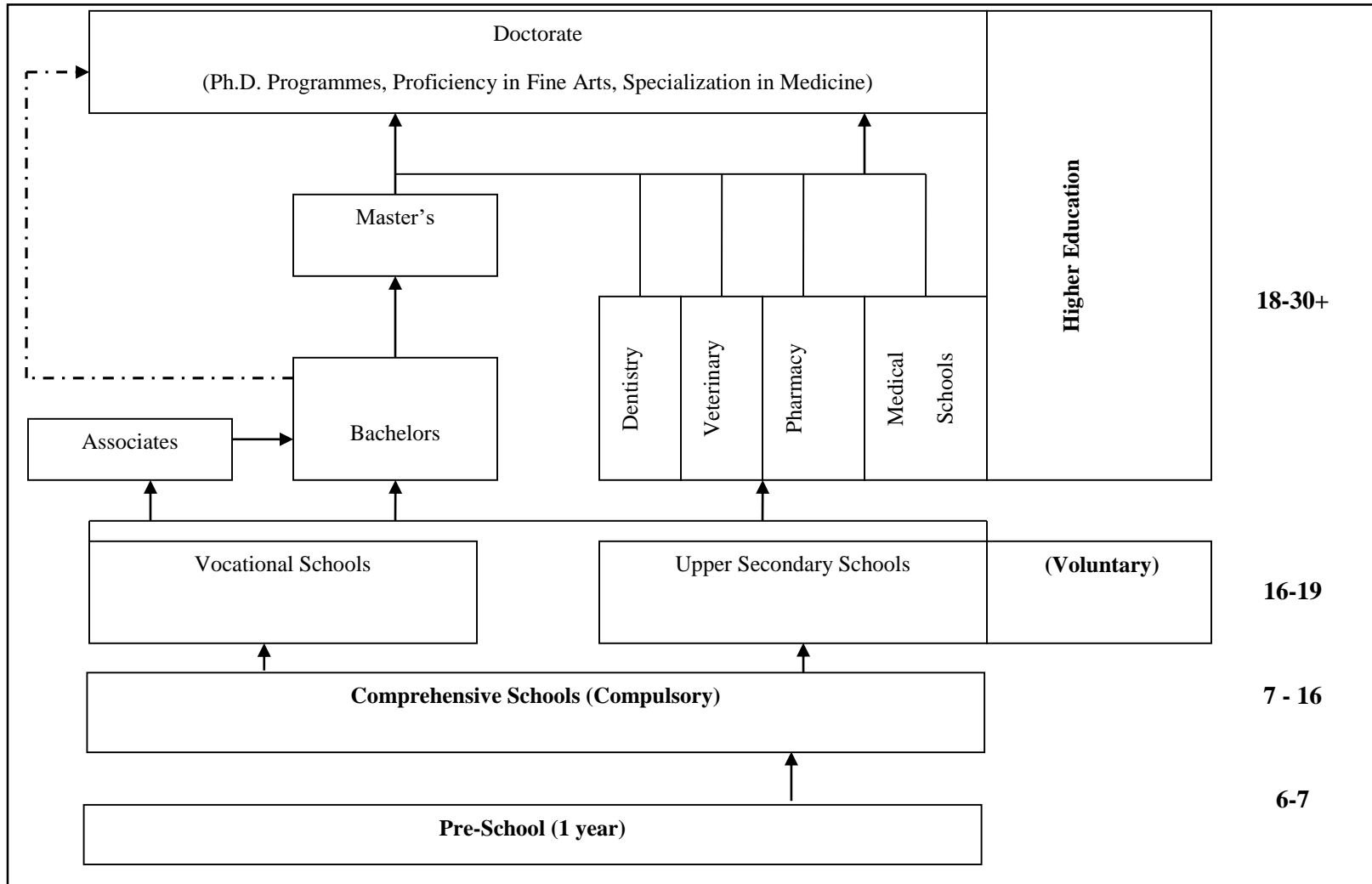


Figure 5. Finnish Educational System (Finnish National Board of Education, 2015).

Both early childhood education and kindergarten mostly focuses on the pupils' age-related development other than stress on academic achievement (Kupiainen et al., 2009).

The core of the Finnish education system is the compulsory nine year basic education between ages 7-16. The main goal of basic education is to support students' growth towards humanity and ethically responsible membership in society and to equip them with the knowledge and skills needed in life. Basic education in Finland is non-selective. Schools do not select their students. Every student is assigned to a nearby school, but they can also participate in another school with some restrictions (Finnish National Board of Education, 2015).

Every school follows a national core curriculum, which includes the goals and core contents of different subjects. The education leaders, usually the local education boards and the schools can independently come up with their own curricula with the condition of staying within the framework of the national core curriculum. This is quite similar to the independence of schools in the US.

### *Singapore*

The main purpose of education in Singapore (Figure 6) is to help students discover their talents, realize their potential, and develop a passion for learning which lasts through their life. The whole school system consists of three hundred sixty six schools which is smaller than some

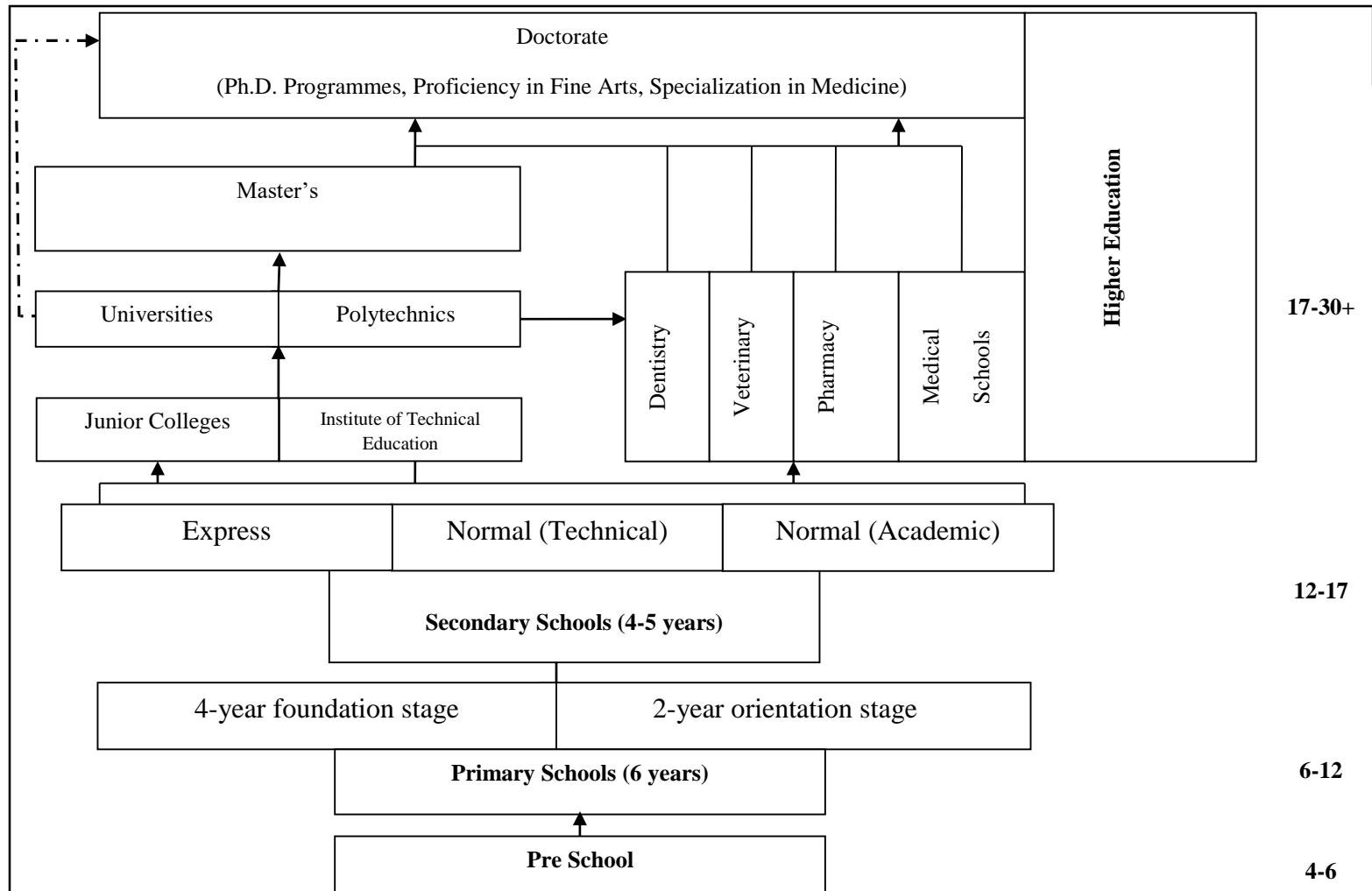


Figure 6. Singapore Educational System (Singapore Ministry of Education, 2015).

district in the US school system or a neighborhood school system in Turkey. Thinking Schools, Learning Nation (1997) and Teach Less, Learn More (2004) are the underlying projects that support the national education system of Singapore (Ng, 2007).

The system starts with pre-school education years where students attend between ages 4 and 6 years old (see Figure 6). There is a broad curriculum applied for three years with the intent to build self-confidence, learn social skills, and develop learning dispositions. These main characteristics underlie a strong foundation for children's future learning. There are total of ten kindergartens. The projected number is fifteen by the end of 2016 (Singapore Ministry of Education, 2015).

Primary education is 6 years of compulsory education where students attend a 4-year foundation stage from Primary 1 to 4 and a 2-year orientation stage from Primary 5 to 6. The overall goal of primary education is to provide students with a good grasp of English, the Mother Tongue, and Mathematics. There are no school fees. Schools apply a subject-based banding type of education which allows students to take a mix of standard and foundation subjects, depending on their mastery levels. Along with their education all students are encouraged to participate in Co-Curricular Activities (CCA) and Community Involvement Programs. At the end of the 6th year of primary education, there is a final examination called the Primary School Leaving Examination (PSLE: Singapore Ministry of Education, 2015).

Secondary education is based on how students perform on the PSLE. Based on their scores, the system places them in the Express, Normal (Academic) or Normal

(Technical) course. It is important that students face challenges based on learning abilities and interests. The different curricular emphases are designed to match students' own learning pace. The total number of years a student spends during his/her secondary education years is 4-5 years. Schools have some fees, but the cost is less than 20 US Dollars a month and is based on family income. There is a national examination called General Certificate of Education 'Ordinary' Levels (GCE "O") (for Express course) or General Certificate of Education 'Normal' Levels (GCE "A") (for Normal course) (Ministry of Education, 2002b). All students take part in at least one CCA; CCA performance is considered for admission to junior colleges, centralized institutes, polytechnics, and institutes of technical education.

Although the main goals of all these different education systems are similar and targeted to make their citizens' lives better, paths they are taking have some commonalities and some differences due to culture, geographic location, and economic and social differences. It is often believed that countries that are at the top of the TIMSS rankings or perform well on any other standardized test study are comfortable with their education system. In reality, the world economy is not where it was twenty years ago. Change is inevitable and so is educational change. There is no educational reform that will hold a nation's future in good hands for years on end. Each program has its own barriers to leap which brings about the need for sustainable educational reform where palsied parts can be seen and renewed over time.

It is anticipated that comparison of these education systems may help policy makers and education leaders better understand what differences exist among them, and what reforms a better ranking country has implemented that are successful. On the other hand, it is also important to see what problems a lower ranking country has and what efforts they are making to fix them. Following was taken from TIMSS 2015 Study Flyer which is published by TIMSS organizers and available online:

- *A major purpose of TIMSS is to provide important background information that can be used to improve teaching and learning in mathematics and science.*
- *TIMSS Advanced measures trends in advanced mathematics and physics for students in their final year of secondary school (twelfth grade in most countries). The assessment provides educational policy makers with valuable information about how many students are excelling at highly specialized material in a global context.*
- *Participation in TIMSS enables evidence-based decisions for educational improvement. High quality, internationally comparative data about student achievement in mathematics and science are important for monitoring and improving the health of a country's education system. Evidence of underperforming areas often spurs education reform, with subsequent assessments being effective monitors of changes in the educational system (TIMSS & PIRLS International Study Center, 2015, p. 1).*



Although TIMSS and similar exams provide useful background information, it is also emphasized that TIMSS organizers provide valid and reliable test results by which a participating system can compare its outcome globally. Also, if a country participates in all the tests conducted, results can be tracked over time for within system comparisons.

Furthermore, it is stated that TIMSS provides evidence of low performing areas which may trigger education reform, with four-year cycle re-assessments being effective monitors of changes in the system (TIMSS & PIRLS International Study Center, 2015). One can argue why relying on results of such a standardized test result to initiate a change in a system might end up being very costly. It is important to validate the quality of these results so participating systems can decide whether to use results to effect change or not. If so, should they link their reforms to future exam results? This study used two statistical procedures called latent class analysis and the mixture Rasch model to examine the item responses of students from four different nations to ascertain whether test results are stable under different analysis models.

### **Definition of Terms**

*LCA* – Latent Class Analysis. LCA is a statistical technique for exploring unknown class membership among participants using categorical and/or continuous observed variables.

*MRM* – The Mixture Rasch Model. The MRM is a Rasch model using item difficulty parameters to make inferences about differential behavior difficulties of similarly constrained or facilitated - latent - classes of people (Rost, 1990).

*TIMSS* – Trends in International Mathematics and Science Study is a large scale assessment provides an international perspective to participant nation and informs educational policy and reforms all over the world.

*PISA* – Programme for International Student Assessment is also a worldwide exam done by the Organisation for Economic Co-operation and Development (OECD) on 15-year old students.

## Chapter Two

### Method

#### Participants

Data used in this study were taken from the TIMSS-2011 8th-grade mathematics section administered in 2011. Students' responses to the items were used for both LCA and the MRM analyses. There were 26,596 8th-grade students from four different nations. Turkey participated with 6,928 students 49% of whom were girls and 51% boys. The USA participated with 10,477 students which were 51% girls and 49% boys. Finland participated with 4,266 students which were 48% girls and 52% boys. Singapore participated with 5,927 students which were 49% girls and 51% boys. The mean age for participating nations was 14.00 for Turkey, 14.20 for USA, 14.80 for Finland and 14.40 for Singapore. For the analysis purposes only 1,225 students from Turkey, 1,990 students from USA, 1,229 students from Singapore, and 768 students from Finland were selected with a total of 5,212 (see Table 1).

Table 1  
*Gender and Age of TIMSS-2011 Subjects (based on booklet selection)*

Nation	Count		Gender (%)				Mean Age	
			Girl		Boy			
	Selected	Population	Selected	Population	Selected	Population	Selected	Population
Turkey	1,225	6,928	48.70	49	51.30	51	14.08	14.00
USA	1,990	10,477	49.70	51	50.30	49	14.22	14.20
Singapore	1,229	5,927	49.40	49	50.60	51	14.39	14.80
Finland	768	4,266	50.30	48	49.70	52	14.75	14.40

Note: Gender is shown in percentages.

Although Finland participated with slightly under 4,500 students, TIMSS administrators asked participating nations to join with at least 4,500 students so that there would be enough respondents for each item (TIMSS 2011 International Results in Mathematics, 2012).

### **Instrument**

The TIMSS-2011 8th-grade mathematics test consisted of 217 items which included 118 multiple-choice items in 14 different booklets. Each booklet contained 10-18 items. Six of the mathematics blocks were released. Eight of them were kept confidential for evaluating trends in 2015. Out of 217 questions, there were 48 released with item text. Some booklets did not contain enough items (i.e., at least 10 items) so they were excluded from the study. Also some booklets had some overlapping items. Only Booklet One, Booklet Four, and Booklet Six were used due to having a larger number of released items in those booklets. The total number of released items included in these booklets is 40. According to TIMSS Technical Report (2012), some items were kept confidential so organizers could use the items in the future for trend analysis purposes. For item specific domain information please see [http://timssandpirls.bc.edu/timss2011/downloads/T11\\_UserGuide.pdf](http://timssandpirls.bc.edu/timss2011/downloads/T11_UserGuide.pdf). Three of the released items are shown in Figures 7, 8, and 9 below:

Content Domain	Main Topic	Cognitive Domain
NUMBER	Fractions and Decimals	Reasoning

Location of N on number line

$P$  and  $Q$  represent two fractions on the number line above.  
 $P \times Q = N$ .

Which of these shows the location of  $N$  on the number line?

A.

B.

C.

D.

Item Number: M032662

<b>Correct Response:</b>	<b>D</b>
--------------------------	----------

**Overall Percent Correct**

Education system	Percent correct
Chinese Taipei-CHN	53
Hong Kong-CHN	47
Singapore	45
Korea, Rep. of	44
Japan	43
Russian Federation	31
Sweden	30
England-GBR	29
Finland	29
Palestinian Nat'l Auth.	28
Israel	27
Oman	26
Syrian Arab Republic	25
Saudi Arabia	25
Jordan	24
Australia	23
Hungary	23
<b>International average</b>	<b>23</b>
<b>United States</b>	<b>22</b>
Qatar	22
Slovenia	21
Bahrain	21
New Zealand	19
Ukraine	19
Lebanon	18
Malaysia	18
Lithuania	18
Macedonia, Rep. of	17
Iran, Islamic Rep. of	16
Morocco	16
Italy	16
Norway	15
Armenia	15
United Arab Emirates	15
Turkey	15
Tunisia	14
Kazakhstan	14
Chile	14
Georgia	13
Ghana	13
Romania	12
Thailand	12
Indonesia	10

Benchmarking education system	Percent correct
Massachusetts-USA	44
Minnesota-USA	38
North Carolina-USA	36
Connecticut-USA	30
Quebec-CAN	29
Ontario-CAN	27
Alberta-CAN	24
Colorado-USA	21
Florida-USA	20
California-USA	19
Indiana-USA	19
Abu Dhabi-UAE	16
Dubai-UAE	14
Alabama-USA	13

▲ Percent higher than International average  
 ● Percent lower than International average

Figure 7. Sample item in reasoning cognitive domain (TIMSS-2011 Report, 2012)

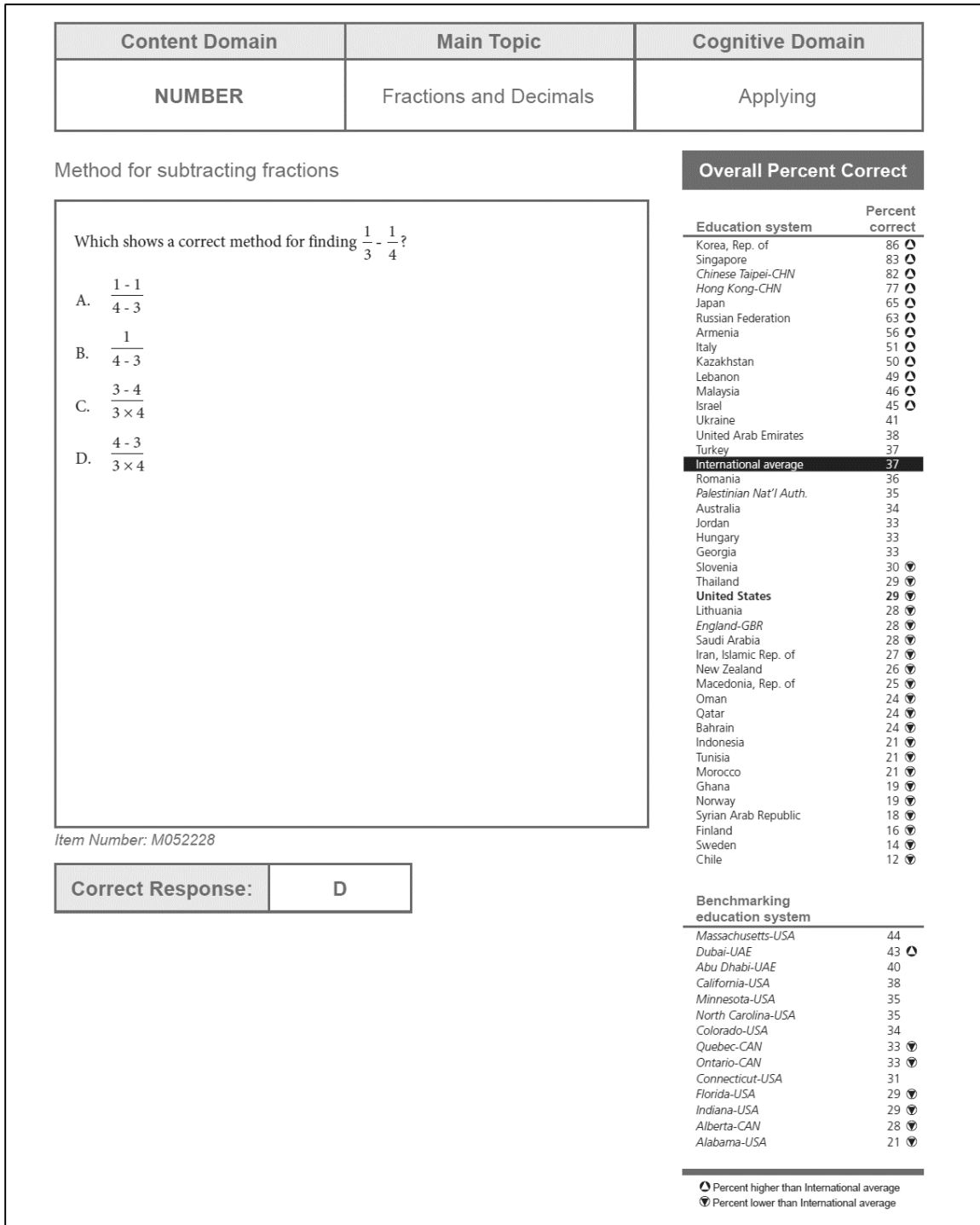


Figure 8. Sample item in applying cognitive domain (TIMSS-2011 Report, 2012)

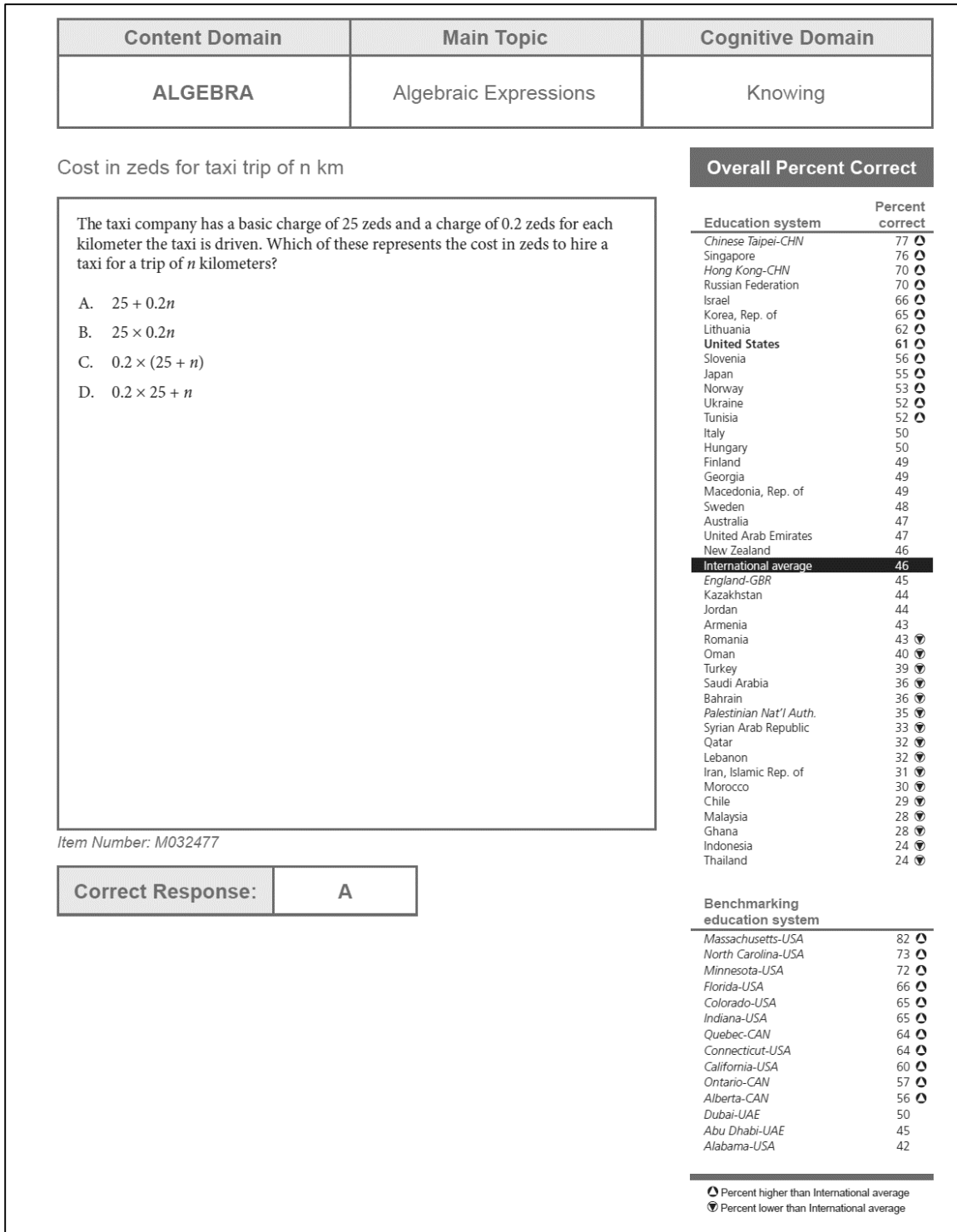


Figure 9. Sample item in knowing cognitive domain (TIMSS-2011 Report, 2012)

The test had two main domains which were content and cognitive areas. The content domain consisted of number, algebra, geometry, data, and chance. The cognitive domain, on the other hand, covered knowing, applying, and reasoning areas. Such differences in domains and areas tested might also lead different classes.

### **Procedure**

An institutional review board (IRB) application was submitted prior to the study (see Appendix A). Since the data were available online to the public, the IRB committee decided that this project was exempt on February 4, 2016 (see Appendix B).

TIMSS 2011 is the fifth stage of the series of international studies done by International Association for the Evaluation of Educational Achievement (IEA). IEA strongly emphasizes that it is important to work with participating countries on a one-on-one basis so any concern or question that arises can be solved quickly. Each participating country assigns a National Research Coordinator to work collaboratively with the organizers of the exam. Participating countries tested the items on a sample of students and submitted the results to the TIMSS-2011 Science and Mathematics Item Review Committee of subject area experts (TIMSS-2011 Technical Report, 20012). Once items were approved the exam was administered. After conducting the test in each participating country, data were released on the official website of TIMSS. Data used in this study were taken from the internet release of item statistics and item responses (<http://timss.bc.edu/timss2011/international-database.html>).



## Analysis

**Research question one.** Does analysis of TIMSS-2011 8th-grade mathematics data from four proposed nations yield multiple latent classes using LCA techniques?

The three booklets from the TIMSS-2011 8th-grade mathematics data were used to run the LCA method described in the previous chapter via *Mplus* Version 7.11 (Muthén & Muthén, 2012a). To find the optimal number of classes, a  $k$ -class model was compared to a  $(k-1)$ -class model by increasing the class number. The Lo-Mendell-Rubin's Likelihood Ratio Test (LMR LR) test was used to compare models where a significant  $p$ -value shows that a  $k$ -class model fits better than the  $(k-1)$ -class model (Wang & Wang, 2012). The LMR LR test is used iteratively until finding a non-significant  $p$ -value between a  $(k+1)$ -class model and a  $k$ -class model which shows the  $(k-1)$ -class model was the optimal number of latent classes. The bootstrap likelihood ratio test (BLRT)  $p$ -value was also calculated from the log-likelihood differences in bootstrap samples from both  $k$ -class and  $(k-1)$ -class models. Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) were checked to see if a best fit model could be clearly identified; the lowest value of these criteria amongst competing LCA models is considered as justification for determination (Samuelsen & Dayton, 2010). For classification purposes, estimated posterior probabilities are used to determine if individuals were assigned into a latent class based on their highest posterior probability value (e.g., Nagin's (2005) criterion for minimum acceptable class membership

classification is exceeded when the average posterior probability was at least 0.70 for all groups).

Finally, each latent class should be defined in a clear and interpretable way such that the differences in the population is described clearly (Wang & Wang, 2012). The best fitting model was also calculated with different sets of random starting values until the best log-likelihood value was the most frequent solution to provide evidence that the global maximum was reached (Samuelsen & Dayton, 2010; Wang & Wang, 2012). If, for any reason, an acceptable model was not found, implications for the validation of TIMSS-2011 would be discussed.

**Research question two.** Does analysis of TIMSS-2011 8th-grade mathematics data from four proposed nations yield multiple latent classes using the MRM techniques?

The three booklets were used to run the mixture Rasch model analysis using WINMIRA (von Davier, 2001a). Since the data were sparse, competing models were selected by means of information criterion values which were the Pearson Chi-square value and Cressie-Read statistic (Cressie & Read, 1984). (With the Cressie-Read statistics, the number of parameters is included in the model as a penalty term for over parameterization (Kang & Cohen, 2007)). WINMIRA calculates the information indices using conditional likelihood estimation (von Davier, 2001b). Information criteria used in this study were the Pearson Chi-square value and Cressie-Read where larger values show better fit.

Once the latent classes were identified, item fit was examined. Item fit statistics are handled slightly differently in the MRM than in a simple Rasch analysis. Each possible latent class yields its own Rasch analysis and its own set of item and person position estimates and fit statistics, along with point-scale indicators (comparable to discrimination indices). An item that overfits ( $p < .05$ ) does not provide new information about the participants. An item that underfits ( $p > .95$ ) has an item discrimination is lower than it is assumed by Rasch model.

**Research question three.** Do LCA and the MRM analysis results differ in terms of :

- a. Item fit parameters for TIMSS-2011 8th-grade mathematics?
- b. Item class parameters for TIMSS-2011 8th-grade mathematics?

Results from analysis of research question one and research question two were compared in order to see if two methods yield different results in terms of number of classes, item parameters, fit indices, and class weights within the classes identified as the best fitting.

**Research question four.** Are there associations between LCA and the MRM latent classes, nation, and gender for TIMSS-2011 8th-grade mathematics?

A four-way frequency table was constructed using class membership results from research questions one and two using SPSS 22. Class membership from LCA, class membership from the MRM, nation, and gender were used to examine associations. A simple cross-tabulation to check whether the frequencies per each cell were adequate to

allow log-linear analysis to be performed. Once cell frequencies were found adequate, model fit was examined using the chi-square statistic. Models tested were nested models, so chi-square differences were tested as models became more parsimonious. Model testing begins with the saturated model and higher order terms are sequentially removed in a backward stepwise fashion until a model is identified that has adequate fit to the data and is the most parsimonious. A nonsignificant chi square value shows that the model fit the data. Significant partial associations were used to identify variable associations necessary to provide a fitting model.

Additionally, data were normally distributed for all booklets. Values for skewness and kurtosis between -2 and +2 are considered acceptable in order to establish data are normally distributed (George & Mallery, 2010). All items for booklets had skewness and kurtosis values within the acceptable range for normal univariate distributions.

## Chapter Three

### Results

**Research question one.** Does analysis of TIMSS-2011 8th-grade mathematics data from four proposed nations yield multiple latent classes using LCA techniques?

The latent class structure of the TIMSS-2011 8th-grade mathematics data was assessed by exploratory LCA analysis with *Mplus* Version 7.11 (Muthén & Muthén, 2012a) with the three-step modeling approach by Wang and Wang (2012): 1) Find the optimal number of latent classes (use fit indices), 2) evaluate the quality of the classification of latent class membership, and 3) define the latent classes.

#### Number of Latent Classes

The item response data from TIMSS-2011 8th-grade mathematics assessments were used as the input for LCA analysis. Each booklet had a different number of items and numbers of participants. Data were recoded into dichotomous responses to get sufficient values in each cell of the contingency table (Collins & Lanza, 2010). The analyzed latent class model can be seen in Figure 10 for Booklet One; the model for the remaining two booklets was identical except for use of different items (see Appendices D and F). Rectangle shapes show observed variables (items) and circle shapes show error

components (e1, e2, e3, etc.), and “C” is the latent construct. To find the optimal number of classes, the fit of the  $k$ -class model was compared with a series of increasing class number models (see Table 2, Table 6 and Table 10). Out of all solutions, classes with the smallest BIC values were selected since BIC works the best with larger sample sizes (Nylund, Asparouhov, & Muthén, 2007). Additionally LMR LR and BLRT  $p$ -values were calculated for model fit decision purposes. Also, the sample *Mplus* input files for each booklet used in this study can be seen in Appendices C, E, and G.

**Booklet One.** The best model fit with the optimal number of classes was decided by analyzing the fit of a series of increasing class number models by comparing the  $k$ -class model with the  $(k-1)$ -class model (Wang & Wang, 2012).

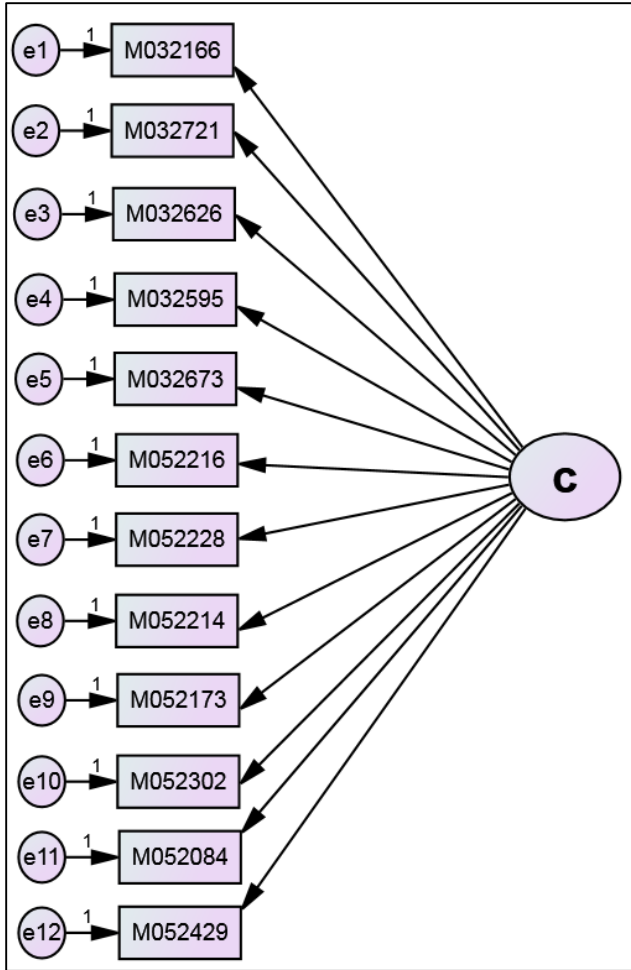


Figure 10. LCA model for Booklet One (Amos Version 22)

The fit statistics and information criterion indices for the models, which ranged from 1 to 4 latent classes, are shown in Table 2. Based on the  $p$ -values of the LMR LR test ( $p = 0.06$ ) and the BLRT test ( $p = 0.07$ ), both were statistically nonsignificant at the 4-class model; hence, the test failed to reject the 3-class model in favor of a four or more class model. Also non-decreasing BIC (22966) of the 4- class model supported evidence for the 3-class model, the non-decreasing AIC (22730) of the 4-class model supported evidence for the 3-class model. Therefore, the fit of the 3-class model was determined to be adequate and the preferred model for further analysis for Booklet One.

Table 2  
*LCA Model Fit Indices for Booklet One*

Model	BIC	AIC	LMR LRT <i>p</i> -value	BLRT <i>p</i> -value
1-class	N/A	N/A	N/A	N/A
2-class	23351	23214	<0.001	<0.001
<b>3-class</b>	<b>22938</b>	<b>22730</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>
4-class	22966	22687	0.06	0.07

*Note.* BIC = the Bayesian information criterion; AIC = Akaike's information criterion; LMR LRT = Lo-Mendell-Rubin Likelihood Ratio Test; BLRT = Bootstrap Likelihood Ratio Test.

**Classification Quality.** Estimated posterior probabilities were used to examine the quality of the classification for the 3-class model for Booklet One. In LCA, membership of the individuals are not determined definitely. However each participant is assigned into the best possible latent class based on their largest posterior probability. Also, the probability of being in the wrong class is low when an individual's highest posterior probability is close to 1.0 (Wang & Wang, 2012). The final class sizes and percentages for the latent classes are given in Table 3. Table 3 shows that, 519 students (29.5%) were assigned to Class 1, 743 students (37.6%) were assigned to Class 2, and 502 students (28.5%) were assigned to Class 3.

Table 3  
*Final Latent Class Size and Percentage for Booklet One*

Classes	Size	Percentage
1	519	29.5 %
2	743	42.1 %
3	502	28.5 %

The average latent class posterior probabilities for the most likely latent class membership are reported in Table 4. The probabilities for most likely latent class



membership for students assigned to the first class was 0.90, while the probability of misclassification was 0.10. Similarly, for students assigned to the second class, the probability of correct class membership was 0.86, while the probability of misclassification was 0.14; for students assigned to the third class, the probability of correct class membership was 0.89, while the probability of misclassification was 0.11. According to Nagin (2005), average latent class probabilities for most likely latent class membership should be 0.70 or above which in this case meets his criterion for all groups.

Table 4  
*Average Latent Class Probabilities for Most Likely Latent Class Membership for Booklet One*

Classes	Probability of Class 1 Membership	Probability of Class 2 Membership	Probability of Class 3 Membership
1	0.90	0.10	0.00
2	0.06	0.86	0.07
3	0.00	0.11	0.89

Another criterion used in this paper is the entropy statistic. Clark (2010) states that medium entropy values (.between .60 and .80) support classification correctness. For Booklet One, entropy was .74 which shows that latent class membership classification quality was adequate enough for the 3-class model.

**Definition of Latent Classes.** The differences in the sample population was explored by analysis of the estimated item-response probability of endorsing “Correct Response” for each of the 12 items. The three latent classes—highly skilled students, moderately skilled students, and somewhat skilled students —were labeled by the researcher based on the observed pattern of item response probabilities. The highly skilled students class, denoted as Class 1 consisting of 519 students, had the highest item-response probabilities for each

of the 12 items. Class 2, which contained 743 students with the second highest item-response probabilities for each of the 12 items, as moderately skilled students; Class 3 was defined as somewhat skilled students, which contained 502 students and had the lowest item-response probabilities for each of the 12 items. The unconditional latent class probabilities and the conditional probabilities for endorsing “Correct Answer” are reported by latent class in Table 5. Conditional probability profiles for endorsing the “Correct answer” for the 3-Class model are tabulated in Figure 11.

Table 5  
*Three-Class Latent Class Membership for Booklet One*

Item	Probability of Class 1	Probability of Class 2	Probability of Class 3
		Unconditional	
	0.29	0.42	0.29
		Conditional “Correct Answer”	
M032166	0.95	0.91	0.41
M032721	0.68	0.37	0.31
M032626	0.94	0.55	0.31
M032595	0.96	0.78	0.33
M032673	0.91	0.66	0.24
M052216	0.99	0.87	0.41
M052228	0.93	0.30	0.14
M052214	0.74	0.39	0.28
M052173	0.63	0.05	0.10
M052302	0.99	0.92	0.52
M052084	0.95	0.64	0.21
M052429	0.94	0.75	0.27

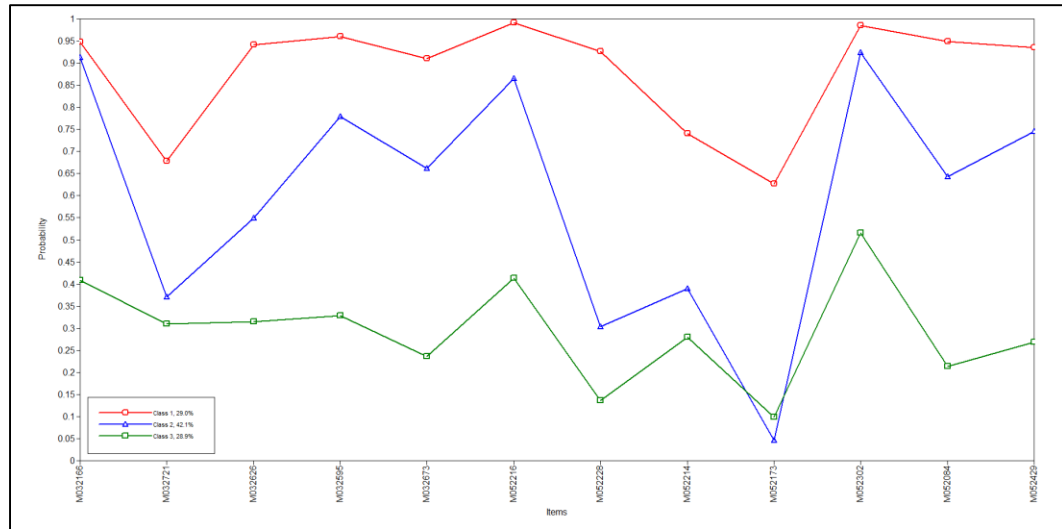


Figure 11. Conditional Probability Profiles of Endorsing “Correct Answer” for 3-Class LCA Model for Booklet One (Mplus Version 7.11)

**Booklet Four.** The best model fit with the optimal number of classes was decided by analyzing the fit of a series of increasing class number models by comparing the  $k$ -class model with the  $(k-1)$ -class model for Booklet Four (Wang & Wang, 2012). The fit statistics and information criterion indices for the models, which ranged from 1 to 4 latent classes, are shown in Table 6. Based on the  $p$ -values of the LMR LR test ( $p = 0.29$ ) and the BLRT test ( $p = 0.14$ ), both were statistically nonsignificant at the 4-class model; hence, the test failed to reject the 3-class model in favor of a four or more class model. Also non-decreasing BIC (21392) of the 4-class model supported evidence for the 3-class model, the non-decreasing AIC (21207) of the 4-class model supported evidence for the 3-class model. Hence, the fit of the 3-class model was decided to be adequate and the selected model for further analysis for Booklet Four.

Table 6  
*LCA Model Fit Indices for Booklet Four*

Model	BIC	AIC	LMR LRT <i>p</i> -value	BLRT <i>p</i> -value
1-class	N/A	N/A	N/A	N/A
2-class	21371	21256	<0.001	<0.001
<b>3-class</b>	<b>21332</b>	<b>21157</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>
4-class	21392	21207	0.29	0.14

*Note.* BIC = the Bayesian information criterion; AIC = Akaike's information criterion; LMR LRT = Lo-Mendell-Rubin Likelihood Ratio Test; BLRT = Bootstrap Likelihood Ratio Test.

**Classification Quality.** Estimated posterior probabilities were used to measure the quality of the classification for the 3-class model for Booklet Four. The final class sizes and percentages for the latent classes are given in Table 7. Table 7 shows that 473 students (27.1%) were assigned to Class 1, 694 students (39.0%) were assigned to Class 2, and 579 students (33.9%) were assigned to Class 3.

Table 7  
*Final Latent Class Size and Percentage for Booklet Four*

Classes	Size	Percentage
1	473	27.1 %
2	694	39.0 %
3	579	33.9 %

The average latent class posterior probabilities for the most likely latent class membership are reported in Table 8. The probability for most likely latent class membership for students assigned to the first class was 0.87, while the probability of misclassification was 0.13. Similarly, for students assigned to the second class, the probability of correct class membership was 0.76, while the probability of misclassification was 0.24; for students assigned to the third class, the probability of correct class membership was 0.80, while the probability of misclassification was 0.20.

All average latent class probabilities for most likely latent class membership exceeded 0.70 which in this case meets Nagin’s (2005) criterion for all groups.

Table 8  
*Average Latent Class Probabilities for Most Likely Latent Class Membership for Booklet Four*

Classes	Probability of Class 1 Membership	Probability of Class 2 Membership	Probability of Class 3 Membership
1	0.87	0.13	0.00
2	0.09	0.76	0.15
3	0.00	0.20	0.80

Clark (2010) states that medium entropy values (.between .60 and .80) support classification correctness. For Booklet Four, entropy was .69 which show that latent class membership classification quality was adequate enough for the 3-class model.

**Definition of Latent Classes.** The differences in the sample population were explored by analysis of the estimated item-response probability of endorsing “Correct Response” for each of the 10 items. The three latent classes—highly skilled students, moderately skilled students, and somewhat skilled students —were labeled by the researcher based on the observed pattern of item response probabilities. The highly skilled students class, denoted as Class 1 consisting of 473 students, had the highest item-response probabilities for each of the 10 items. Class 2, which contained 694 students with the second highest item-response probabilities for each of the 10 items, as moderately skilled students; Class 3 was defined as somewhat skilled students, which contained 579 students and had the lowest item-response probabilities for each of the 10 items. The unconditional latent class probabilities and the conditional probabilities for endorsing “Correct Answer” are

reported by latent class in Table 9. Conditional probability profiles for endorsing the “Correct answer” for the 3-Class model are shown in Figure 12.

Table 9  
*Three-Class Latent Class Membership for Booklet Four*

Item	Probability of Class 1	Probability of Class 2	Probability of Class 3
	Unconditional		
	0.27	0.40	0.33
	Conditional “Correct Answer”		
M032094	0.99	0.74	0.38
M032662	0.69	0.15	0.11
M032419	0.87	0.59	0.30
M032477	0.98	0.60	0.25
M032324	0.76	0.32	0.19
M032116	0.88	0.52	0.29
M032100	0.89	0.69	0.34
M032402	0.90	0.62	0.40
M032397	0.84	0.70	0.33
M032132	0.85	0.65	0.36

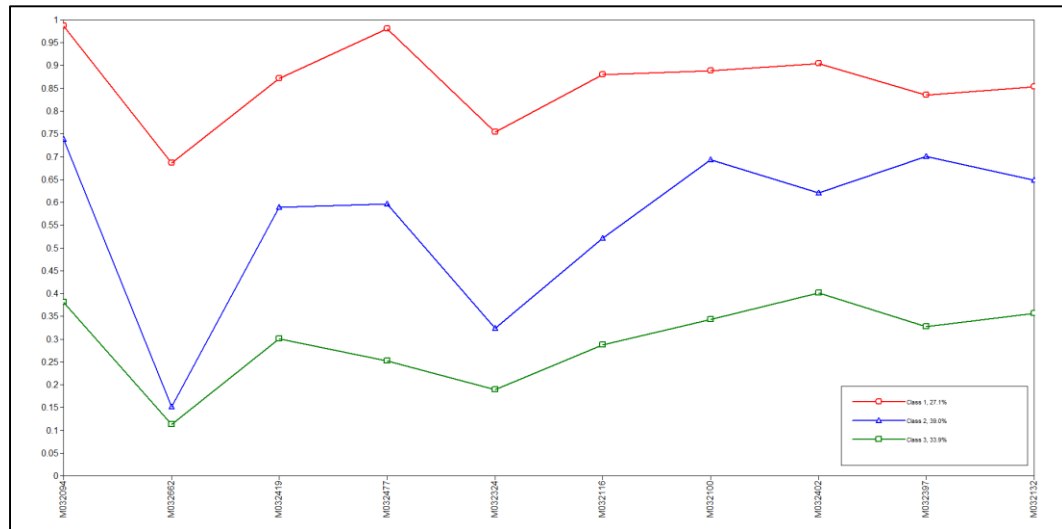


Figure 12. Conditional Probability Profiles of Endorsing “Correct Answer” for 3-Class LCA Model for Booklet Four (Mplus Version 7.11).

**Booklet Six.** The best model fit with the optimal number of classes was examined by analyzing the fit of a series of increasing class number models by comparing the k-class model with the (k-1)-class model for Booklet Four (Wang & Wang, 2012). The fit statistics and information criterion indices for the models, which ranged from 1 to 4 latent classes, are shown in Table 10. Based on the *p*-values of the LMR LR test (*p* = 0.09) and the BLRT test (*p* = 0.07), both were statistically non-significant at the 3-class model; hence, the test failed to reject the 2-class model in favor of a three or more class model. Also non-decreasing BIC (33827) of the 3-class model supported evidence for the 2-class model, the non-decreasing AIC (33522) of the 3-class model supported evidence for the 2-class model. Therefore, the fit of the 2-class model was decided to be adequate and the selected model for further analysis for Booklet Six.

Table 10  
*LCA Model Fit Indices for Booklet Six*

Model	BIC	AIC	LMR LRT <i>p</i> -value	BLRT <i>p</i> -value
1-class	N/A	N/A	N/A	N/A
<b>2-class</b>	<b>34491</b>	<b>34290</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>
3-class	33827	33522	0.09	0.07
4-class	33803	33395	0.16	0.11

*Note.* BIC = the Bayesian information criterion; AIC = Akaike's information criterion; LMR LRT = Lo-Mendell-Rubin Likelihood Ratio Test; BLRT = Bootstrap Likelihood Ratio Test.

**Classification Quality.** Estimated posterior probabilities were used to measure the quality of the classification for the 2-class model for Booklet Four. The final class sizes and percentages for the latent classes are given in Table 11. Table 11 shows that,

813 students (47.9%) were assigned to Class 1, and 889 students (52.1%) were assigned to Class 2.

Table 11  
*Final Latent Class Size and Percentage for Booklet Six*

Classes	Size	Percentage
1	813	47.9 %
2	889	52.1 %

The average latent class posterior probabilities for the most likely latent class membership are reported in Table 12. The probabilities for most likely latent class membership for students assigned to the first class was 0.96, while the probability of misclassification was 0.04. Similarly, for students assigned to the second class, the probability of correct class membership was 0.96, while the probability of misclassification was 0.04. Most likely latent class membership was 0.70 or above for all groups.

Table 12  
*Average Latent Class Probabilities for Most Likely Latent Class Membership for Booklet Six*

Classes	Probability of Class 1 Membership	Probability of Class 2 Membership
1	0.96	0.04
2	0.04	0.96

For Booklet Six, entropy was .86 which shows that latent class membership classification quality was adequate enough for the 2-class model.

**Definition of Latent Classes.** The differences in the sample population was explored by analysis of the estimated item-response probability of endorsing “Correct Response” for



each of the 18 items. The two latent classes—highly skilled students, and moderately skilled students—were labeled by the researcher based on the observed pattern of item response probabilities. The highly skilled students class, denoted as Class 1 consisting of 813 students, had the highest item-response probabilities for each of the 18 items. Class 2, which contained 889 students with the lower item-response probabilities for each of the 18 items, as moderately skilled students. The unconditional latent class probabilities and the conditional probabilities for endorsing “Correct Answer” are reported by latent class in Table 13. Conditional probability profiles for endorsing the “Correct answer” for the 2-Class model are shown in Figure 13.

Table 13  
*Two-Class Latent Class Membership for Booklet Six*

Item	Probability of Class 1	Probability of Class 2
	Unconditional	
	0.48	0.52
	Conditional “Correct Answer”	
M042041	0.97	0.61
M042024	0.95	0.45
M042016	0.77	0.40
M042077	0.89	0.36
M042235	0.95	0.40
M042067	0.68	0.28
M042150	0.65	0.34
M042260	0.91	0.71
M032352	0.92	0.48
M032738	0.96	0.57
M032295	0.99	0.65
M032331	0.54	0.17
M032623	0.77	0.17
M032679	0.81	0.35
M032047	0.67	0.41
M032398	0.72	0.36
M032507	0.68	0.20
M032424	0.84	0.37

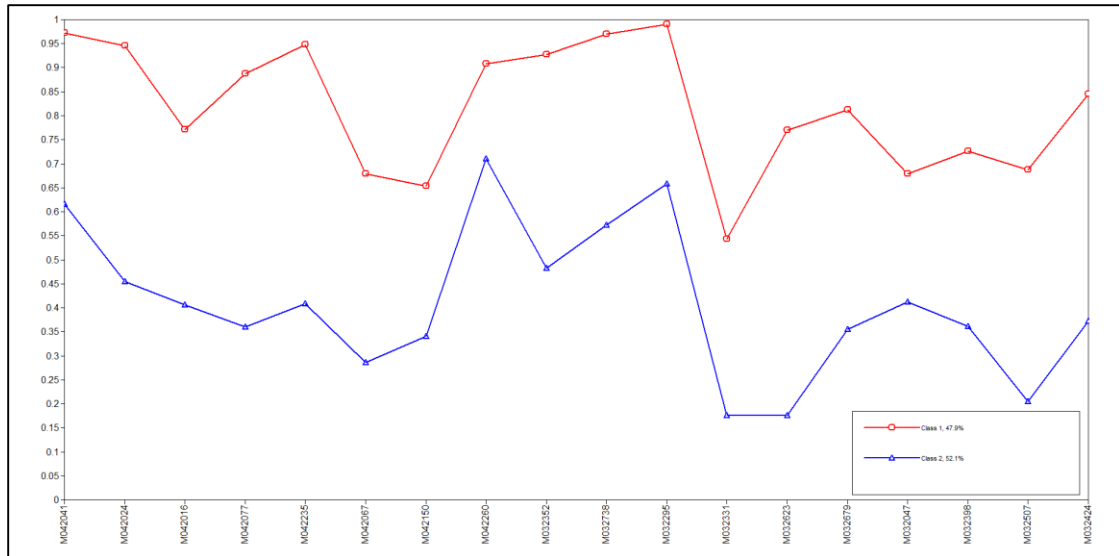


Figure 13. Conditional Probability Profiles of Endorsing “Correct Answer” for 2-Class LCA Model for Booklet Six (Mplus Version 7.11).

**Research question two.** Does analysis of TIMSS-2011 8th-grade mathematics data from four proposed nations yield multiple latent classes using the MRM techniques?

### Number of latent classes

To find the appropriate number of latent classes, competing models with one, two, three, and four latent classes were fit to the data for three booklets. Table 14 shows *p*-values of the Pearson Chi-square and Cressie-Read for the four models. [It should be noted that there was agreement between the Pearson and Cressie-Read values in identifying the number of latent classes for all booklets.] These fit indices were employed due to data being sparse (von Davier, 2001b). Table 14 suggests that there were different numbers of latent classes for different booklets based on larger values of the Pearson Chi-

square and Cressie-Read. Therefore, the models with higher values of the Pearson Chi-square and Cressie-Read were selected for each booklet.

Table 14  
*p-values of Model Fit Indices for the MRM*

Model	Booklet One		Booklet Four		Booklet Six	
	<u>Cressie</u> <u>Read</u>	<u>Pearson</u> <u>X<sup>2</sup></u>	<u>Cressie</u> <u>Read</u>	<u>Pearson</u> <u>X<sup>2</sup></u>	<u>Cressie</u> <u>Read</u>	<u>Pearson</u> <u>X<sup>2</sup></u>
1-Class	0.00	0.00	0.00	0.00	0.00	0.05
2-Class	0.10	0.48	<b>0.13</b>	<b>0.15</b>	<b>0.48</b>	<b>0.90</b>
3-Class	<b>0.20</b>	<b>0.58</b>	0.00	0.15	0.45	0.80
4-Class	0.00	0.18	0.10	0.03	0.08	0.68

Results showed that the mean of the raw scores of class 1 was high (M=9.02 SD=2.35), class 2 was medium (M=7.18, SD=2.26), and class 3 was low for Booklet One (M=3.42, SD=1.61). Also, for booklet three, class 1 was low (M=3.97, SD=1.77), class 2 was high (M=6.39, SD=1.32), for Booklet Six class 1 was low (M=13.42, SD=3.16), class 2 was high (M=6.66, SD=2.63). Comparing item parameters across different classes is critical when deciding on number of classes. This procedure provides critical information about the qualitative differences within the latent classes. This comparison supplies information about item difficulties where the researcher can focus on the items that are relatively more difficult in one class compared to other ones (Baghaei & Carstensen, 2013). These MRMs are closely related to latent class analysis. The following paragraphs focus on the latent class and item parameter results for each booklet used in this study.

**Booklet One.** The dataset consisted of 12 items with 1764 participants. To determine the appropriate number of classes, one, two, three, and four latent class solutions were fit to

the data. Table 14 provides  $p$ -values of Cressie-Read and Pearson Chi-square statistics.  $P$ -values for Booklet One of Cressie-Read and Pearson Chi-square were .20 and .58. Since the three class model had the highest  $p$ -value, a three-class solution was selected for Booklet One. Class size values for each class shows that class 1 was expected to include about 42% of the sample. Class 2 was expected to include about 36% of the sample. Class 3 was expected to include 22% of the sample in the data set. According to the Q-index, there was no need to remove any items since all of the items fit each class well ( $.05 < p < .95$ ) (See Table 15).

Table 15  
*Item fit assessed by the Q-index for all classes of Booklet One*

Item	Class -1			Class -2			Class -3		
	Q-index	Zq	$p$ (X>Zq)	Q-index	Zq	$p$ (X>Zq)	Q-index	Zq	$p$ (X>Zq)
M032166	0.25	0.78	0.21	0.21	0.66	0.25	0.24	-0.89	0.81
M032721	0.14	0.84	0.20	0.19	1.29	0.10	0.18	-0.44	0.67
M032626	0.11	-0.28	0.62	0.17	0.17	0.43	0.24	0.44	0.33
M032595	0.13	-0.18	0.57	0.15	-0.32	0.63	0.25	-0.01	0.50
M032673	0.14	0.30	0.38	0.13	-0.66	0.75	0.24	0.00	0.49
M052216	0.21	0.06	0.48	0.16	0.30	0.38	0.27	0.21	0.41
M052228	0.09	-0.56	0.71	0.12	-0.99	0.84	0.21	-0.17	0.57
M052214	0.50	-0.61	0.73	0.12	0.50	0.20	0.27	0.65	0.25
M052173	0.06	-0.93	0.82	0.13	-1.22	0.89	0.22	0.19	0.42
M052302	0.23	0.23	0.40	0.18	0.44	0.33	0.31	-0.07	0.53
M052084	0.12	-0.22	0.59	0.15	-0.20	0.58	0.26	0.17	0.43
M052429	0.17	0.59	0.27	0.16	0.02	0.49	0.29	0.09	0.46

Figure 14 shows that the three classes had different item difficulty parameters. The lines display items on which the three classes seem to converge and also to diverge. Item difficulty estimates were substantially different for the majority of items. It can be concluded that all classes found the items to be relatively easy as logit position was generally negative (see Table 16 for specific values including standard error).

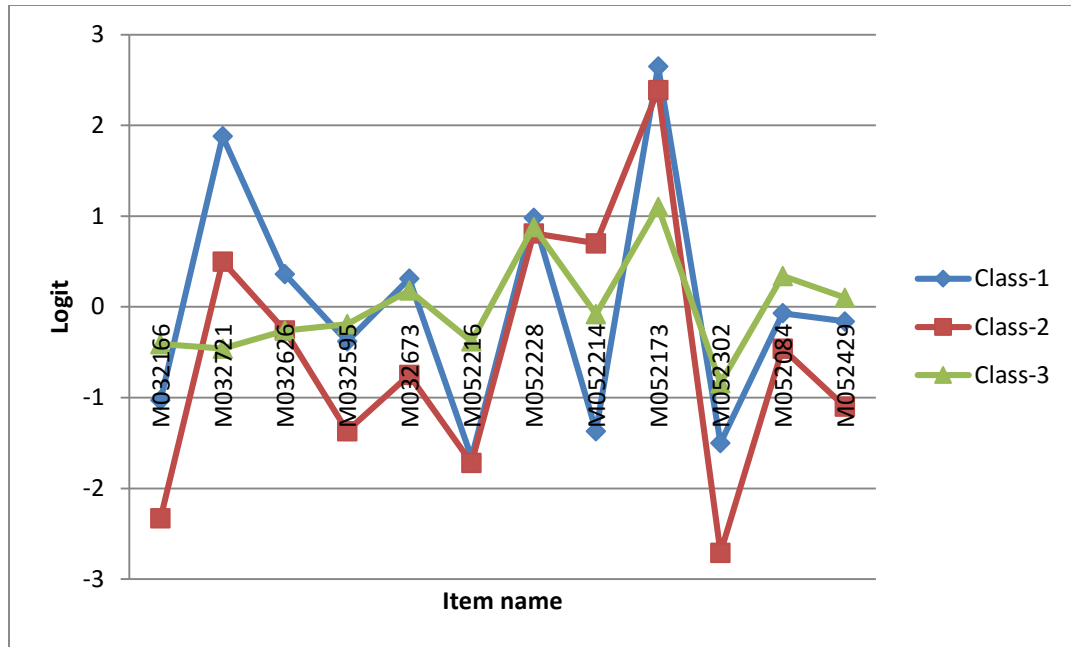


Figure 14. Class specific item parameter profiles for Booklet One.

Table 16  
Item parameters of Booklet One by classes

Item	Class-1		Class-2		Class-3	
	Estimate	Error	Estimate	Error	Estimate	Error
M032166	-1.03	0.13	-2.33	0.14	-0.41	0.11
M032721	1.88	0.10	0.50	0.10	-0.46	0.11
M032626	0.36	0.10	-0.26	0.10	-0.26	0.11
M032595	-0.38	0.11	-1.37	0.11	-0.19	0.12
M032673	0.31	0.10	-0.75	0.10	0.18	0.12
M052216	-1.66	0.16	-1.72	0.12	-0.38	0.11
M052228	0.98	0.09	0.81	0.10	0.88	0.15
M052214	-1.37	0.14	0.70	0.10	-0.08	0.11
M052173	2.65	0.10	2.39	0.14	1.10	0.16
M052302	-1.50	0.15	-2.71	0.16	-0.84	0.10
M052084	-0.07	0.11	-0.46	0.09	0.34	0.13
M052429	-0.16	0.11	-1.10	0.11	0.10	0.12

**Booklet Four.** The dataset consisted of 10 items with 1746 participants. To determine the appropriate number of classes, one, two, three, and four latent class solutions were fit

to the data. Table 14 provides  $p$ -values of Cressie-Read and Pearson Chi-square statistics.  $P$ -values for Booklet Four of Cressie-Read and Pearson Chi-square were .13 and .15. Since the two class model had the highest  $p$ -value, a two-class solution was selected for Booklet Four. It is important to note that since fit index values were close, as a general rule a more parsimonious model was selected. Class size values for each class presents that class 1 was expected to include about 66% of the sample. Class 2 was expected to include about 34% of the sample. The class sizes indicate that about 66 percent and 34 percent of the sample can be fitted by a mixed Rasch model which was assumed to hold in these classes. According to the  $Q$ -index, there was one item (M032662) with a  $Z_q$  value of 2.37 and  $p$ -value of .01 which shows lower discrimination in class one. In such cases, item removal is suggested from the scale only after examining the items content and additional information from the estimated model (von Davier, 2001b). Item category values for this item were acceptable. Out of 1,746 responses 1,251 students answered the item false and 495 students answered correct. Additionally, the item parameter value for class one was also acceptable with a value of .13. After examining the item category values and item fit, it is decided not to remove the item from analysis. All of the other items fit each class well ( $.05 < p < .95$ ) (See Table 17).

Table 17

*Item fit assessed by the Q-index for all classes of Booklet Four.*

Item	Class -1			Class -2		
	Q-index	Zq	$P$ ( $X > Zq$ )	Q-index	Zq	$P$ ( $X > Zq$ )
M032094	0.25	-0.04	0.51	0.38	-0.11	0.54
<b>M032662</b>	<b>0.32</b>	<b>2.37</b>	<b>0.01</b>	0.17	0.08	0.47
M032419	0.24	-0.58	0.71	0.12	0.02	0.49
M032477	0.27	0.83	0.20	0.39	-0.54	0.71
M032324	0.25	0.30	0.37	0.16	0.09	0.46
M032116	0.25	0.33	0.37	0.15	-0.01	0.51
M032100	0.22	-0.95	0.83	0.17	0.30	0.38
M032402	0.27	0.80	0.21	0.14	-0.01	0.50
M032397	0.21	-1.45	0.92	0.17	0.00	0.50
M032132	0.25	-0.26	0.60	0.18	0.14	0.44

Figure 15 shows that the two classes had similar item difficulty parameters for the first six items and different item difficulty parameters for the last four items. These four items were slightly easier for first class then for the second class. The lines display items on which the two classes seem to diverge and later to converge. The majority of items were not markedly different in difficulty across classes. In general all classes found the items to be relatively easy as logit position was generally negative (see Table 18 for specific values including standard error).

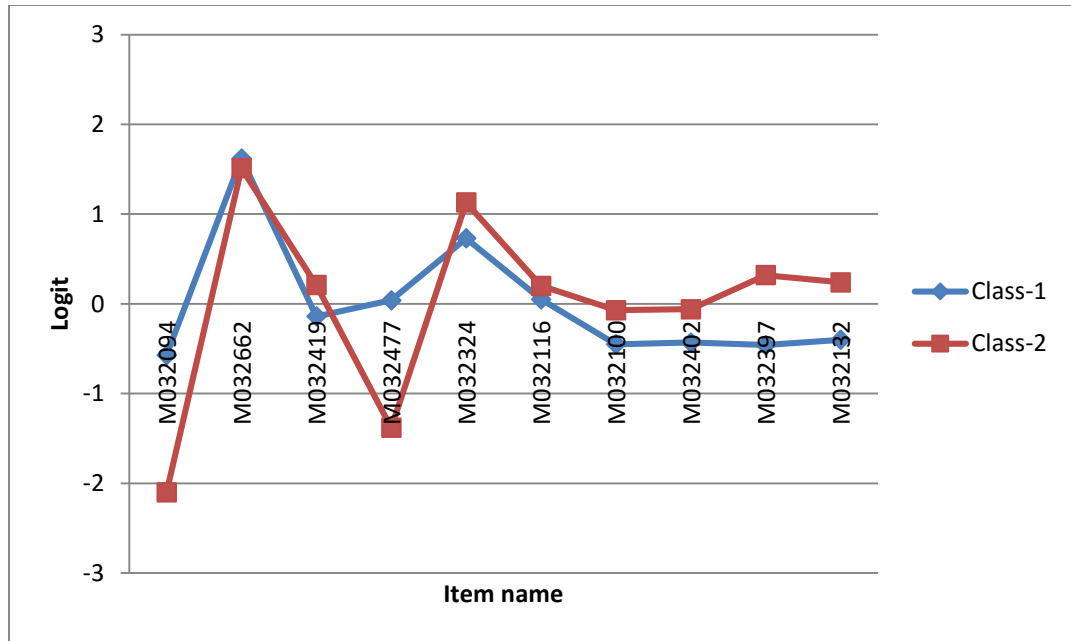


Figure 15. Class specific item parameter profiles for Booklet Four.

Table 18  
Item parameters of Booklet Four by classes

Item	Class-1		Class-2	
	Estimate	Error	Estimate	Error
M032094	-0.57	0.06	-2.10	0.29
M032662	1.62	0.09	1.51	0.09
M032419	-0.14	0.07	0.21	0.12
M032477	0.04	0.07	-1.38	0.21
M032324	0.73	0.07	1.13	0.10
M032116	0.05	0.07	0.20	0.12
M032100	-0.45	0.06	-0.07	0.13
M032402	-0.43	0.06	-0.06	0.13
M032397	-0.46	0.06	0.32	0.11
M032132	-0.40	0.06	0.24	0.12

**Booklet Six.** The dataset consisted of 18 items with 1701 participants. To determine the appropriate number of classes, one, two, three, and four latent class solutions were fit to the data. Table 14 provides  $p$ -values of Cressie-Read and Pearson Chi-square statistics. P-



values for Booklet One of Cressie-Read and Pearson Chi-square were .48 and .90. Since the two class model had the highest  $p$ -value, a two-class solution was fitted for Booklet Six. Class size values for each class presents that class 1 was expected to include about 62% of the sample. Class 2 was expected to include about 38% of the sample. The class sizes indicate that about 62 percent and 38 percent of the sample can be fit by a mixed Rasch model which was assumed to hold in these classes. According to the Q-index, there were two items showing lower discrimination values as follows: M042077 with a  $Z_q$  value of 1.99 and  $p$ -value of .02 and M042067 with a  $Z_q$  value of 2.05 and  $p$ -value of .02 in class two. Based on von Davier (2001b), items were examined and it was decided that there was no need for removal of both of the items. Item category values for both items were acceptable. For item M042077, out of 1,701 responses 659 students answered the item false and 1,042 students answered correct. For item M042067, out of 1,701 responses 894 students answered the item false and 807 students answered correct. Additionally, item parameter values for class two were also acceptable and as follows .29 and .31. All of the other items fit each class well ( $.05 < p < .95$ ) (See Table 19).

Table 19  
*Item fit assessed by the Q-index for all classes of Booklet Six.*

Item	Class -1			Class -2		
	Q-index	Zq	$P$ ( $X > Zq$ )	Q-index	Zq	$P$ ( $X > Zq$ )
M042041	0.19	-0.23	0.59	0.26	-1.00	0.84
M042024	0.16	-0.31	0.62	0.29	-0.23	0.59
M042016	0.21	1.46	0.07	0.27	-0.64	0.74
<b>M042077</b>	<b>0.15</b>	<b>-0.43</b>	<b>0.67</b>	<b>0.37</b>	<b>1.99</b>	<b>0.02</b>
M042235	0.24	0.22	0.41	0.31	-0.40	0.66
<b>M042067</b>	<b>0.13</b>	<b>-1.06</b>	<b>0.85</b>	<b>0.36</b>	<b>2.06</b>	<b>0.02</b>
M042150	0.19	0.94	0.17	0.30	0.13	0.45
M042260	0.20	0.64	0.26	0.30	-0.19	0.57
M032352	0.18	0.02	0.49	0.30	0.02	0.49
M032738	0.24	0.25	0.40	0.24	-1.40	0.92
M032295	0.46	0.11	0.46	0.29	-0.41	0.65
M032331	0.14	-0.55	0.71	0.34	1.59	0.06
M032623	0.13	-1.16	0.88	0.32	0.49	0.31
M032679	0.16	-0.31	0.62	0.34	0.94	0.17
M032047	0.17	0.36	0.36	0.26	-0.40	0.65
M032398	0.16	-0.32	0.63	0.27	-40.00	0.66
M032507	0.15	-0.44	0.67	0.30	0.34	0.37
M032424	0.20	0.96	0.17	0.26	-0.86	0.81

In general, the two classes had similar item difficulty parameters for the items (see Figure 16). Specifically, items M042041, M042024, M042235, M042067, M032738, M032623, M032679, M032507, and M032424 were similar in difficulty level for both classes. On the other hand, items M042016, M042077, M042150, M042260, M032352, M032295, M032331, M032047, and M032398 showed different difficulty levels. In general all classes found the items to be relatively hard as logit position was generally positive (see Table 20 for specific values including standard error).

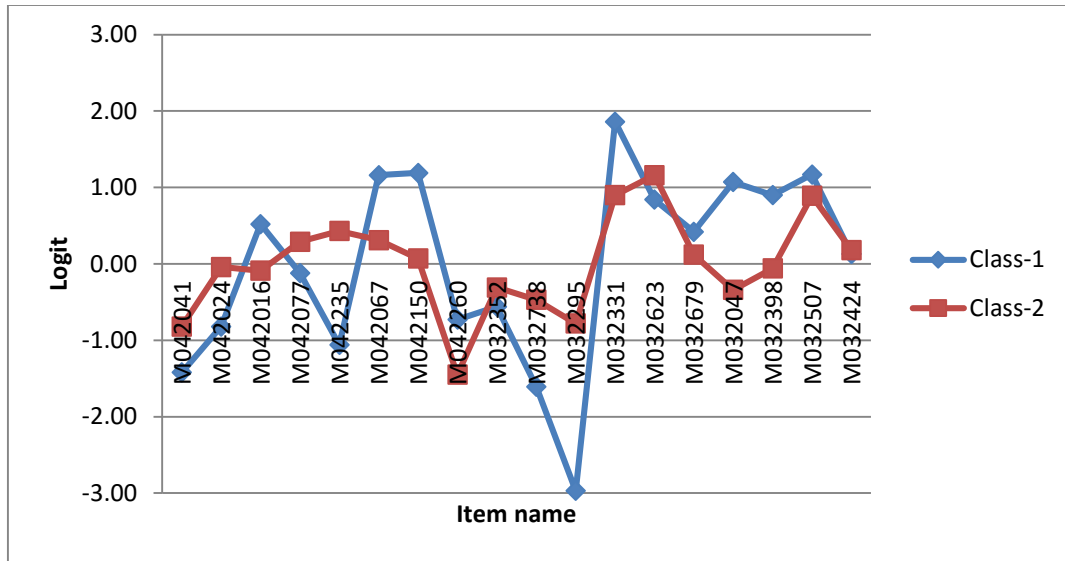


Figure 16. Class specific item parameter profiles for Booklet Six.

Table 20  
Item parameters of Booklet Six by classes

Item	Class-1		Class-2	
	Estimate	Error	Estimate	Error
M042041	-1.42	0.13	-0.82	0.08
M042024	-0.82	0.10	-0.04	0.09
M042016	0.52	0.08	-0.09	0.09
M042077	-0.12	0.09	0.29	0.09
M042235	-1.06	0.11	0.43	0.09
M042067	1.16	0.07	0.31	0.09
M042150	1.19	0.10	0.07	0.09
M042260	-0.73	0.09	-1.45	0.09
M032352	-0.56	0.14	-0.31	0.08
M032738	-1.61	0.24	-0.47	0.08
M032295	-2.97	0.07	-0.78	0.08
M032331	1.86	0.07	0.90	0.10
M032623	0.84	0.08	1.16	0.11
M032679	0.42	0.07	0.12	0.09
M032047	1.07	0.07	-0.34	0.08
M032398	0.90	0.07	-0.06	0.09
M032507	1.17	0.07	0.89	0.10
M032424	0.13	0.08	0.18	0.09

**Research question three.** Do LCA and the MRM analysis results differ in terms of :

- a. Item fit parameters for TIMSS-2011 8th-grade mathematics?
- b. Item class parameters for TIMSS-2011 8th-grade mathematics?

**Booklet One.** Results from LCA analysis and the MRM analysis were compared. Both LCA and the MRM analysis provided a three-class solution for the data. Although item parameters were not comparable, standard errors of the items had similar values for both analyses (see Table 21).

Table 21  
*Item parameter comparisons of Booklet One LCA and MRM by Class*

Item	Class-1 LCA/MRM				Class-2 LCA/MRM				Class-3 LCA/MRM			
	Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error
M032166	2.90	-1.03	0.23	0.13	-2.35	-2.33	0.18	0.14	0.37	-0.41	0.15	0.11
M032721	0.74	1.88	0.12	0.10	0.52	0.50	0.10	0.10	0.80	-0.46	0.10	0.11
M032626	2.77	0.36	0.32	0.10	-0.20	-0.26	0.10	0.10	0.78	-0.26	0.11	0.11
M032595	3.17	-0.38	0.29	0.11	-1.25	-1.37	0.15	0.11	0.71	-0.19	0.12	0.12
M032673	2.32	0.31	0.19	0.10	-0.66	-0.75	0.13	0.10	1.17	0.18	0.13	0.12
M052216	2.69	-1.66	0.18	0.16	-1.86	-1.72	0.18	0.12	0.34	-0.38	0.12	0.11
M052228	2.54	0.98	0.30	0.09	0.83	0.81	0.14	0.10	1.84	0.88	0.15	0.15
M052214	1.05	-1.37	0.13	0.14	0.45	0.70	0.10	0.10	0.94	-0.08	0.11	0.11
M052173	0.52	2.65	0.15	0.10	3.02	2.39	0.29	0.14	2.21	1.10	0.16	0.16
M052302	2.16	-1.50	0.14	0.15	-2.50	-2.71	0.18	0.16	-0.06	-0.84	0.14	0.10
M052084	2.92	-0.07	0.27	0.11	-0.59	-0.46	0.13	0.09	1.30	0.34	0.13	0.13
M052429	2.67	-0.16	0.24	0.11	-1.07	-1.10	0.13	0.11	1.00	0.10	0.13	0.12

Additionally, although the Bayesian information criterion was not used in the MRM for identifying the model fit, both methods produced very close BIC values (see Table 22).

Table 22  
*LCA and MRM 3 Class Model BIC Fit Indices for Booklet One*

Model	LCA	MRM
3-class	22938	22970

*Note.* BIC = the Bayesian information criterion.

Furthermore, two analyses had somewhat different solutions for the class weights. Latent class analysis put the most cases into the middle class. The mixture Rasch model sorted classes based on similarity in their response patterns (see Table 23).

Table 23  
*LCA and MRM 3 Class Model Class Sizes for Booklet One*

Class	LCA	MRM
1	28.0%	42.0%
2	42.0%	36.0%
3	28.0%	22.0%

**Booklet Four.** Results from LCA analysis and the MRM analysis were compared. The LCA model provided a three-class solution. On the other hand, the MRM analysis provided a two-class solution for the data. Since the solutions were based on different number of classes both item parameters and standard errors of the items were not comparable (see Table 24).

Table 24  
*Item parameter comparisons of Booklet Four LCA and MRM by Class*

Item	Class-1 LCA/MRM				Class-2 LCA/MRM				Class-3 LCA/MRM			
	Estimate		Error		Estimate		Error		Estimate		Error	
M032094	-2.26	-0.57	0.30	0.06	-1.03	-2.10	0.23	0.29	0.48	N/A	0.14	N/A
M032662	-0.78	1.62	0.16	0.09	1.72	1.51	0.26	0.09	2.06	N/A	0.18	N/A
M032419	-1.91	-0.14	0.23	0.07	-0.36	0.21	0.13	0.12	0.84	N/A	0.16	N/A
M032477	-2.90	0.04	0.32	0.07	-0.39	-1.38	0.20	0.21	1.08	N/A	0.16	N/A
M032324	-1.11	0.73	0.17	0.07	0.73	1.13	0.15	0.10	1.45	N/A	0.15	N/A
M032116	-1.99	0.05	0.23	0.07	-0.08	0.20	0.15	0.12	0.91	N/A	0.14	N/A
M032100	-2.07	-0.45	0.19	0.06	-0.81	-0.07	0.18	0.13	0.65	N/A	0.15	N/A
M032402	-2.24	-0.43	0.25	0.06	-0.49	-0.06	0.15	0.13	0.39	N/A	0.12	N/A
M032397	-1.62	-0.46	0.15	0.06	-0.85	0.32	0.18	0.11	0.72	N/A	0.16	N/A
M032132	-1.76	-0.40	0.16	0.06	-0.61	0.24	0.17	0.12	0.59	N/A	0.13	N/A

However, for identifying the model fit, both methods produced very close BIC values for different solutions based on different number of classes (see Table 25).

Table 25  
*LCA and MRM 3 vs. 2 Class Model BIC Fit Indices for Booklet Four*

Model	LCA	MRM
2-class	N/A	21301
3-class	21332	N/A

*Note.* BIC = the Bayesian information criterion.

Additionally, two analyses had different solutions for the class weights. Latent class analysis put the most cases into the middle class in a three-class model. The mixture Rasch model sorted classes based on similarity in their response patterns and places most of the cases into the first class in a two-class model (see Table 26).

Table 26  
*LCA and MRM 3 vs. 2 Class Model Class Sizes for Booklet Four*

Class	LCA	MRM
1	27.1%	66.0%
2	39.0%	34.0%
3	33.9%	N/A

**Booklet Six.** Results from LCA analysis and the MRM analysis were compared. Both models supported a two-class solution. Similar to Booklet One, item parameters were not comparable but standard errors of the items had similar values for both analyses (see Table 27).

Table 27

*Item parameter comparisons of Booklet Six LCA and MRM by Class*

Item	Class-1 LCA/MRM				Class-2 LCA/MRM			
	Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error
M042041	-3.53	-1.42	0.27	0.13	-0.47	-0.82	0.08	0.08
M042024	-2.84	-0.82	0.18	0.10	0.18	-0.04	0.08	0.09
M042016	-1.21	0.52	0.09	0.08	0.37	-0.09	0.07	0.09
M042077	-2.06	-0.12	0.15	0.09	0.57	0.29	0.08	0.09
M042235	-2.89	-1.06	0.19	0.11	0.37	0.43	0.09	0.09
M042067	-0.74	1.16	0.10	0.07	0.91	0.31	0.07	0.09
M042150	-0.63	1.19	0.08	0.10	0.66	0.07	0.07	0.09
M042260	-2.29	-0.73	0.13	0.09	-0.89	-1.45	0.08	0.09
M032352	-2.55	-0.56	0.17	0.14	0.06	-0.31	0.08	0.08
M032738	-3.44	-1.61	0.25	0.24	-0.29	-0.47	0.08	0.08
M032295	-4.58	-2.97	0.40	0.07	-0.65	-0.78	0.09	0.08
M032331	-0.17	1.86	0.09	0.07	1.54	0.90	0.09	0.10
M032623	-1.20	0.84	0.12	0.08	1.54	1.16	0.10	0.11
M032679	-1.46	0.42	0.11	0.07	0.59	0.12	0.07	0.09
M032047	-0.75	1.07	0.08	0.07	0.35	-0.34	0.07	0.08
M032398	-0.97	0.90	0.10	0.07	0.57	-0.06	0.07	0.09
M032507	-0.78	1.17	0.09	0.07	1.35	0.89	0.09	0.10
M032424	-1.69	0.13	0.11	0.08	0.52	0.18	0.08	0.09

Although, as data were sparse, the MRM did not use the BIC value for model fit purposes, both methods produced very close BIC values for the two-class model (see Table 28).

Table 28

*LCA and MRM 2 Class Model BIC Fit Indices for Booklet Six*

Model	LCA	MRM
2-class	34491	33709

*Note.* BIC = the Bayesian information criterion.

Once again, the two analyses had different solutions for the class weights. Latent class analysis put the most cases into the second class in a two-class model. The mixture Rasch model sorted classes based on similarity in their response patterns and placed most of the cases into the first class in a two-class model (see Table 29).

Table 29  
*LCA and MRM 2 Class Model Class Sizes for Booklet Six*

Class	LCA	MRM
1	47.9 %	62.0 %
2	52.1 %	38.0 %

**Research question four.** Are there associations between LCA and the MRM latent classes, nation, and gender for TIMSS-2011 8th-grade mathematics?

**Booklet One.** A four way log-linear analysis was performed with variables nation, gender, LCA class membership, and the MRM class membership. Hierarchical, nested models were fitted. In a hierarchical model it is sufficient to list the highest order terms. K-way effects were examined to see the contribution of each level of interaction. The likelihood ratio chi-square with no parameters and only the mean was 2582.89. The value for the first order effect was 2248.50. The difference  $2582.89 - 2248.50 = 334.38$  is displayed on the first line of the table. The difference is a measure of how much the model improved when first order effects were included. The significant  $p$  value ( $< .001$ ) means that the hypothesis of first order effects (main marginals) being zero is rejected. In other words, there was a first order effect. Similar reasoning is then applied to the question of second order effects. The addition of a second order effect improved the



likelihood ratio chi-square by 2218.10. This was also significant. But the addition of a third and a fourth order term did not significantly improve fit ( $p > .05$ ).

Table 30  
*K-Way and Higher-Order Effects for Booklet One*

	K	df	Likelihood Ratio	
			Chi-Square	$p$
K-way Effects	1	8	334.38	<.001
	2	23	2218.10	<.001
	3	28	25.93	.58
	4	12	4.46	.97

Table 31 shows that there were statistically significant associations between nation and LCA class membership ( $p < .05$ ), nation and the MRM class membership ( $p < .05$ ), and LCA class membership and MRM class membership ( $p < .05$ ) for Booklet One. All other interactions between other variables were not statistically significant ( $p > .05$ ).

Table 31  
*Partial Associations for Booklet One*

Effect	df	Partial Chi-Square	Sig.
NATION*ITSEX*LCA	6	6.81	.34
NATION*ITSEX*MRM	6	8.60	.20
NATION*LCA*MRM	12	9.90	.63
ITSEX*LCA*MRM	4	1.07	.90
NATION*ITSEX	3	3.46	.33
<b>NATION*LCA</b>	<b>6</b>	<b>300.25</b>	<b>&lt;.001</b>
ITSEX*LCA	2	2.68	.26
<b>NATION*MRM</b>	<b>6</b>	<b>41.85</b>	<b>&lt;.001</b>
ITSEX*MRM	2	5.67	.06
<b>LCA*MRM</b>	<b>4</b>	<b>1237.21</b>	<b>&lt;.001</b>
NATION	3	181.54	.00
ITSEX	1	.06	.81
LCA	2	59.35	.00
MRM	2	93.44	.00

*Note.* NATION= Countries, ITSEX=Gender, LCA= Latent Class Analysis  
 Group Membership, MRM= Mixed Rasch Model Group Membership

To further analyze the interactions, a custom model was created using two way interactions between nation, LCA class membership, and the MRM class membership variables. In Table 32, the goodness of fit test showed that the model fit the data adequately ( $p > .05$ ).

Table 32  
*Goodness-of-Fit Tests for 2-way Interaction Model for Booklet One*

	Chi-Square	df	<i>p</i>	Adjusted	
				df <sup>a</sup>	<i>p</i>
Likelihood Ratio	43.09	48	<b>.67</b>	40	.34

a. One degree of freedom is subtracted for each cell with an expected value of zero. The unadjusted df is an upper bound on the true df, while the adjusted df may be an underestimate.

**Associations.** A crosstab analysis was run to see the exact membership percentages between interactions. If LCA group membership could be explained by the nation variable, class should be associated with country. Although class membership somewhat reflect countries' success rates in TIMSS as indicated by the latent class, where USA and Finland fell in the table suggests that LCA classes were not a product of the nation variable (see Table 33). However, based on where the nations' academic performance stands, the table supports the idea that latent class identification is based on skill level for Booklet One.

Table 33  
*Crosstabulation of Nation vs. LCA Class Membership for Booklet One*

		LCA GROUP MEMBERSHIP				
			Class 1	Class 2	Class 3	Total
NATION	Turkey	Count	99	91	248	438
		% within NATION	22.6%	20.8%	56.6%	100.0%
	USA	Count	114	392	147	653
		% within NATION	17.5%	60.0%	22.5%	100.0%
	Singapore	Count	285	102	30	417
		% within NATION	68.3%	24.5%	7.2%	100.0%
	Finland	Count	21	158	77	256
		% within NATION	8.2%	61.7%	30.1%	100.0%
Total		Count		743	502	1764
		% within NATION	29.4%	42.1%	28.5%	100.0%

On the other hand, the MRM group membership values more closely paralleled the nation variable but, again, not purely relying on it (see Table 34). Finland and USA provided similar results where Singaporean students were mostly in class 1(65.0 %) and Turkish students were mostly in class 3 (50.7 %).

Table 34

*Crosstabulation of Nation vs. MRM Class Membership for Booklet One*

		MRM GROUP MEMBERSHIP				
		Class 1	Class 2	Class 3	Total	
NATION	Turkey	Count	122	94	222	438
		% within NATION	27.9%	21.5%	50.7%	100.0%
	USA	Count	229	319	105	653
		% within NATION	35.1%	48.9%	16.1%	100.0%
	Singapore	Count	271	121	25	417
		% within NATION	65.0%	29.0%	6.0%	100.0%
	Finland	Count	79	125	52	256
		% within NATION	30.9%	48.8%	20.3%	100.0%
Total		Count		659	404	1764
		% within NATION	39.7%	37.4%	22.9%	100.0%

Furthermore, analysis for Booklet One shows that LCA class memberships and the MRM class memberships overlapped by 70%. For class 1, the agreement level was 74.0 %. For class 2, the agreement level was 60.6 %. For class 3, the agreement level was 80.1 % (see Table 35).

Table 35

*Crosstabulation of LCA Class Membership vs. MRM Class Membership for Booklet One*

		MRM GROUP MEMBERSHIP				
		Class 1	Class 2	Class 3	Total	
LCA GROUP MEMBERSHIP	Class 1	Count	384	135	0	519
		% within LCA GROUP MEMBERSHIP	74.0%	26.0%	0.0%	100.0%
	Class 2	Count	291	450	2	743
		% within LCA GROUP MEMBERSHIP	39.2%	60.6%	0.3%	100.0%
	Class 3	Count	26	74	402	502
		% within LCA GROUP MEMBERSHIP	5.2%	14.7%	80.1%	100.0%
Total		Count	701	659	404	1764
		% within LCA GROUP MEMBERSHIP	39.7%	37.4%	22.9%	100.0%

**Booklet Four.** A four way log-linear analysis was conducted with variables nation, gender, LCA class membership, and the MRM class membership. The likelihood ratio chi-square with no parameters and only the mean was 2326.18. The value for the first order effect was 1897.99. The difference  $2326.18 - 1897.99 = 428.19$  is displayed on the first line of Table 36. The significant  $p$  value ( $< .001$ ) shows that there was a first order effect. The addition of a second order effect improved the likelihood ratio chi-square by 1894.55. This was also significant. But the addition of a third and a fourth order term did not significantly improve fit ( $p > .05$ ).

Table 36  
*K-Way and Higher-Order Effects for Booklet Four*

	K	df	Likelihood Ratio	
			Chi-Square	$P$
K-way Effects	1	7	428.19	<.001
	2	17	1894.55	<.001
	3	17	3.43	1.00
	4	6	0.02	1.00

Table 37 shows that there were statistically significant associations between nation and LCA class membership ( $p < .05$ ), nation and the MRM class membership ( $p < .05$ ), LCA class membership and MRM class membership ( $p < .05$ ), and nation and gender ( $p < .05$ ) for Booklet Four. All other interactions between other variables were not statistically significant ( $p > .05$ ).

Table 37  
*Partial Associations for Booklet Four*

Effect	df	Partial Chi-Square	<i>p</i>
LCA*NATION*MRM	6	.00	1.00
LCA*NATION*ITSEX	6	3.25	.78
LCA*MRM*ITSEX	2	.00	1.00
NATION*MRM*ITSEX	3	1.33	.72
<b>LCA*NATION</b>	<b>6</b>	<b>65.39</b>	<b>&lt;.001</b>
<b>LCA*MRM</b>	<b>2</b>	<b>1362.86</b>	<b>&lt;.001</b>
<b>NATION*MRM</b>	<b>3</b>	<b>10.46</b>	<b>.02</b>
LCA*ITSEX	2	4.41	.11
<b>NATION*ITSEX</b>	<b>3</b>	<b>10.05</b>	<b>.02</b>
MRM*ITSEX	1	.13	.72
LCA	2	42.13	<.001
NATION	3	216.13	<.001
MRM	1	169.78	<.001
ITSEX	1	.15	.70

*Note.* NATION= Countries, ITSEX=Gender, LCA= Latent Class Analysis Group Membership, MRM= Mixed Rasch Model Group Membership

To further analyze the interactions of nation, LCA class membership, and the MRM class membership variables a custom model was created with the significant two-way associations. In Table 38, the goodness of fit test showed that the model fit the data adequately ( $p > .05$ ).

Table 38  
*Goodness-of-Fit Tests for 2-way Interaction Model for Booklet Four*

	Chi-Square	df	<i>p</i>	Adjusted	
				df <sup>a</sup>	<i>p</i>
Likelihood Ratio	7.99	26	1.00	10	.63

a. One degree of freedom is subtracted for each cell with an expected value of zero. The unadjusted df is an upper bound on the true df, while the adjusted df may be an underestimate.

**Associations.** A crosstab analysis was run to see the exact membership percentages between interactions. Table 39 results support latent class identification as based on skill level for Booklet Four.

Table 39

*Crosstabulation of Nation vs. LCA Class Membership for Booklet Four*

		LCA GROUP MEMBERSHIP				
		Class 1	Class 2	Class 3	Total	
NATION	Turkey	Count	52	142	210	404
		% within NATION	12.9%	35.1%	52.0%	100.0%
	USA	Count	138	302	242	682
		% within NATION	20.2%	44.3%	35.5%	100.0%
	Singapore	Count	233	126	51	410
		% within NATION	56.8%	30.7%	12.4%	100.0%
	Finland	Count	50	124	76	250
		% within NATION	20.0%	49.6%	30.4%	100.0%
Total		Count	473	694	579	1746
		% within NATION	27.1%	39.7%	33.2%	100.0%

Since there was an interaction between nation and gender variables for Booklet Four, a crosstabulation analysis was done to see the levels. For this booklet, there were more male Turkish students than females (57.9% vs. 42.1%). Both USA and Singapore had almost equal percentages for gender (see Table 40). However, there were more girls than boys for Finland (54.0% vs. 46.0%).

Table 40

*Crosstabulation of Nation vs. Gender for Booklet Four*

		GENDER		Total	
		GIRL	BOY		
NATION	Turkey	Count	170	234	404
		% within NATION	42.1%	57.9%	100.0%
	USA	Count	353	329	682
		% within NATION	51.8%	48.2%	100.0%
	Singapore	Count	207	203	410
		% within NATION	50.5%	49.5%	100.0%
	Finland	Count	135	115	250
		% within NATION	54.0%	46.0%	100.0%
Total		Count		865	881
		% within NATION		49.5%	50.5%

The MRM analysis had a two-class solution for Booklet Four (see Table 41). Based on that, a majority of the Turkish, American, and Finnish students fell into the first class, 82.9 %, 70.5% and 72.8 respectively. However, Singaporean students were mostly in class 2 (64.6 %).

Table 41  
*Crosstabulation of Nation vs. MRM Class Membership for Booklet Four*

		MRM GROUP MEMBERSHIP			
		Class 1	Class 2	Total	
NATION	Turkey	Count	335	69	404
		% within NATION	82.9%	17.1%	100.0%
	USA	Count	481	201	682
		% within NATION	70.5%	29.5%	100.0%
	Singapore	Count	145	265	410
		% within NATION	35.4%	64.6%	100.0%
	Finland	Count	182	68	250
		% within NATION	72.8%	27.2%	100.0%
Total		Count	1143	603	1746
		% within NATION	65.5%	34.5%	100.0%

Additionally, a crosstab analysis for Booklet Four was done to see LCA class memberships and the MRM class membership agreement level. Although LCA and MRM analysis provided a different number of classes for Booklet Four, LCA's class one (highly skilled students) overlapped 100 % with MRM class two. LCA class two (moderate skill students) overlapped with both MRM class one (81.3%) and class two (18.7%). LCA class three (somewhat moderate skilled students) overlapped with only MRM class one (see Table 42).



Table 42

*Crosstabulation of LCA Class Membership vs. MRM Class Membership for Booklet Four*

		MRM GROUP MEMBERSHIP		
		Class 1	Class 2	Total
LCA GROUP MEMBERSHIP	Count	0	473	473
	% within LCA GROUP MEMBERSHIP	0.0%	100.0%	100.0%
	Class 1			
Class 2	Count	564	130	694
	% within LCA GROUP MEMBERSHIP	81.3%	18.7%	100.0%
	Class 2			
Class 3	Count	579	0	579
	% within LCA GROUP MEMBERSHIP	100.0%	0.0%	100.0%
	Class 3			
Total	Count	1143	603	1746
	% within LCA GROUP MEMBERSHIP	65.5%	34.5%	100.0%

**Booklet Six.** A four way log-linear analysis was done with variables nation, gender, LCA class membership, and the MRM class membership. The likelihood ratio chi-square with no parameters and only the mean was 1829.25. The value for the first order effect was 1535.37. The difference 293.88 is displayed on the first line of the table (see Table 43). The significant  $p$  value ( $< .001$ ) shows that there was a first order effect. The addition of a second order effect improved the likelihood ratio chi-square by 1526.405. This was also significant. But the addition of a third and a fourth order term did not significantly improve fit ( $p > .05$ ).

Table 43

*K-Way and Higher-Order Effects for Booklet Six*

	K	df	Likelihood Ratio	
			Chi-Square	$p$
K-way Effects	1	6	293.882	<.001
	2	12	1526.405	<.001
	3	10	8.957	1.000
	4	3	0.005	1.000

Table 44 shows what interactions were significant. Similar to the other booklets, there were statistically significant associations between nation and LCA class membership ( $p < .05$ ), nation and the MRM class membership ( $p < .05$ ), and LCA class membership and MRM class membership ( $p < .05$ ). All other interactions between other variables were not statistically significant ( $p > .05$ ).

Table 44  
*Partial Associations for Booklet Six*

Effect	df	Partial Chi-Square	<i>p</i>
NATION*ITSEX*LCA	3	1.98	.57
NATION*ITSEX*MRM	3	3.52	.31
NATION*LCA*MRM	3	3.83	.28
ITSEX*LCA*MRM	1	.92	.33
NATION*ITSEX	3	.94	.81
<b>NATION*LCA</b>	<b>3</b>	<b>104.55</b>	<b>&lt;.001</b>
ITSEX*LCA	1	.04	.82
<b>NATION*MRM</b>	<b>3</b>	<b>72.23</b>	<b>&lt;.001</b>
ITSEX*MRM	1	.17	.68
<b>LCA*MRM</b>	<b>1</b>	<b>1019.36</b>	<b>&lt;.001</b>
NATION	3	185.68	<.001
ITSEX	1	.42	.51
LCA	1	3.48	.06
MRM	1	104.28	<.001

*Note.* NATION= Countries, ITSEX=Gender, LCA= Latent Class Analysis Group Membership, MRM= Mixed Rasch Model Group Membership

Table 45 shows the goodness of fit test indicating that the model including the main marginal and the significant associations fit the data adequately ( $p > .05$ ).

Table 45  
*Goodness-of-Fit Tests for 2-way Interaction Model for Booklet Six*

	Chi-Square	df	<i>p</i>
Likelihood Ratio	4.22	3	<b>.24</b>

**Associations.** A crosstab analysis was done to investigate the exact membership percentages between interactions. The agreement between nation and LCA class membership variables is shown in Table 46. Most of the Turkish students fell into the second class of the LCA (73.9%). On the other hand, most of the Singaporean students fell into the first group of the LCA (80.3%). However, Finland and USA had comparable percentages for both classes (46.2% and 53.8% for Finland, 41.1% and 58.9% for USA).

Table 46  
*Crosstabulation of Nation vs. LCA Class Membership for Booklet Six*

		LCA GROUP MEMBERSHIP			
		Class 1	Class 2	Total	
NATION	Turkey	Count	100	283	383
		% within NATION	26.1%	73.9%	100.0%
	USA	Count	269	386	655
		% within NATION	41.1%	58.9%	100.0%
	Singapore	Count	323	79	402
		% within NATION	80.3%	19.7%	100.0%
	Finland	Count	121	141	262
		% within NATION	46.2%	53.8%	100.0%
Total		Count		889	1702
		% within NATION	47.8%	52.2%	100.0%

Additionally, the MRM analysis also had a two-class solution for Booklet Six (see Table 47). A majority of the American, Singaporean, and Finnish students fell into the first class, 64.3 %, 86.1% and 62.1% respectively. However, Turkish students were mostly in class 2 (65.8 %).

Table 47

*Crosstabulation of Nation vs. MRM Class Membership for Booklet Six*

		MRM GROUP MEMBERSHIP			
		Class 1	Class 2	Total	
NATION	Turkey	Count	131	252	383
		% within NATION	34.2%	65.8%	100.0%
	USA	Count	421	234	655
		% within NATION	64.3%	35.7%	100.0%
	Singapore	Count	346	56	402
		% within NATION	86.1%	13.9%	100.0%
	Finland	Count	162	99	261
		% within NATION	62.1%	37.9%	100.0%
Total		Count	1060	641	1701
		% within NATION	62.3%	37.7%	100.0%

A crosstab analysis for Booklet Six was also run to see the LCA class membership and the MRM class membership agreement level. Table 48 shows that LCA's class one overlaps perfectly with MRM's class two. However LCA's class two fell into both MRM class one and class two, 81.3% and 18.7% respectively.

Table 48

*Crosstabulation of LCA Class Membership vs. MRM Class Membership for Booklet Six*

		MRM GROUP MEMBERSHIP			
		Class 1	Class 2	Total	
LCA GROUP MEMBERSHIP	Class 1	Count	0	473	473
		% within LCA GROUP MEMBERSHIP	0.0%	100.0%	100.0%
	Class 2	Count	564	130	694
		% within LCA GROUP MEMBERSHIP	81.3%	18.7%	100.0%
Total		Count	1143	603	1746
		% within LCA GROUP MEMBERSHIP	65.5%	34.5%	100.0%

## **Chapter Four**

### **Discussion**

This chapter presents a summary of the paper, important findings for each research question, limitations of the research, and recommendations for further research.

#### **Summary of the Study**

This study compared the results of LCA and the MRM analyses for a major international assessment in mathematics. Although the two approaches and the outcomes in terms of class designations overlapped, assumptions about the nature of the data and the information derived from each analysis differed. The literature review summarized the theory and application of latent class analysis and the mixture Rasch model in identifying latent classes in the social sciences. Also, a log-linear analysis was conducted to understand the interactions between latent classes identified by LCA and the MRM. The data set used in the study was from four diverse countries (Turkey, USA, Finland, and Singapore) participating in TIMSS-2011. There are instructional differences and historical performance differences for each country and analyses yielded results associated mostly with nation of the participants.

The TIMSS-2011 8th-grade mathematics section contained 48 released items within 14 different booklets. Since booklets had overlapping items, to cover the largest

number of items within single booklets, booklets one, four, and six were selected for analysis. This method resulted in the coverage of 40 single items.

Latent class analysis is used to determine if individuals can be divided into subgroups or latent classes based on an unobserved construct (Collins & Lanza, 2010). The analyses run by booklet revealed the number of underlying subgroups and suggested a potential meaning for classes. Models used in the study explored different subgroups within the data set based on participants' responses. On the other hand, the mixture Rasch model is also used for similar purposes to understand the nature of the data. Mixture Rasch models, which combine Rasch models with latent class analysis, have been used to identify latent classes based on use of different problem-solving techniques or who use different skills in response to test items. For each technique, different fit indices were used to find the best model. The Bayesian information criterion (BIC) was used for LCA. Cressie Read and chi-square were used for the MRM.

Results of log-linear analysis showed that overall the two techniques provided similar but not identical results. There were significant interactions between nation and identified latent classes for both LCA and the MRM. Also a crosstab analysis uncovered the agreement level of 2-way interactions from log-linear analysis showing the level of agreement between identified classes of LCA and nation and also identified classes of MRM and nation as well as identified classes of LCA vs. the MRM.

### **Important Findings**

**Research question one.** Does analysis of TIMSS-2011 8th-grade mathematics data from four proposed nations yield multiple latent classes using LCA techniques?

The three booklets from the TIMSS-2011 8th-grade mathematics data were used to run the LCA method explained in the previous chapter via *Mplus* Version 7.31 (Muthén & Muthén, 2012a). The three-step modeling approach by Wang and Wang (2012) was used to explore possible latent structures. The main goal of the present research question was to identify distinct latent classes and descriptive features of the dataset. For each booklet, analysis yielded a different number of subgroups with adequate model fit and adequate quality, and seemed to define latent classes of different ability levels ranging between low to high ability level.

The results suggest that TIMSS-2011 8th-grade mathematics data yield different subgroups based on ability levels of students. The interpretation of the classes was based on the demographic information related to students' background. The students selected for the study were from different nations with different rankings on TIMSS achievement. Wang and Wang (2012) emphasize that, for a successful LCA model it is important to construct the definition of the latent classes in an interpretable manner.

Initial thought about the LCA part of the study was if there were multiple latent classes, this would be considered a test validity issue. But analysis showed that the identified latent classes can be generally defined by the nation variable. For example, analysis of the data suggest that there were three possible latent classes for Booklet One. Results showed that 519 students (29.5%) were assigned to Class 1, 743 students (37.6%)

were assigned to Class 2, and 502 students (28.5%) were assigned to Class 3. Results of the analysis assigned one fourth of the students to the higher ability group, half of them to the moderate ability group, and one fourth of them to the low ability group for Booklet One.

The ability levels of the students overlap with where their nation stands on the TIMSS mean score table. The three class (also 2-class) solutions suggested that latent classes are tied to the nation variable, in which instructional differences, class sizes, number of hours, money spent on education, etc. makes the main difference. So for LCA analysis, it is clear that multiple latent classes are associated with data being obtained from different nations that have clear performance differences on the test. Also, literature on TIMSS results suggests that the mean achievement score differences between countries that occur on an international test are the result of multiple practice differences within the system itself (Carnoy, & Rothstein, 2013; Stigler, Gonzales, Kwanaka, Knoll, & Serrano, 1999; Yücel, Karadağ, & Turan, 2013). So the results of the LCA analysis became a confirmatory result for the nature of the classes identified within the study.

**Research question two.** Does analysis of TIMSS-2011 8th-grade mathematics data from four proposed nations yield multiple latent classes using the MRM techniques?

Response data from the three booklets were used to run the mixture Rasch model analysis using WINMIRA (von Davier, 2001a). The main goal of the research question two was to find whether distinct latent subgroups were identified. For each booklet, MRM yielded a different number of latent classes with adequate model fit based on the



item difficulty and response patterns of the students. Once the latent classes were identified, item fit was examined. Item fit statistics are handled slightly differently in a MRM than in a simple Rasch analysis. Each possible latent class yielded its own Rasch analysis and its own set of item and person position estimates and fit statistics, along with point-scale indicators.

Results of the current study show that examinees systematically differed in the ways they understand or solve items. For each booklet there was more than one class. MRM, in this case, provides valuable information to the field. Having distinct classes means that pupils may employ different strategies to solve test items, an idea which has been emphasized by cognitive psychologists doing psychometric studies for the past decades (Mislevy & Huang, 2007).

In general, items were relatively easy for two booklets and relatively difficult for one booklet. For example, raw mean scores for Booklet One show that participants in classes one and two had substantially higher mean scores than participants in class two. Analysis showed that the mean of the raw scores of class 1 was high ( $M=9.02$ ,  $SD=2.35$ ), class 2 was medium ( $M=7.18$ ,  $SD=2.26$ ), and class 3 was low for Booklet One ( $M=3.42$ ,  $SD=1.61$ ). Also, for booklet three, class 1 was low ( $M=3.97$ ,  $SD=1.77$ ), class 2 was high ( $M=6.39$ ,  $SD=1.32$ ), for Booklet Six class 1 was low ( $M=13.42$ ,  $SD=3.16$ ), class 2 was high ( $M=6.66$ ,  $SD=2.63$ ). Also item logit positions show that the same item was relatively hard for participants in different classes. For example, item M032047 of Booklet Six had an item difficulty parameter of 1.07 and a standard error of 0.07 for class

one and an item difficulty parameter of -0.34 and standard error of 0.08 for class two. The difference in logit measures were around more than three standard errors different which can be interpreted as item being comparatively hard for the students in class one for Booklet Six (Masters & Keeves, 1999). This also supports the idea that the appearance of distinct classes is related to person ability.

The MRM can be used for different purposes. In this study, validity of the test was a concern of the researcher. Possible latent classes were seen as threats to validity. Messick (1989) states that, validity is not just guessing about some behavior but exploration of the strategies and processes that take part in the minds of participants during the exam. Similar to research question one, the reason why students belong to different latent subgroups could be because they are from different educational systems, which in this case they were (Mislevy & Huang, 2007).

Having different latent classes means that construct validity may be called into question. In this case, this means that the same construct is not being measured similarly for all students. Additionally the interpretation of the construct does not apply to all latent classes similarly. Since there were threats to construct validity by the existence of different latent classes, examinee classification can be done prior to interpretation. Interpreting the tested construct would be more appropriate based on the latent class. Response patterns seemed different, possibly for different achievement groups. If there were no major differences between the item logit scores of the latent classes of the MRM analysis (see Figure 14, 15 and 16), results could be interpreted as the latent classes

describe the same dimension. In such a case results of the MRM analysis would support construct validity. However, differences in item logit positions per class suggest further investigation into test validity.

**Research question three.** Do LCA and the MRM analysis results differ in terms of:

- a. Item fit parameters for TIMSS-2011 8th-grade mathematics?
- b. Item class parameters for TIMSS-2011 8th-grade mathematics?

Results from the analyses of research question one and research question two were compared to see if LCA and the MRM differ as far as item fit and item class parameters.

For item parameters, both of the techniques calculate item logit values and standard errors. For LCA, item parameter estimates are on the logit scale, and therefore, can be somewhat difficult to interpret. The same information is given in a more interpretable scale under the MRM where item parameters are products of item difficulty measure for each class. However standard errors of the parameters have very close results for all booklets (see Tables 21, 24 and 27)

The decision on number of classes differs in the two techniques. BIC and AIC were used to evaluate fit for LCA. On the other hand, since Winmira2001 considered data as being sparse, Cressie-Read and Chi-square values were used for model fit purposes. However, based on BIC values, both techniques provided similar results (see Tables 22, 25, and 28). So it can be concluded that selecting one model over another model did not

depend on fit values. Since a qualitative conclusion is important for LCA, model fit is not enough by itself. There are also other combinations of different values such as average estimated posterior probabilities for quality (Nagin, 2005) and entropy value (Clark, 2010). Moreover, latent classes should be defined in an interpretable way as well. For the MRM, the solution is simpler. If there is model fit based on fit indices the next step is simply interpretation of the model.

The two analyses had somewhat different solutions for the class weights for all booklets. It can be interpreted that latent class analysis puts the most cases into the middle class for three class solutions and to the second class for two class solution. LCA uses response probabilities in which students have the same probability of giving the correct answer within the same class. As a result of this, students in the same class have no quantitative differences. The only difference created and shown by LCA is between groups which is a product of qualitative differences. In our case, this would be interpreted as item correct response values based on students' background. However, the mixture Rasch model, regardless of number of classes within the solution, sorts classes based on similarity in their response patterns which results in the placement of cases with an order where most student fall in to the first class, than second, than third etc. Since there are differences between item parameters within the same class for the MRM, interpretation changes and relies on two things: one being latent class membership and two being the class specific quantitative person parameter (Büsch, Hagemann, & Bender, 2010).

**Research question four.** Are there associations between LCA and the MRM latent classes, nation, and gender for TIMSS-2011 8th-grade mathematics?

A four way log-linear analysis was performed with variables nation, gender, LCA class membership, and the MRM class membership. Results from the analyses of research question one and research question two as well as gender and nation were used to see if there were significant interactions between these variables. Results found significant interactions between nation vs. LCA class membership, nation vs. the MRM class membership, and also LCA class membership vs. the MRM class membership for all booklets. There was also another significant interaction between nation and gender for Booklet Six. A follow-up crosstab analysis was also run to help interpret the interaction.

The results revealed that qualitative meaning of the latent class analysis rests on the idea of students being from different educational systems. Although there were four nations, none of the analyses created four distinct groups. This is likely because there were students who are from different nations but within the same ability level. Also LCA class memberships only reflect countries' success rates in TIMSS. To some extent where the USA and Finland fell in the association table suggests that LCA classes were not a product of the nation variable (see Tables 33, 39 and 46).

Similarly, there was a significant interaction between nation and the MRM class membership. However, the MRM results overlapped with nation data better than the LCA results (see Tables 34, 41, and 47). Based on where nations stand within the TIMSS mean score placement, Singapore was in mostly one class, USA and Finland were

combined in another class, and Turkey was by itself for three class solutions. This was one advantage of the dataset where latent classes could be qualitatively defined based on students coming from different nations since the nation variable was provided before the analysis.

Another interaction was found between nation and gender for Booklet Four. For this booklet, there were more male Turkish students than females (57.9% vs. 42.1%). Both USA and Singapore had almost equal percentages for gender (see Table 40). However, there were more girls than boys for Finland (54.0% vs. 46.0%). There was also significant interaction between LCA class membership results and the MRM class membership results for all booklets. The results of crosstab analysis showed that there was agreement between class memberships of both techniques (see Tables 35, 42, and 48).

Research question one and two explored whether analyses of TIMSS-2011 mathematics data yielded multiple latent classes with LCA and the MRM techniques. Past studies have found that LCA and the MRM techniques commonly revealed the existence of distinct latent classes within international large scale assessments. Studies have shown that TIMSS and like tests (PISA, PIRLS) contain distinct latent classes, mostly based on item difficulty and related to student background (e.g., Choi et al., 2015; Oliveri, et al., 2014; Zhang et al., 2015). Findings of this study tally with the results of past studies conducted using similar datasets.

Also a similar comparison study was conducted and found that a MRM-related technique presented a clearer interpretation of latent classes (Choi et al., 2010). The present study also suggests that, although two techniques provide similar results, the MRM provides more interpretable results in terms of definition of the latent classes.

### **Implications for Research**

There are several connections between LCA and MRM. They are similar in the assumption that observed data structures result from a latent construct. With both LCA and the MRM, observed variables are assumed independent, conditional on values of the latent variable. For LCA, the latent variable concludes the data structure is nominal (latent class membership). On the other hand, for the MRM, the latent variable that determines data structure is continuous (Collins & Lanza, 2010 ; von Davier, 2001a).

The MRM can be called the “super combination” version of the Rasch and latent class analysis, because participants are assigned to the qualitatively scaled latent class variables based on their item response patterns; and they are also assigned quantitatively to a latent class based on the number of items solved (Rost, 1990; Tenenbaum, Strauss, & Büsch, 2007). On the other hand, results of LCA provided sample information (proportion of people in each class), item information (probability of correct response for each item from examinees from each class) and examinee information (posterior probability of class membership for each examinee in each class).

While using LCA, it was noticed that LCA can be an exploratory procedure for understanding data, since classes are not known prior to analysis and class characteristics are not known until after analysis. However the MRM was found more useful in the TIMSS-2011 dataset because students' educational background was the main difference between examinees. When examinees systematically differ in the ways they understand or solve items, or coming from different educational backgrounds, the variance within the dataset might lead to differences in item solution techniques which most likely causes difference in item position parameters and therefore to different latent classes.

LCA is useful in examining the relationships between indicator variables due to class membership only. Also it calculates class membership probabilities instead of fixed class memberships. For the dataset used in this paper, LCA created a three class solution for Booklet One where the MRM also provided the same number of classes. For Booklet One, since LCA does not arrange classes by the size of them and tries to emphasize the qualitative underlying idea of the dataset, class loadings were mostly on second class. The MRM, on the other hand align class loadings in an order starting with the biggest class. This gives one advantage to the MRM over LCA. It makes easier to interpret the class loadings of the MRM results.

At this point, researchers might emphasize that LCA provides statistical optimization. However gaining statistical optimization may mean that classification interpretability and usability can be lost where there is no background information supplied (Dallas & Wilse, 2013). The solution the mixture Rasch model provides on this



matter is using item difficulty parameters. For example, for Booklet Four, the MRM provided a two class solution where class sizes were 66.0% and 34.0% respectively. Since the main product of each class is item difficulty parameters, interpretation of classes is derived from differences in item difficulties (see Table 18).

The findings of this paper do not reveal unequivocally whether a model based on primarily qualitative differences (LCA), that is, different strategies, instructional differences, curriculum etc. or a model including additional factors of quantitative differences within strategies (MRM) should be used with this particular dataset. Both of the tests provided similar results with more or less similar interpretations. Both techniques resulted in models that fit the data similarly. Nonetheless, for tests similar to TIMSS exams, item difficulty parameters can be useful for educational researchers, suggesting MRM analysis may be more productive.

### **Implications for Turkish Educators**

One of the reasons for using the TIMSS-2011 data set was to be able to explore where Turkish students stand compared to their rivals within the TIMSS participants. According to the crosstab analysis of nation vs. the MRM results, Turkish students mostly fell into class three where student ability level was lower than the other participating nations for Booklet One (50.7%). Most of the items in this booklet were from the knowing cognitive domain with number and algebra content domain. For Booklet Four, where a two class solution was advised by the MRM analysis, Turkish students were mostly in class one (82.9%). For this booklet, selected items were slightly

easier for students in class-one (see Figure 15). Especially, items M032324, M032116, M032100, M032402 and M032397, which are in the geometry content domain, were easier for Turkish pupils. Additionally, for Booklet Six, Turkish students were mostly in class two (65.8%). Again, geometry content-related items were easy for this class. However, the algebra content domain related questions were harder for this class (see Figure 16).

Overall, the MRM analysis and crosstab analysis of nation vs. the MRM results showed that most items were either easier or somewhat easier for classes where Turkish students are the majority. The average TIMSS-2011 mathematics score of 452 ( $M=500$ ,  $SD=100$ ) for Turkish 8th-graders could be influenced by lower scores on open-ended questions with which Turkish students are not familiar since the educational system mostly relies on multiple choice-based large-scale assessments.

As a result of this study, Turkish educators should note that although results from multiple choice item place Turkey at a higher level, Turkey's place within the TIMSS-2011 results acknowledge that students show weaker performance levels for the rest of the test. Students being focused on test solving techniques more than real-life-related open-ended problems indicate that Turkey should consider evolving its high school and university entrance method to a more modern system than is the case currently. Also, recent education developments will not be felt by industry or society since students are mainly interested in the exam results more than what is being taught at schools.

However, it is important to note that limitations of the TIMSS exams are important. While TIMSS provides informative knowledge for countries, it is important to understand that inferences can be made for only those narrowly defined populations regarding its performance on a narrowly defined set of topics (Rutkowski & Rutkowski, 2016). Turkey is one of the biggest participants of the TIMSS exams. Making policy change recommendations based on such results should be very carefully examined by all parties of the education system.

### **Limitations of the Study**

As with any statistical approach that uses binary variables, recoding categorical responses into dichotomous responses was one of the limitations of the study since student responses might result in different classification based on the multiple choice responses. In any latent class model, the issue of reification is of great importance. Also using a real world dataset limited the radius of effect area of the study since conclusions are limited to the current data.

Sampling techniques of TIMSS organizers is also another limitation. One simple example shows that number of students in Turkish and American educational systems are more than the whole population of Singapore and Finland. TIMSS requires each participant country to join with at least 4.500 students. Although this number covers most of the Singaporean and Finnish 8th-grade population, it is still small for systems like the US or Turkey (Rutkowski & Rutkowski, 2016). In this case, generalizability of the results are questionable.

Additionally, TIMSS organizers only released a small portion of the items. Running the analysis only on this small set of released items also limited the interpretation and generalizability of the results. Sample size was reduced due to the number of released items. There were only three booklets used out of 14 booklets. As a result of this, results found here may not be representative of other booklets in TIMSS-2011. Therefore results of this study may show some limitations regarding generalizability to all of TIMSS. It is also important to note that countries in this study differed greatly in ability levels of students. This also brings up generalizability questions related to whether it is possible to find distinct latent classes if countries with more similar scores had been examined.

### **Recommendations for Further Research**

This study provides useful information about two commonly used techniques in educational research. Since the data used in this study are from a real data set, none of the techniques were tested under controlled circumstances such as different levels of amount and type of missing data, presence of outliers, sample size (bigger, smaller), item distributions, score distributions, etc. Monte Carlo simulation studies are recommended to see if the results differ under these different conditions.

Further, TIMSS multiple-choice items were dichotomous; use of items with varied responses scales is also recommended, as are studies with item content very different from a mathematics achievement test. For example, studies are recommended that compare LCA and MRM when the construct assessed is a personality variable or

attitudinal as well as achievement. The comparison of both techniques is limited to dataset used in this study. Therefore, it is suggested that same study can be done using other type of questionnaires.

For international exams, cross-cultural comparisons are important. Participating countries were selected to create variance. Hence, the same study should be conducted to see if countries whose education systems are similar also provide similar results. Future research should also be conducted using data from other countries using different languages as well as at different ability levels. Researchers who are interested in comparison of LCA and MRM can replicate the study using data from countries that differ on characteristics other than education system. Moreover, same study can be done using items which have similar item difficulty parameters to see if latent subgroups occur in the data. Also it might be useful to group item based on their content area to see if both techniques produce same results. Furthermore, studies are recommended to see if the results differ based on the science section of the test.

## References

- Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (pp. 267–281), Akademinai Kiado.
- Atkin, J. M., & Black, P. (1997). Policy perils of international comparisons: The TIMSS case. *Phi Delta Kappan*, 79(1), 22-28. Retrieved from <http://0-search.proquest.com.bianca.penlib.du.edu/docview/218481513?accountid=14608>
- Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation*, 18(5), 1-13.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345-370.
- Buyruk, H. (2015, June). *Current Developments in School Education in Turkey: education 'reforms' and teacher trade union responses*. In FORUM (Vol. 57, No. 2, pp. 147-166). Symposium Journals.
- Büsch, D., Hagemann, N., & Bender, N. (2010). The dimensionality of the Edinburgh Handedness Inventory: An analysis with models of the item response theory. *Laterality*, 15(6), 610-628.

- Carnoy, M., & Rothstein, R. (2013). What do international tests really show about US student performance. *Economic Policy Institute*, 28.
- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2), 195-212.
- Cho, S., J., Cohen, A.S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture IRT measurement model. *Applied Psychological Measurement*, 34, 583–604.
- Choi, Y. J., Alexeev, N., & Cohen, A. S. (2015). Differential Item Functioning Analysis Using a Mixture 3-Parameter Logistic Model With a Covariate on the TIMSS 2007 Mathematics Test. *International Journal of Testing*, 15(3), 239-253.
- Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis*. Hoboken, NJ: John Wiley & Sons, Inc.
- Clark, S. L. (2010). *Mixture modeling with behavioral data* (Doctoral dissertation). Retrieved from <http://0-search.proquest.com.bianca.penlib.du.edu/pqdtft/docview/230964800/fulltextPDF/13A850185757B2D864B/>
- Cressie, T. R. C. & Read, N. A. C. (1984). Multinomial goodness-of-fit statistics. *Journal of the Royal Statistical Society Series B*, 46, 440–464.
- Dallas, A. D., & Willse, J. T. (2013). Survey analysis with mixture Rasch models. *Journal of Applied Measurement*, 15(4), 394-404.

- Embretson, S. E. (2006). Mixed Rasch models for measurement in cognitive psychology. In von Davier, M. & Carstensen C. H. *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York, NY: Springer
- Fischer, G. H., & Molenaar, I. W. (Eds.). (2012). *Rasch models: Foundations, recent developments, and applications*. NY: Springer Science & Business Media.
- Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement, 75*(2), 208-234.
- George, D., & Mallery, M. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update* (10a ed.) Boston: Pearson.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2), 215-231.
- Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications* (pp. 2-34). Springer New York.
- Gupta, S. C., & Kapoor, D. V. (2000). *Fundamentals of mathematical statistics: A modern approach*. India: Sultan Chand.
- Haberman, S. J. (1979). *Analysis of qualitative data. vol. 2, new developments*. London: Academic Press.



Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Thousand Oaks, CA: Sage Publications, Inc.

Harrington, D. (2008). *Confirmatory factor analysis*. NY: Oxford University Press.

Higginbotham, D. L. (2013). *An assessment of character and leadership development latent factor structures through confirmatory factor, item response theory, and latent class analyses* (Unpublished doctoral dissertation). University of Denver.

Holliday, W. G. (1999). Questioning the TIMSS. *The Science Teacher*, 66(1), 34-37.

Retrieved from <http://0->

[search.proquest.com.bianca.penlib.du.edu/docview/214627161?accountid=14608](http://search.proquest.com.bianca.penlib.du.edu/docview/214627161?accountid=14608)

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

<https://ies.ed.gov>

<http://www.minedu.fi/OPM/Koulutus/lukiokoulutus/?lang=en>

<http://www.moe.gov.sg/education/>

[http://www.oph.fi/english/education\\_system/basic\\_education](http://www.oph.fi/english/education_system/basic_education)

<http://www.tuik.gov.tr/UstMenu.do?metod=temelist>

Kalkınma Bakanlığı, (2013), “2013 Yılı Programı”, *10. Kalkınma Planı* (2014-2018)

[http://www.kalkinma.gov.tr/DocObjects/view/15089/Onuncu\\_Kalk%C4%B1nma\\_Plan\\_%C4%B.pdf](http://www.kalkinma.gov.tr/DocObjects/view/15089/Onuncu_Kalk%C4%B1nma_Plan_%C4%B.pdf)

Kamakura, W. A., & Wedel, M. (2000). Factor analysis and missing data. *Journal of Marketing Research*, *37*(4), 490-498.

Kamerman, S. B. (2000). Early childhood education and care: an overview of developments in the OECD countries. *International Journal of Educational Research*, *33*(1), 7-29..

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331-358.

Knoke, D., & Burke, P. J. (1980). *Log-Linear Models*. Sage Publications, Inc. Newberry Park, CA.

Kupiainen, S., Hautamäki, J., & Karjalainen, T. (2009). *The Finnish education system and PISA*. Helsinki: Ministry of Education.

Lazarsfeld, P. F. (1950). Logical and mathematical foundation of latent structure analysis. In *Measurement and Prediction* (SA Stouffer *et al.*, eds.) *4*, 362-412. Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. (1959). Latent structure analysis. *Psychology: A study of a science*, *3*, 476-542.

- Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22(3), 249-264.
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3), 767-778.
- Lord, F. M. (1952). A theory of test score (Psychometric Monograph No. 7). *Psychometric Society*, 35. Iowa City, IA.
- Masters, G. N., & Keeves, J. P. (1999). *Advances in measurement in educational research and assessment*. Pergamon. New York, NY.
- McCutcheon AL. *Latent Class Analysis*. Newbury Park, Calif: Sage Publications; 1987.
- McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York: John Wiley & Sons.
- MEB-EARGED (2003). Üçüncü Uluslararası Fen ve Matematik Çalışması (TIMSS 1999) Ulusal Rapor [Third international mathematics and science study (TIMSS 1999) national report]. Ankara: Turkey.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. New York: Macmillan.
- Ministry of Education and Culture, Finland. (2012). *Finnish Education in a Nutshell*. Helsinki, Finland.

- Ministry of Education, Singapore. (2002b). *Junior college/upper secondary education review report*. Singapore: Author.
- Mislevy, R., & Huang, C., W. (2007). Measurement models as narrative structures. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 15-35). New York: Springer Verlag.
- Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report*. Boston: Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., & Smith, T. A. (2000). *TIMSS 1999 international mathematics report*. Boston: International Study Center.
- Muthén, B. O. (2001). Latent variable mixture modeling. In G. Marcoulides & R. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1-33). Mahwah, NJ: Lawrence Erlbaum Associates.

- Muthén, B. O. (2004). *Mplus technical appendices* (3rd ver.). Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (2012a). *Mplus* (Version 7.31) [Computer Software]. Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (1998). 2014. *Mplus User's Guide, 7th edition*. Muthén & Muthén, Los Angeles.
- Nagin, D. (2005). *Group-based modeling of development*. Harvard University Press.
- Ng, I. (2007). Intergenerational income mobility in Singapore. *The BE Journal of Economic Analysis & Policy*, 7(2).
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14(4), 535-569.
- OECD. (2009a). *PISA 2009 Assessment framework – key competencies in reading, mathematics and science*. Paris: OECD.
- OECD. (2010). *Improving health and social cohesion through education*. Paris: OECD.
- Oliveri, M. E., Ercikan, K., Zumbo, B. D., & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments—Comparing a latent class to a manifest DIF approach. *International Journal of Testing*, 14(3), 265-287.

- Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen, Denmark: Danish Institute for Educational Research*, 1960.
- Robitaille, D.F., & Robeck, E. D., (1996). The character and the context of TIMSS. In D.F. Robitaille and R.A. Garden (Eds.), *Research questions and study design. TIMSS Monograph N. 2*. Vancouver, Canada: Pasific Educational Press.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271-282.
- Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In *Rasch Models* (pp. 257-268). NY: Springer.
- Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). New York: Waxmann.
- Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, 45(4), 252-257.
- Samuelsen, K. M., & Dayton, C. M. (2010). Latent class analysis. In G. Hancock & R. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 173-184). New York: Routledge.

- Schneider, M. (2009). The international PISA test. *Education Next*, 9(4) Retrieved from <http://0-search.proquest.com.bianca.penlib.du.edu/docview/1237827029?accountid=14608>
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Shen, C. (2002). Revisiting the relationship between students' achievement and their self-perceptions: A cross-national analysis based on TIMSS 1999 data. *Assessment in Education: Principles, Policy & Practice*, 9(2), 161-184.
- Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt, Germany: Peter Lang.
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.
- Stigler, J. W., Gonzales, P., Kwanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States*. A Research and Development Report.
- Tenenbaum, G., Strauss, B., & Büsch, D. (2007). Applications of generalized Rasch models in the sport, exercise, and the motor domains. In *Multivariate and mixture distribution Rasch models* (pp. 347-356). Springer New York.

- TIMSS & PIRLS International Study Center. (2015). *TIMSS 2015 20 Years of Achievement Trends*. [Brochure]. Retrieved from [http://timssandpirls.bc.edu/home/pdf/T2015\\_TIMSS.pdf](http://timssandpirls.bc.edu/home/pdf/T2015_TIMSS.pdf).
- Toker, T., & Green, K, E. (2012, April). *Cognitive Diagnostic Assessment of TIMSS2007 Mathematics Achievement Items for 8th Graders in Turkey*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.
- Toker, T., & Green, K. (2014). An application of cognitive diagnostic assessment on TIMSS-2007 8th-grade mathematics items. *International Journal of Academic Research*, 6(4), 139-145. doi:10.7813/2075-4124.2014/6-4/B.22
- US Census Bureau. (2015). *Statistical abstract of the United States*. US Government Printing Office.
- Vermunt, J. K. (1997b). *EM: A general program for the analysis of categorical data*. WORC Research Paper, Tilburg University
- Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. *The Sage encyclopedia of social sciences research methods*, 549-553.
- von Davier, M. (2001). *WINMIRA* [Computer software]. Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften




- von Davier, M. (2001b). *WINMIRA user manual* [Computer software manual]. Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften
- von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–115). New York: Springer.
- Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Hoboken, NJ: John Wiley & Sons, Inc.
- Yücel, C., Karadağ, E., & Turan, S. (2013). TIMSS 2011 ulusal ön değerlendirme raporu. *Eskişehir Osmangazi Üniversitesi Eğitim Fakültesi, Eğitimde Politika Analizi Raporlar Serisi I*. Eskişehir.
- Zhang, D., Orrill, C., & Campbell, T. (2015). Using the mixture Rasch model to explore knowledge resources students invoke in mathematics and science assessments. *School Science and Mathematics, 115*(7), 356-365.

# Appendices

## Appendix A

### University of Denver Institutional Review Board Application

 UNIVERSITY OF DENVER RESEARCH & SPONSORED PROGRAMS University of Denver Institutional Review Board (IRB)	
C.4 Does this project involve access to identifiable private data or specimens from living individuals?	NO
C.5 Was the original data collection for the current project?	NO
C.6 Does this project consist exclusively of interviewing or surveying subjects about his/her area of expertise, with a focus on policies, practice, and/or procedures (e.g. the collected data does not focus on personal opinion or private information)?	NO
C.7 Is the project meant to record the stories, knowledge or experiences of individuals? Oral histories typically do not intend to answer a research question or hypothesis.	NO
<b>Section D: Determining if DU is Engaged in Research</b>	
DU is usually considered engaged in research if any of the following are true:	
<ul style="list-style-type: none"><li>• DU is the primary awardee on the grant or contract</li><li>• DU is the only site</li><li>• DU employees, students, or agents are obtaining consent/assent</li><li>• DU employees, students, or agents are interacting or intervening with human subjects. Examples include online surveys, interviews, participant observations, invasive/non-invasive study procedures, manipulating the subject's environment for research purposes, or</li><li>• DU employees, students, or agents are obtaining or receiving identifiable, private information or biological samples</li></ul>	
Does conducting this project cause DU to be engaged in research?	Yes <input type="checkbox"/> No <input checked="" type="checkbox"/>
<b>If a protocol (research narrative) exists for this project it must be submitted for review. Submit this request along with any supplemental documents that may aid in review of your project to the University of Denver IRB at: <a href="mailto:IRBAdmin@du.edu">IRBAdmin@du.edu</a></b>	
-15 DU IRB/Office of Research Compliance Determination of Human Subject Research Form, v.1, dated 5-13	



University of Denver Institutional Review Board (IRB)

questions or sample questions if they have been developed. List all attachments:

B.3 Are all of the data used in this project publicly available, (e.g. blog, aggregate data, etc.)?

Yes  No

**Section C: Is this Project Human Subjects Research as Defined by Federal Regulations?**

Research is defined in the Code of Federal Regulations, 45 CFR 46.102(d), as a systematic investigation designed to develop or contribute to generalizable knowledge

The Belmont Report states "... the term "research" designates an activity designed to test a hypothesis or formal protocol that sets forth an objective and a set of procedures to reach that objective."

Research generally does not include operational activities such as routine outbreak investigational and disease monitoring and studies for internal management purposes such as program evaluation, quality assurance, quality improvement, fiscal or program audits, marketing studies or contracted for services.

**Generalizable knowledge** is information where the intended use of the research findings can be applied to populations or situations beyond that study. Note that publishing the results of a project does not automatically meet the definition of generalizable knowledge.

C.1 Do you have a specific research question or hypothesis?

NO

C.2 Is your primary intent to generate knowledge that can be applied broadly to the group/condition under study?

NO

**Human subject is defined in the Code of Federal Regulations, 45 CFR 46.102(f)(1 or 2), as a living individual about whom an investigator obtains data through intervention or interaction or identifiable private information.**

The specimen(s)/data/information must be collected from or be about live subjects. Research on cadavers, autopsy specimens or specimens/information from subjects now deceased is not human subjects research.

C.3 Does this project involve intervention or interaction with a living individual or group of individuals? (e.g. confidential surveys, interviews, medical or educational testing)

NO



C.4 Does this project involve access to identifiable private data or specimens from living individuals? NO
C.5 Was the original data collection for the current project? NO
C.6 Does this project consist exclusively of interviewing or surveying subjects about his/her area of expertise, with a focus on policies, practice, and/or procedures (e.g. the collected data does not focus on personal opinion or private information)? NO
C.7 Is the project meant to record the stories, knowledge or experiences of individuals? Oral histories typically do not intend to answer a research question or hypothesis. NO
<b>Section D: Determining if DU is Engaged in Research</b>
DU is usually considered engaged in research if any of the following are true: <ul style="list-style-type: none"><li>• DU is the primary awardee on the grant or contract</li><li>• DU is the only site</li><li>• DU employees, students, or agents are obtaining consent/assent</li><li>• DU employees, students, or agents are interacting or intervening with human subjects. Examples include online surveys, interviews, participant observations, invasive/non-invasive study procedures, manipulating the subject's environment for research purposes, or</li><li>• DU employees, students, or agents are obtaining or receiving identifiable, private information or biological samples</li></ul>
Does conducting this project cause DU to be engaged in research?      Yes <input type="checkbox"/> No <input checked="" type="checkbox"/>
<b>If a protocol (research narrative) exists for this project it must be submitted for review. Submit this request along with any supplemental documents that may aid in review of your project to the University of Denver IRB at: <a href="mailto:IRBAdmin@du.edu">IRBAdmin@du.edu</a></b>

## Appendix B

### University of Denver Institutional Review Board Approval Letter



UNIVERSITY of  
DENVER

OFFICE OF RESEARCH AND SPONSORED PROGRAMS

February 4, 2016

Turker Toker  
RE: Determination of Proposed Project  
Project Title: A Comparison of Latent Class Analysis and the Mixed Rasch Model: A Cross-Cultural Comparison of 8<sup>th</sup> Grade Math Achievement in the Fourth International Mathematics and Science Study (TIMSS-2011)

Dear Mr. Toker,

Thank you for submitting the IRB Determination Form, dated 01/28/16, to the University of Denver Institutional Review Board for evaluation to determine if the above-referenced project qualifies as human subject research. Based on the information provided, it has been determined that the proposed project does not require IRB review. This determination is based on whether this proposed project is research with human subjects defined by the federal regulations.

The IRB Determination Form was evaluated and it was assessed that the proposed project outlined on the involves a comparison of two statistical methods based on publicly available international exam data. The information provided on the form clearly stated that there is no specific research question and all data obtained is publically available. This proposed project does not meet the regulatory definition of research with human subjects.

The Regulatory Definition of Research and Human Subject  
Federal research regulations define **research** as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge."

During the review of this proposed project, it was noted that publicly available data will be utilized to compare two statistical methods. The intent of the project does not intend to contribute to generalizable knowledge and therefore does not fulfill the definition of research.

Per the regulations, **Human subject** means a living individual about whom an investigator (whether professional or student) conducting research obtains 1) data through intervention or interaction with the individual, or 2) identifiable private information. This project will not obtain identifiable data and does not involve interacting with individuals which does not qualify this project as involving human subjects.

In order for a project to require IRB review, the proposed research must qualify under **both** definitions of being research and involving human subjects. This research project does NOT fulfill the regulatory definition of research or human subjects per the federal regulation definition.

Wray Speed Building, 222 1/2 39<sup>th</sup> St. University Blvd | Denver, CO 80202-4600 | Phone: 303.871.2121 | [www.ucd.edu/irb](http://www.ucd.edu/irb)

My evaluation, based only on the information provided, determined that the proposed project does not require IRB review. If you have questions regarding this determination or believe that this proposed project does qualify as human subject research, please feel free to contact me directly at 303-871-4049 or via e-mail at: [mary.travis@du.edu](mailto:mary.travis@du.edu).

Sincerely,



Mary Travis  
Director, Research Integrity & Education  
Office of Research and Sponsored Programs  
University of Denver

## Appendix C

LCA 2 Class Model Specification for Booklet One (Other Classes Similar) (Mplus

Version 7.11)

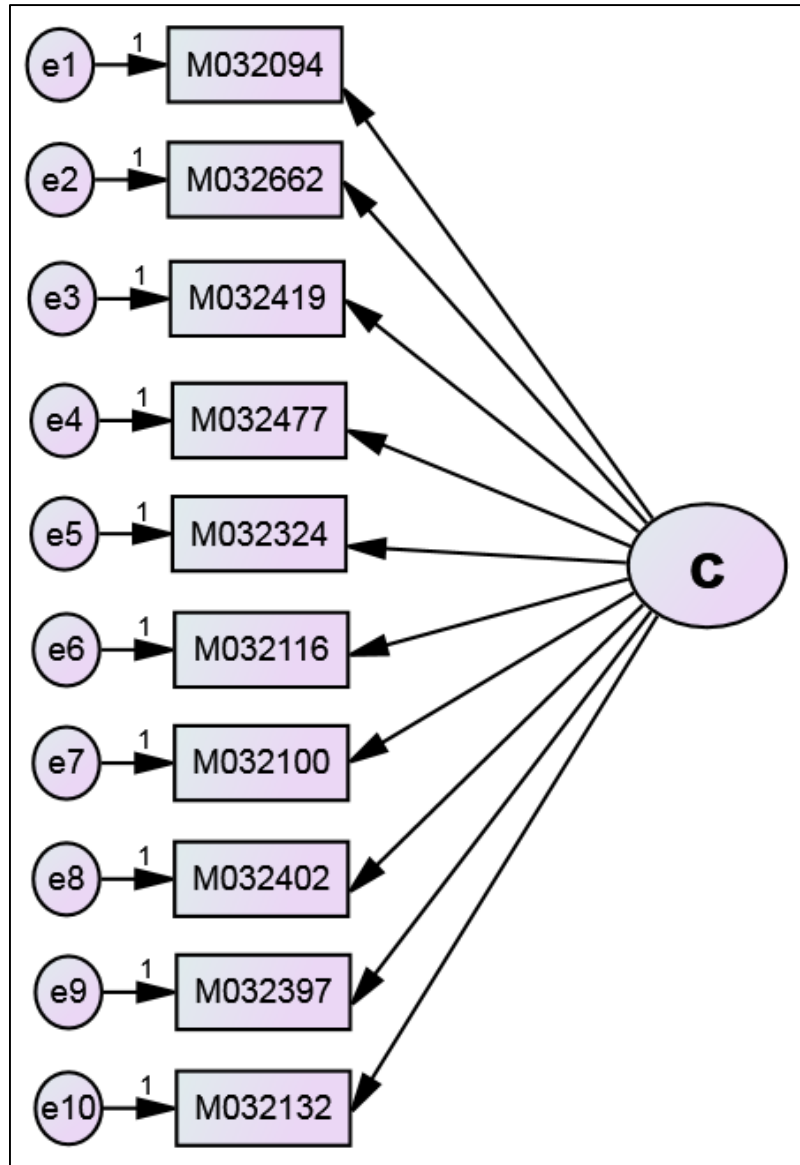
```
Mplus VERSION 7.11
MUTHEN & MUTHEN
05/17/2016 11:26 AM

INPUT INSTRUCTIONS

Title:
Booklet 1 2 Class Solution Latent Class Analysis.
Data:
File is Booklet1 2 class.dat;
Variable:
names          = IDSTUD M066 M021 M026 M095 M173 M016 M028 M014 M073 M002 M084 M029;
usevariables = M066 M021 M026 M095 M173 M016 M028 M014 M073 M002 M084 M029;
categorical = M066 M021 M026 M095 M173 M016 M028 M014 M073 M002 M084 M029;
classes = c(2);
Analysis:
Type=mixture;
Plot:
type is plot3;
series is M066 (1) M021 (2) M026 (3) M095 (4) M173 (5) M016 (6)
          M028 (7) M014 (8) M073 (9) M002 (10) M084 (11) M029 (12);
Savedata:
file is booklet1_2class_save.txt ;
save is cprob;
format is free;
output:
tech11 tech14;
```

## Appendix D

LCA model for Booklet Four (Amos Version 22)





## Appendix E

LCA 3 Class Model Specification for Booklet Four (Other Classes Similar) (*Mplus*

Version 7.11)

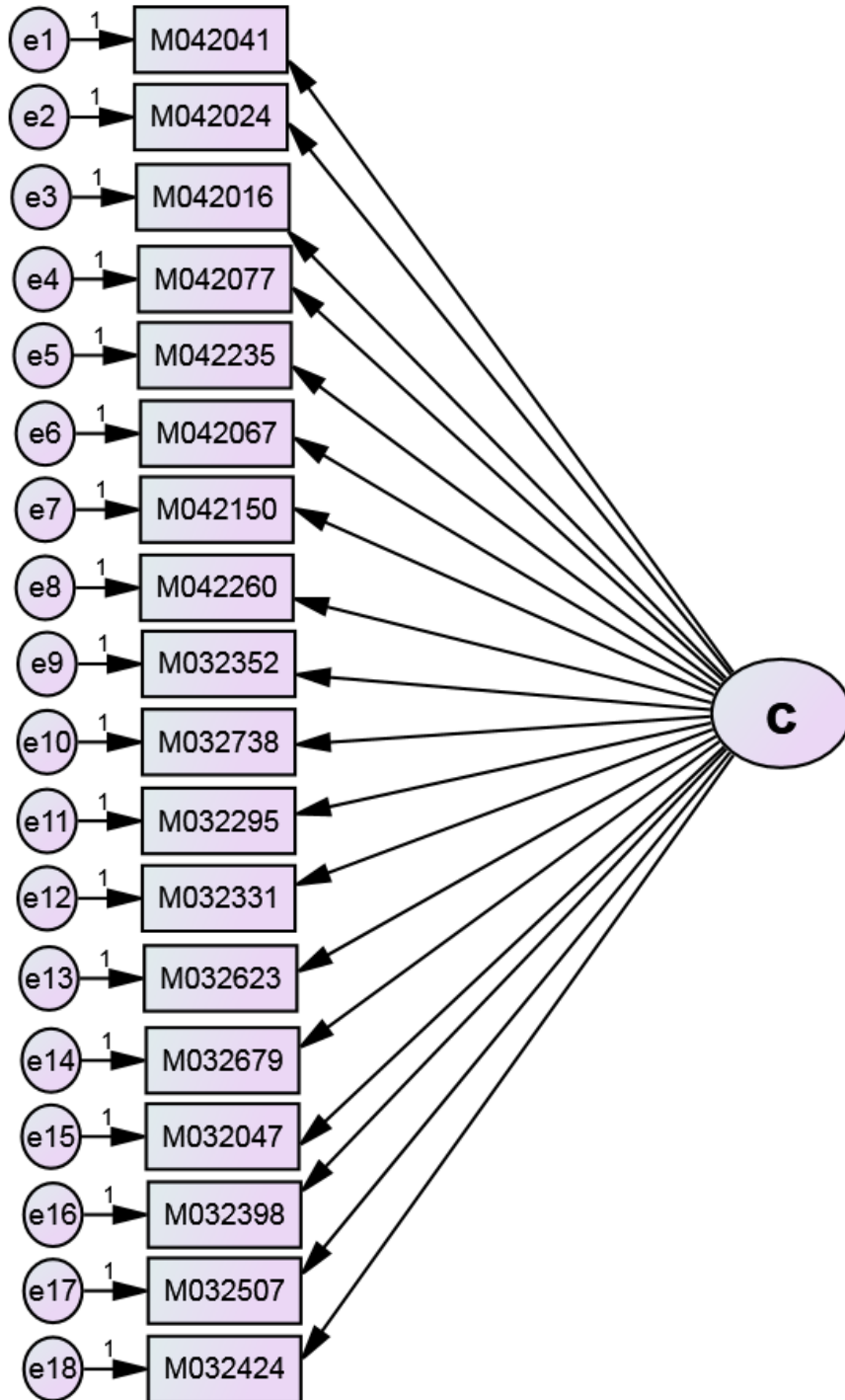
```
Mplus VERSION 7.11
MUTHEN & MUTHEN
05/17/2016 11:29 AM

INPUT INSTRUCTIONS

Title:
Booklet 4 3 Class Solution Latent Class Analysis.
Data:
  File is Booklet 4 3 Class.dat;
Variable:
  names          = IDSTUD M094 M062 M019 M077 M024 M016 M000 M002 M097 M032;
  usevariables   = M094 M062 M019 M077 M024 M016 M000 M002 M097 M032;
  categorical    = M094 M062 M019 M077 M024 M016 M000 M002 M097 M032;
  classes       = c(3);
Analysis:
  Type=mixture;
Plot:
  type is plot3;
  series is M094 (1) M062 (2) M019 (3) M077 (4) M024 (5) M016 (6)
           M000 (7) M002 (8) M097 (9) M032 (10);
Savedata:
  file is booklet4_3class_save.txt ;
  save is cprob;
  format is free;
output:
  tech11 tech14;
```

## Appendix F

LCA model for Booklet Six (Amos Version 22)



## Appendix G

LCA 4 Class Model Specification for Booklet Six (Other Classes Similar) (*Mplus*

Version 7.11)

```
Mplus VERSION 7.11
MUTHEN & MUTHEN
05/17/2016 11:32 AM

INPUT INSTRUCTIONS

Title:
Booklet 6 4 Class Solution Latent Class Analysis.
Data:
File is Booklet 6 4 Class.dat;
Variable:
names          = IDSTUD M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18;
usevariables = M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18;
categorical    = M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18;
classes = c(4);
Analysis:
Type=mixture;
Plot:
type is plot3;
series is M1 (1) M2 (2) M3 (3) M4 (4) M5 (5) M6 (6)
          M7 (7) M8 (8) M9 (9) M10 (10) M11 (11) M12 (12) M13 (13)
          M14 (14) M15 (15) M16 (16) M17 (17) M18 (18);
Savedata:
file is booklet6_4class_save.txt ;
save is cprob;
format is free;
output:
tech11 tech14;
```

## Appendix H

### Turkish Translation of Acknowledgements

#### (Türkçe Teşekkür Sayfası)

Doktora eğitimim boyunca yanımda olan ve benden desteklerini esirgemeyen herkese teşekkürlerimi sunarım. Özellikle şu kısa ömrüm boyunca hayatıma anlam kazandıran, kendi kariyerini feda edip birlikte bu başarıları kazanmamıza ön ayak olan ve çalışmam süresince beni yalnız bırakmayıp bana rehberlik yapan eşim Pınar Toker, sana ne kadar teşekkür etsem de yetmez. Varlıklarıyla evimizin neşesi olan kızlarım Erem Leyl ve Hazel Hacer'e öncelikle anlayışları dolayısıyla teşekkür ederim.

Bunun yanında değerli tez danışmanım Kathy Green'e mastır ve doktora öğrenimim boyunca sağladığı katkılar dolayısıyla teşekkür ederim. Ayrıca eğitim hayatımın başında müşerref olduğum çok değerli ilkokul öğretmenim Ayşegül Özat'a da ekmiş olduğu tohumlar ve o tohumların yeşermesi için göstermiş olduğu sevgi, sabır ve ilgi için çok teşekkür ederim.

Hassaten, geceler boyunca uykusuz kalarak dualarıyla beni yalnız bırakmayan, hayatımın her aşamasında koşulsuz desteğini yanımda hissettiğim annem Emine Kumru'ya da en kalbi şükranlarımı sunarım. Ayrıca birlikte acı, tatlı günler geçirdiğim değerli kız kardeşlerim Cemile ve Duygu'ya da teşekkür ederim.

Ayrıca Milli Eğitim Bakanlığı Yüksek Öğrenim Genel Müdürlüğü ve New York Eğitim Ateşeliği çalışanlarına da teşekkürü bir borç bilirim. Bilhassa, ülkemın renkleri olan Yüce Türk Milleti'nin Edirne'den Kars'a Türkmen'iyle, Kürt'üyle, Laz'ıyla,

Çerkez' iyle, Yörük' üyle, Avşar' ıyla, Roman' ıyla her bir ferdine en derin duygularla teşekkürlerimi sunarım. Sizlerin vergileriyle kazandığım bu başarının ülkeme faydalı hizmetlere dönüşmesi umuduyla...