UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
An investigation of the relationship of student performance to their opportunity-to-learn in PISA 2012 mathematics: The case of Indonesia

Permalink
https://escholarship.org/uc/item/7jg468vz

Author
Wihardini, Diah

Publication Date
2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

An Investigation of the Relationship of Student Performance to their Opportunity-to-learn in PISA 2012 Mathematics: The Case of Indonesia

By

Diah Wihardini

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Sophia Rabe-Hesketh
Professor Bruce Fuller
Professor Nicholas Jewell

Fall 2016

An Investigation of the Relationship of
Student Performance to their Opportunity-to-learn
in PISA 2012 Mathematics: The Case of Indonesia

Copyright 2016

by

Diah Wihardini

Abstract

An Investigation of the Relationship of
Student Performance to their Opportunity-to-learn
in PISA 2012 Mathematics: The Case of Indonesia

By
Diah Wihardini

Doctor of Philosophy in Education
University of California, Berkeley

Professor Mark Wilson, Chair

Consisting of three chapters, this dissertation investigates the possibilities for using data from a large-scale international test (specifically the PISA 2012 test in mathematics) for designing effective policies in education, specifically for developing countries, like Indonesia. The first chapter investigated the internal structure of the PISA 2012 math test and examined how the outcomes were related to the high-stakes 9th grade national examination in mathematics by using common persons linking. In addition, the associations of selected student- and school-level background variables were examined to better explain the variability of student mathematics performance across four content areas included in the test. Next, the second chapter proposed a new model for measuring the students' opportunity-to-learn (OTL) based on provided indicators of the mathematics classroom learning experience. The best-fit model was then chosen due to its strength in explaining the variability of student performance. Finally, the last chapter investigated selected student and school background information to provide insights on how they were associated with aspects of the proposed OTL measures before and after a school clustering effect was added. The data analyses were performed using a multidimensional partial credit model with latent regression based on the Indonesian data. The findings revealed that both the mathematics test and the OTL measures were multidimensional. No correlation was found between the 9th graders' performance in PISA and in the national examination in mathematics. Amongst the selected background variables, SES and grade levels showed stronger associations with the student performance and the OTL level across their sub-domains or aspects, respectively. Students with higher SES were found to attend public schools that offered better OTL. Incorporating the school clustering effect, it was shown that the students' OTL varied relatively more between schools, rather than within schools. In conclusion, this dissertation provides useful information on how the Indonesian students performed in the assessment framework laid out by PISA mathematics and how they perceived their classroom learning experience as indicators of their OTL. This information can be used to leverage and target evidence-based policy decisions to improve the quality of the national education system.

To my mother and late father.

Table of Contents

Contents……..……..... ii
List of Figures……..……..... v
List of Tables……..……..... viii
Acknowledgement……..……..... xi
Executive Summary……..……. xii
Introduction……..……..... xvi

1 Multidimensional Rasch Analysis of Indonesian Students' Performance on PISA 2012 Mathematics

1.1 Introduction……..…… 1
1.2 Background……..……..... 3
1.2.1 Brief description of PISA 2012……. 3
1.2.2 Brief description of the National Education system in Indonesia…. 6
1.3 Data sample……. 9
1.3.1 Description of PISA 2012 items…… 9
1.3.2 Description of selected background variables……..... 10
1.3.3 Description of the National Examination items and data…….. 14
1.4 Methods……... 15
1.5 Findings and Discussion……. 22
1.5.1 Research Aim 1 – Multidimensionality of PISA 2012 Mathematics 22
1.5.2 Research Aim 2 – Effects of background variables using latent regression…….. 30
1.5.3 Research Aim 3 – Linking with the 2012 9th grade National Examination in mathematics…….. 34
1.6 Limitations and Future Directions…….. 37
1.7 Conclusion…….…. 38
Appendix A.1 Technical Notes on Booklet Effect…….. 41
Appendix A.2 Item Classifications…….. 42

2 Unpacking the Opportunity-to-learn Measures in PISA 2012: The Case of Indonesian Students

2.1 Introduction……. 50
2.2 Definitions of Opportunity-to-learn (OTL)……. 51
2.2.1 Importance of Opportunity-to-learn…….. 51
2.2.2 Measures of Opportunity-to-learn……. 54
2.2.3 Opportunity-to-learn in PISA 2012……... 60
2.3 Research Objectives…… 63

2.4 Data Sample…… 65
2.5 Methods……... 65
2.6 Findings and Discussion……. 68
2.6.1 Research Aim 1 – PISA-defined measures of opportunity-to-learn. 68
2.6.2 Research Aim 2 – Proposed alternative models……... 74
2.6.2.1 Alternative Model 1……... 74
2.6.2.2 Alternative Model 2……... 75
2.6.2.3 Alternative Model 3……... 78
2.6.2.4 Results from the Alternative Models……. 83
2.6.3 Discussion…….. 86
2.6.3.1 Item fit……... 86
2.6.3.2 Reliability…….. 87
2.6.3.3 Evidence of validity…….. 87
2.7 Limitations and Future Directions…….. 100
2.8 Conclusion……... 101
Appendix B Description of Opportunity-to-learn Related Items in PISA 2012…….. 103

3 Investigating Effects of Background Variables on the Indonesian Students Opportunity-to-learn in PISA 2012 Mathematics
3.1 Introduction……. 109
3.2 Concept of Opportunity-to-learn (OTL) …… 110
3.2.1 Measures of opportunity-to-learn in PISA 2012……... 112
3.2.1.1 Aspect of Content Exposure (CE)……. 114
3.2.1.2 Aspect of Direct Instruction (DI)……..……. 115
3.2.1.3 Aspect of Student-oriented Instruction (SI)…..…….. 115
3.2.1.4 Aspect of Higher-order Assessment (HA)……. 116
3.2.1.5 Aspect of Teacher Support/Feedback (SF)…… 118
3.2.2 Background information related to the opportunity-to-learn measures……... 119
3.3 Research Objectives……..... 122
3.4 Brief Description of Education System in Indonesia…….. 122
3.5 Data Sample……..... 125
3.5.1 Description of opportunity-to-learn related items in PISA 2012….. 125
3.5.2 Description of the selected background variables……..... 127
3.6 Methods……..... 132
3.7 Findings and Discussion……..... 136
3.7.1 Research Aim 1 – Effects of background variables……... 136
3.7.2 Research Aim 2 – School clustering effect…… 142
3.8 Limitations and Future Directions……..... 148

3.9 Conclusion……..... 149

References……..... 152

List of Figures

Figure	1.1	Descriptive figures of the Indonesian students participating in PISA 2012: (a) the distribution of the students' SES across different school locations, (b) the scatter plot of the students' math scores (from 1 plausible values) against their socio-economic and cultural status, and (c) the distribution of the students' math scores across different school type and locations…….	13
Figure	1.2	Illustration of unidimensional model defined for PISA 2012 mathematics domain……...	17
Figure	1.3	Illustration of the consecutive model approach defined for PISA 2012 mathematics subscales……...	18
Figure	1.4	Illustration of between-item four-dimensional model defined for PISA 2009 mathematics domain……..……..	18
Figure	1.5	A Wright Map of the unidimensional model: (a) using the international pooled sample of countries taking the standard booklets (each 'X' represents approximately 2158 cases) and (b) using the Indonesian sample (each 'X' represents 34 cases)……..	25
Figure	1.6	A Wright Map of the between-item 4-dimensional DDA-adjusted model using the Indonesian student scores in PISA 2012 mathematics……	29
Figure	1.7	Illustration of the between-item 2-dimensional model proposed to relate the 9th grade National Examination (NE) math results with the mathematics performance as defined by PISA 2012……	35
Figure	1.8	A Wright Map of the between-item 2-dimensional DDA-adjusted model using the Indonesian 9th grade student scores in the 2012 national math examination and the PISA 2012 mathematics…….	36
Figure	2.1	The framework of Classroom Learning Assessment Scoring System – Secondary (CLASS-S) instrument, as adapted from Allen et al. (2013, p. 77)……	60
Figure	2.2	Illustration of (a) the unidimensional, (b) consecutive models, (c) between-item 3-dimensional model, and (d) between-item 5-dimensional model for measuring the opportunity-to-learn in an effective classroom learning environment as defined in PISA 2012..	70
Figure	2.3	Illustration of the simple structural path model of the PISA-defined aspects of opportunity to learn having the Indonesian students' math performance in PISA 2012 as the effect variable……..	74
Figure	2.4	Illustration of between-item 4-dimensional Alternative Model 1 proposed for measuring OTL in an effective classroom learning environment……	75
Figure	2.5	Illustration of between-item 4-dimensional Alternative Model 2 proposed for measuring OTL in an effective classroom learning	76

environment. ……..
Figure 2.6 Illustration of between-item 5-dimensional Alternative Model 3 proposed for measuring OTL in an effective classroom learning environment. …….. 79
Figure 2.7 Illustration of between-item 6-dimensional model, consisting of the 5-dimensional Alternative Model 3 of OTL and unidimensional Math model, proposed for relating OTL in an effective classroom learning environment to the student's math literacy skills assessed in PISA 2012. ……. 86
Figure 2.8 A Wright-map of the between-item 5-dimensional Alternative Model 3……... 88
Figure 2.9 The Wright Map of the 1st dimension – *Content Exposure* aspect – of the between-item 5-dimensional Alternative Model 3……... 91
Figure 2.10 The Wright Map of the 2nd dimension – *Direct Instruction* aspect – of the between-item 5-dimensional Alternative Model 3……... 93
Figure 2.11 The Wright Map of the 3rd dimension – *Student-oriented Instruction* aspect – of the between-item 5-dimensional Alternative Model 3 …….. 95
Figure 2.12 The Wright Map of the 4th dimension – *Higher-order Assessment* aspect – of the between-item 5-dimensional Alternative Model 3…. 97
Figure 2.13 The Wright Map of the 5th dimension – *Teacher's Support/Feedback* aspect – of the between-item 5-dimensional Alternative Model 3. …….. 99
Figure 2.14 Illustration of the simple structural path Alternative Model 3 of the OTL aspects having the Indonesian students' math performance in PISA 2012 as the effect variable……. 100

Figure 3.1 Illustration of the original definition of opportunity-to-learn (OTL) aspects in PISA 2012 and their relationships with the proposed measures……..…... 114
Figure 3.2 Illustration of the proposed measures of the inter-related aspects of opportunity-to-learn (OTL) using PISA 2012 survey items……... 119
Figure 3.3 Descriptive figures of the Indonesian students participating in PISA 2012: (a) the distribution of the students' SES across different school locations, (b) the scatter plot of the students' math scores (from 1 plausible values) against their socio-economic and cultural status, and (c) the distribution of the students' math scores across different school type and locations……. 131
Figure 3.4 Distribution of the sampled students' socio-economic status (SES) who attended rural and non-rural schools……... 141
Figure 3.5 Mean raw score distribution of the students' perspectives on the Content Exposure (CE) aspect with respect to grade levels……... 144
Figure 3.6 Mean raw score distribution of the students' perspectives on the Direct Instruction (DI) aspect with respect to grade levels…… 145
Figure 3.7 Mean raw score distribution of the students' perspectives on the 146

Student-oriented Instruction (SI) aspect with respect to grade levels
Figure 3.8 Mean raw score distribution of the students' perspectives on the Higher-order Assessment (HA) aspect with respect to grade levels.. 146
Figure 3.9 Mean raw score distribution of the students' perspectives on the Teacher support/Feedback (SF) aspect with respect to grade levels.. 147

List of Tables

Table	1.1	Cluster rotation design of standard test booklets in PISA 2012 as adapted from OECD (2014)…….	6
Table	1.2	Performance of Indonesian students in PISA mathematics…….	7
Table	1.3	Frequency distribution of Indonesian students across standard booklets in PISA 2012……..	10
Table	1.4	The distribution of Indonesian students across gender and grade levels in PISA 2012. Grade 10 is the modal grade (bolded) for both Indonesia and International data……..	11
Table	1.5	The percentage of sampled student distribution across secondary school types, programs, and locations in PISA 2012, calculated with respect to the whole country sample……	12
Table	1.6	Comparison of model runs on PISA 2012 mathematics performance of Indonesian students……..	20
Table	1.7	Corrected correlation matrix of the 4-dimensional (unconditional and conditional) non-DDA adjusted models……	27
Table	1.8	Regression coefficients and effect sizes of the covariates being controlled for by the DDA-adjusted 4-D latent multiple regression model……	32
Table	A.1.1	The standard booklet parameter estimates in comparison with the published values in the PISA 2012 Technical Report (OECD, 2014).	41
Table	A.2.1	Item classification for PISA 2012 math items included in the standard booklets, as modified from OECD (2014). The bolded items are polytomous items, while the rest is dichotomous…….	42
Table	A.2.2	Potential item mapping of the 9th grade national examination in mathematics across the PISA 2012 math content areas……...	46
Table	2.1	Kurz's opportunity-to-learn (OTL) indices, as adapted from Kurz et al. (2014)…….....	58
Table	2.2	Definition of opportunity-to-learn (OTL) aspects in PISA 2012…….	61
Table	2.3	Comparison of models using 62 opportunity-to-learn related items as defined in PISA 2012……...	71
Table	2.4	Correlation (disattenuated) matrix of the 5-dimensional opportunity-to-learn model…….....	72
Table	2.5	Correlation (disattenuated) matrix of the DDA-adjusted 5-dimensional opportunity-to-learn model with the student math literacy as one single dimension…..……..	72
Table	2.6	Structural path model for the 5-dimensional opportunity-to-learn (OTL) model using two-stage least squares having the Indonesian students' math performance in PISA 2012 as the effect variable……	73

Table	2.7	Items related to the Teacher's Emotional Support (ES) dimension….	76
Table	2.8	Items related to the Instructional Support (IS) dimension……...	77
Table	2.9	Items related to the Classroom Organization (CO) dimension……	78
Table	2.10	Items related to the Direct Instruction (DI) dimension……	80
Table	2.11	Items related to the Student-oriented Instruction (SI) dimension……	81
Table	2.12	Items related to the Higher-order Assessment (HA) dimension……..	82
Table	2.13	Items related to the quality of teacher's support/feedback (SF) dimension…….	82
Table	2.14	Comparison of the alternative models using 57 opportunity-to-learn related items as defined in PISA 2012…….	83
Table	2.15	Correlation (disattenuated) matrix of the 5-dimensional opportunity-to-learn model developed from the Alternative Model 1 and Math….	84
Table	2.16	Correlation (disattenuated) matrix of the 5-dimensional opportunity-to-learn model developed from the Alternative Model 2 and Math….	85
Table	2.17	Correlation (disattenuated) matrix of the 6-dimensional opportunity-to-learn model developed from the Alternative Model 3 and Math….	85
Table	2.18	Regression output of the structural path model for the Alternative Model 3 having the Indonesian students' math performance in PISA 2012 as the effect variable……	85
Table	2.19	Covariance matrix of the 5-dimensional DDA-adjusted* Alternative Model 3……	99
Table	B.1	Itemset ST61 related to Content Exposure Aspect – Experience with Math Task sub-aspect……...	103
Table	B.2	Itemset ST62 related to Content Exposure Aspect – Familiarity with Math Concepts sub-aspect……	103
Table	B.3	Itemset ST73, ST74, ST75, and ST76 related to Content Exposure Aspect – Exposure to Types of Math Tasks sub-aspect……...	104
Table	B.4	Itemset S79 related to Teaching Practices Aspect……...	106
Table	B.5	Itemset ST77 related to Teaching Quality Aspect – Teacher's Emotional Support Sub-aspect…….	107
Table	B.6	Itemset ST81 related to Teaching Quality Aspect – Disciplinary Climate Sub-aspect……...	107
Table	B.7	Itemset ST80 related to Teaching Quality Aspect – Cognitive Activation Sub-aspect……..	107
Table	3.1	Items related to the Direct Instruction (DI) aspect……...	116
Table	3.2	Items related to the Student-oriented Instruction (SI) aspect……...	116
Table	3.3	Items related to the Higher-order Assessment (HA) aspect…….	117
Table	3.4	Items related to the quality of teacher's support/feedback (SF) aspect……	118
Table	3.5	The distribution of opportunity-to-learn related items (coded as ST61 – ST80) based on the rotational design of the PISA 2012 Student Questionnaire……..	126

Table	3.6	List of student and school background variables used in the data analysis…….....	127
Table	3.7	The distribution of Indonesian students across gender and grade levels in PISA 2012. Grade 10 is the modal grade (bolded) for both Indonesia and International data……..	128
Table	3.8	The percentage of sampled student distribution across secondary school types, programs, and locations in PISA 2012, calculated with respect to the whole country sample…….....	129
Table	3.9	The percentage of sampled schools that offered additional math lessons in PISA 2012……	132
Table	3.10	Effect sizes of the multidimensional latent regression with partial credit model (DDA-adjusted) of the Opportunity-to-learn measures in PISA 2012 addressing Research Aim 1 (N = 4800)……	138
Table	3.11	Effect sizes of the multidimensional latent regression with partial credit model (DDA-adjusted) of the Opportunity-to-learn measures in PISA 2012 addressing Research Aim 2 (N = 5477)……	143
Table	3.12	The change in proportion of explained variance for the multidimensional latent regression with partial credit models across the opportunity-to-learn aspect/dimension……	147

Acknowledgement

I would like to express my sincere gratitude to my academic advisor, Professor Mark Wilson, for his never-ending assistance and support throughout the course of my graduate study in the Quantitative Methods and Evaluation (QME) program at UC Berkeley. Without his patience and guidance, this dissertation would have been neither enjoyable nor of any benefit to myself or others.

I wish to thank Fulbright, who has provided the opportunity and funding for coming to the United States to study at UC Berkeley, and also to the Berkeley Evaluation and Assessment Research (BEAR) Center for funding, exposure to real world applications of knowledge, and valuable research work experience.

To Professor Bruce Fuller, Professor Sophia Rabe-Hesketh, and Professor Nicholas Jewell, I thank you for the excellent teaching, academic aspiration, mentorship and valuable learning experiences at UC Berkeley. To Dr. Karen Draney – my project manager and supervisor at the BEAR Center, I thank you for providing mentorship and encouragement to effectively fulfill my roles as a mother, a graduate student, and a research assistant.

To my previous academic supervisors: Dr. Kelvin Gregory and Dr. Michael Teubner, I thank you for laying down a good foundation and introducing the passion for learning during my journey of academic capacity-building in Australia. I finally made it!

I also wish to express my special appreciation and thanks to my great friends: Katherine Castellano, Laura Pryor, Perman Gochyyev, and Yukie Toyama as well as Michela Azzariti, Ria Umar, Yvette Leung, and their families for their friendships and continuous supports during the ups-and-downs throughout the years in Berkeley. My special thanks are also due to my colleagues: Amy Arneson, James Mason, Jin Ho Kim, Joon Ho Lee, Leah Feuerstahler, and many others for continual help and for making the academic rigor at Berkeley both bearable and enjoyable.

I am fully indebted to my beloved mother for always having the time and energy to commute between Indonesia-Australia-North America for the past ten years and for her unconditional love and continuous prayers. Matur nuwun, Ibu! Also, to my parents-in-law who have offered support and prayers for my family's success in studying and living abroad, I thank you.

Last, but not least, to my husband – Henri Jufry, whose love, patience, understanding, and support are very much appreciated, I give my deepest thanks and love. I truly appreciate your sacrifices in supporting me to achieve my dream for all these years. To my daughters: Kirana and Nayra, I thank you for having been patiently (or not) watching your mother constantly working in front of the computer since you were born! I would not know how I would have survived this endeavor without you in my life. My time is all yours now.

Executive Summary

You must be the change you wish to see in the world.
(Gandhi)

This dissertation began with the primary aim to illustrate the utility of PISA outcomes for designing effective policy tools in education, specifically for developing countries, such as Indonesia. Techniques introduced in this dissertation can give examples of how to increase the utility of participation in the high cost and rigorous international large-scale assessment program by utilizing the results for in-country development. Having sound and solid understanding of how the students performed and why such performance varied, can allow one to promote more equitable and higher educational opportunities to improve student achievement, which in turn can leverage the quality of human resources as needed for national development. For the purposes of the dissertation, I used the PISA 2012 mathematics data from Indonesian students and employed the MRCMLM framework (Adams, Wilson, & Wang, 1997) to implement the explanatory item response modelling approach in addressing the research aims.

Summary

In Chapter 1, I investigated the internal structure of the PISA 2012 mathematics test, the effects of selected student- and school-level background information on the latent mathematics ability across the content areas, and the correlation between the 9th grade national examination and the PISA outcomes in mathematics. The findings include the following points. First, the postulated multidimensional structure of the PISA 2012 mathematics overarching four content areas, i.e. *Space and Shape*, *Change and Relationship*, *Quantity*, and *Uncertainty and Data* as intended (OECD, 2014), is supported. The provision of sub-scores representing the students' mathematics literacy in each of the content areas was found to be justified. Using Wright maps, I also pointed out in which content areas the Indonesian students were lacking. Second, statistically significant effects of most of the selected background variables, such as SES, gender, grade levels, and school type and location, were found after applying a latent regression on the students' scores. In particular, this study revealed that high grade-retention[1] might promote an adverse effect to the students' academic success. Finally, there was no correlation found between the 9th graders' latent mathematics ability represented by PISA and the one represented by the national examination. This I interpreted as being due to the almost perfect passing rate of the national examination, i.e. there is almost no variability in the scores.

[1] The 15-year-olds Indonesian students included in the sample were distributed across 7th to 12th grade, which were mostly due to strict grade progression.

As past studies have indicated that good student learning experiences lead to success in student achievement (Gamoran, 1987; Stevens & Grymes, 1993; Floden, 2002; Schmidt & Maier, 2009), in Chapter 2 I evaluated the definition and measurement of the aspects of students' opportunity-to-learn (OTL) as defined by PISA 2012. To better explain the variability of the Indonesian students' performance, I then proposed an alternative approach to measure the OTL aspects, but still used the items provided. The proposed OTL measures consist of such aspects as *Content Exposure*, *Direct Instruction*, *Student-oriented Instruction*, *Higher-order Assessment*, and *Teacher Support/Feedback*. These measures are derived from effective classroom learning factors as identified in the literature (see Hattie (2009), Kurz (2011), Allen et al. (2013), and OECD (2014)). I found that the aspects of the proposed OTL measures could explain the variability of student performance better than the other possible OTL measures being evaluated in the chapter. Furthermore, the Wright maps suggested that in the typical math lessons, (1) there was a lack of math word problems being exposed, (2) teacher-dominant practices were common, (3) there was a lack of task differentiation based on the students' learning capacity, (4) the stated cognitively-activated learning opportunities happened infrequently, and (5) teachers did not praise and point out students' strength and weaknesses often enough.

The proposed model of OTL measures, then, became the foundation for the base model used in Chapter 3 for investigating the associations of selected student- and school-level background variables on these aspects of OTL. The selected background variables included gender, SES, grade levels, school type, program, and location, availability of additional math lessons offered by the school, and proportion of certified teachers. Statistically significant associations were found between the OTL aspects and the grade-related variables. Grade retention, if not accompanied by an increase in instructional support, was associated with reduced exposure to expected curriculum contents, and thus may lead to an adverse effect on student learning (McCoy & Reynolds, 1999; Silberglitt, Appelton, Burns, & Jimerson, 2006; Allen et al., 2009). Similar with findings from Chapter 1, SES was also found to be statistically significant and positively associated with most aspects of OTL, suggesting that students with high SES were more likely to attend public schools that provided good opportunities to experience and be exposed to the prescribed mathematics contents and types of problems/tasks, and a learning environment that supported direct instructional strategies and higher-order assessment/school work. Furthermore, after incorporating the school clustering as fixed effects in the latent regression model, the results show that the Indonesian students' OTL did indeed vary strongly between schools and that there was a considerable variability in the implemented curriculum between schools.

Limitations and Future Directions

Findings from this dissertation should be considered in light of several limitations, which eventually lead to an avenue of future research. Some of the limitations and future research are highlighted as follows.

First, the person/school clustering as a result of the two-stage sampling has not been fully taken into account in the investigation of how the background variables correlated with

the students' mathematics literacy estimates across the four content areas, although such approach was implemented across the OTL aspects by incorporating school clustering as fixed effects. In further work, one could expand the multidimensional latent regression model to a three-level hierarchical linear model that takes into account the school clustering effect better in estimating the effects of background variables on both the students' math ability and their OTL, while better handling of the complex data structure due to the matrix sampling design of the test and survey forms.

Second, the use of anchoring item parameters from the international calibration sample for estimating the students' math ability has produced misfit. More investigations on how and why the models show misfit are warranted to give a better rationale behind the low scores for a particular set of test-takers.

Next, the limited sample of students who took both PISA test and the national examination in the same year might have contributed to ineffectiveness of the common person linking approach to provide acceptable criterion-related validity evidence of the PISA test on Indonesian students. Finding a larger and more representative sample of these link students could help increase the utility of the PISA test along with the associated vast amount of background variables. Although PISA does not assess school curriculum, benchmarking any national examination results with the PISA outcomes can give essential information on how comparable the local education system is with global results.

Furthermore, the OTL-related indicators were obtained from student self-report that can potentially be subject to social desirability bias, specifically if the sampled students systematically inflate or deflate their perspectives on their own classroom learning environment. Although the 15-year-old students were asserted to be mature and experienced enough to recall their classroom learning experiences (Schmidt, Zoido, & Cogan, 2013), one should note that the students' rating of the OTL-related items might also represent their accumulative perceptions of OTL throughout their schooling. To strengthen the students' self assessment of their OTL, albeit being time and cost intensive, a structured or semi-structured observational survey of classroom interaction could be proposed to better portray the effectiveness of the learning environment. Another future study could also explore other techniques that can effectively overcome potential bias due to the effects of social-desirability on items.

Finally, the cross-sectional nature of PISA data provides a limitation as the timing of the one-time data collection is not guaranteed to be representative. Therefore, the PISA outcomes should always be interpreted with caution as they might not fully represent the whole picture of the students' performance and learning environment being investigated. Should it be required to assess the development or improvement of the student performance and OTL over a period of time for a national policy evaluation, for instance, expanding it to a longitudinal data collection by tracking some subsets of students or schools within a country on each round of PISA may be an option. By doing so, the effectiveness of a curriculum and/or school policy reform can be assessed better.

Implications

Despite of the limitations as described above, the reported findings have the following major implications for the evaluation and redesign of educational policy reforms in Indonesia.

First, the results can be seen as contributing to development of the national mathematics curriculum at all grade levels, as successful outcomes on the PISA's four overarching mathematics content areas would depend upon the strength of their foundational knowledge taught in earlier years. The multidimensional analysis at item level can provide substantial insights on the difference in the student performance across content areas. The relative differences in the student performance across the mathematics content areas can be utilized to improve the curriculum structure and/or the content pedagogical knowledge on the seemingly more difficult concepts. In-depth descriptions of student performance on each assessed construct may influence the assessment framework and inform changes in the instructional and pedagogical content of the teachers. A similar implication can also be derived from the development of the multidimensional OTL measures. The relative differences in the students' perspectives of their OTL aspects as illustrated by the students' responses to some particular OTL-related items can imply ways to better allocate resources to improve the students' learning environment, which in turn may influence their academic achievement.

Second, by analyzing the effect of background variables on student performance as well as their association with the students' OTL, the diverse socio-economic gaps and vast differences in local communities can be seen as important hurdles for the implementation of the curriculum for students of all backgrounds. The relative differences of the effects across the mathematics content areas and the OTL aspects can then be utilized to hypothesize ways to improve the curriculum structure, content pedagogical knowledge, and school delivery standards, particularly since only 30% of schools in Indonesian has achieved the national delivery standards (Kompas, 2016b).

Third, utilizing PISA outcomes and relating them with the national examination outcomes at 9^{th} grade, or even 12^{th} grade level, may become a sound policy developmental tactic as it provides criterion-related validity evidence for justifying the cost and benefit of participating in the international large-scale assessment and validating the needs for rigorous education reforms. Furthermore, many critics of the new curriculum change and even those who are in support of the moratorium[2] on the national examinations might welcome more information on studies about the benchmarks upon which the Indonesian students' performance is based.

In conclusion, this dissertation provides useful information on how the Indonesian students actually performed in the assessment framework as laid out by PISA mathematics and perceived their classroom learning experience and environment as indicators of their opportunity to learn. These information can be used to leverage and target evidence-based policy decisions to improve the quality of the national education system.

[2] The Indonesian Minister of Education and Culture proposed a Presidential bill to halt the 2017 national examinations (Kompas, 2016a), but his proposal was then rejected by the Presidential Office (Kompas, 2016b).

Introduction

One of the key motivations for participating in a high cost and rigorous international large-scale assessment like PISA (Program for International Student Assessment) is to benefit from the abundant cross-national and local information about the "economies [of] education outcomes" (Lockheed, 2013). Such benefits should outweigh the massive human and financial resources that participation in international testing involves. The PISA test provides aggregate information about student cognitive performance on particular knowledge domains, but it also typically collects abundant contextual information obtained from students, parents, teachers, and schools. In order that the anticipated benefit of participating in PISA outweighs the cost and potential risk, specifically for developing countries with limited capacity in making evidence-based policy decisions, the interpretation of the PISA outcomes should always be kept in its proper perspective: to inform better policy decisions for improvement. Hence, for my dissertation, I aim to illustrate the utility of the PISA data to (1) provide insights on the extent to which the participating students performed across the assessed knowledge domain, (2) explain the aspects of the students' opportunity to learn the subject matter and how this opportunity can explain the variability of the student performance, and (3) investigate the association of selected background information on the students' performance and their opportunity to learn.

PISA is run every three years by the Organization for Economic Co-operation and Development (OECD) and administered across its country members and several partner countries and economies. It assesses 15-year-olds because, at this age, students have nearly completed the end of compulsory education in most of the participating countries; and some are expected to continue to higher education and/or be ready to join the labor-market under various circumstances (OECD, 2013). Student performance at this stage can also reflect the effectiveness of their learning and academic attainment in the preceding levels. Although the PISA surveys three subjects, mathematics, reading, and science, on every round, it has an alternating primary focus on each round. Having a different emphasis on each round gives time benefits for the participating countries to focus policy improvement efforts on only one single subject area and allow a sufficient time-frame for monitoring trends in their student performance. Therefore, PISA outcomes are often used to influence the decision-making processes for the improvement of national educational policies (Breakspear, 2012; OECD, 2013b; Sjoberg, 2012). Since PISA[3] 2012 focused on mathematics literacy, different areas of mathematical content knowledge and classroom learning practices were dominant in the test and the background questionnaire. My dissertation, therefore, focuses on mathematics learning.

[3] Although the latest round of PISA was held in 2015, the data has yet been made publicly available.

With the signing of the ASEAN[4] Economic Community treaty that allows free trade and exchange of work force among its country members starting in 2016, Indonesia – the fourth largest country in the world in terms of population – does not have any other choice but to ensure that its school graduates possess high quality skills to those from other ASEAN countries or even from the rest of the world. This rationale may have prompted Indonesia to fully participate in PISA as it specifically provides a broad assessment of comparative learning outcomes toward mastery of life skills, not just the mastery of school curriculum. Unfortunately, Indonesia always falls short in the PISA tests. Having great diversity and vast gaps in the socio-economic status, the Indonesian students have performed very poorly and were always ranked close to the bottom in comparison to students from other countries. For the purpose of my dissertation, I used the Indonesian data from the mathematics test to get a deeper understanding how student and school background information can explain the variability in the student performance and learning opportunity in mathematics. The findings can help provide sound evidence for the Indonesian government in making sound policy reforms to leverage the quality of its human resources.

To address the three research aims as described further in the dissertation chapters, I used the explanatory item response modelling approach, specifically the multidimensional random coefficients multinomial logit modelling (MRCMLM) approach (Adams, Wilson, & Wang, 1997) to examine the students' use of score categories within and across different content areas, calibrate the item parameters, and estimate the students' latent construct on mathematics literacy and their opportunity to learn (OTL). Furthermore, by adding latent regression approach to the MRCMLM, the student/school background information accompanying the test could be associated across the sub-domains of the mathematics literacy and the aspects of OTL. Evaluating the relative differences of the effects of such background information gives informed knowledge to hypothesize ways for improving the curriculum structure, the content pedagogical knowledge, and the school delivery standards. All data analyses were carried out on ConQuest 4 (Adams, Wu, & Wilson, 2015).

This three-paper dissertation is organized into three chapters with the following organization. First, *Chapter 1* investigated how the students' performance information can be disaggregated to the subscale level and how their background information can explain the variability of the student academic performance. In addition to the investigation of the internal structure of PISA math test for construct validity evidence, I also presented the examination of criterion-related validity evidence by investigating how the PISA outcomes relate to the high-stakes national examination results. Second, using the students' self-report perspectives on their classroom learning practices, in *Chapter 2* I re-defined and presented several possible measures of the aspects of OTL. A best-fit model for measuring OTL was then selected depending on its ability to better explain the variability of student performance, especially since I interpreted this to be a necessary condition for a successful education process. Next, *Chapter 3* investigated and discussed the utility of student and school background information to provide insights on how they were associated with the degree of OTL. By understanding deeply how the background information is correlated with each

[4] ASEAN stands for the Association of Southeast Asian Nations, which promotes intergovernmental and economic development collaboration among such country as Brunei Darussalam, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, and Vietnam.

aspect of OTL before and after accounting for the school clustering effect, specific student and/or school policy can be targeted to improve the curricular and resource allocation management, which in turn leverages the degree of students' OTL.

Chapter 1

Multidimensional Rasch Analysis of Indonesian Students' Performance on PISA 2012 Mathematics

1.1. Introduction

Explanatory item response models as a modern measurement approach can help promote the understanding of the results in participating in any series of international large-scale student assessments (ILSA). Such benefits should outweigh the massive human and financial resources such participation normally involves. One of most highly anticipated benefits is to provide reliable and valid information on how one country's student performance compares to those in other countries. Such comparison allows national policy-makers to review the effectiveness of their education system in terms of curriculum, teaching practices and school resources, and how these educational components compare to global trends. Typically, the ILSA test score represents aggregate information about student performance. Our study applies multidimensional Rasch models (Rasch, 1960) with latent regression to deconstruct these scores in terms of the assessed content areas while simultaneously controlling for student and school contextual variables. With this modelling approach, I seek to explain how students perform on each item across different subdomains and under diverse contexts. This explanatory function contributes to inherent advantages of item response models for a wide variety of evaluations (De Boeck and Wilson, 2004), specifically when related to the student's home and school backgrounds. As a result, it provides information to support targeted curricular improvements for students from many backgrounds.

This chapter aims to investigate how this performance information can be disaggregated to the subscale level and how the student/school background information accompanying the test might explain the variability of the student academic performance. In addition to the investigation of the internal structure of the ILSA test for construct validity evidence, this research will also examine criterion-related validity evidence by investigating how the ILSA outcomes relate to Indonesia's high-stake national examination results. Focusing on the Program for International Student Assessment (PISA) test in mathematics, I use data on Indonesian students from the PISA 2012 round and from the 2012 national examination. A detailed investigation on how PISA scores can help us understand the student performance, how the diverse student backgrounds may influence the performance, and how PISA is related to the student performance in a high-stakes national examination will benefit policy makers with regard to the national education policy reforms, specifically in developing countries with vast socio-economic gaps.

Indonesia – a populous and diverse country in South East Asia – has always had low performance on international assessment studies, particularly in the triennial PISA. Started in 2000, PISA measures the readiness of 15-year-olds to meet future challenges through their capacities in three major knowledge domains: reading, mathematics, and science. Despites the fact that passing rates on the national examinations (NEs) often reached almost 100% (Kompas, 2012; Tempo, 2012; Kompas, 2013; MoCEI, 2014)[1], the Ministry of Education and Culture of Indonesia (MoCEI) has often used PISA outcomes in the narrative of the education reform needs, particularly in the needs for major curriculum change in order to better prepare its younger generation for global competition (MoCEI, 2013a, 2013b; BAPPENAS, 2016). The MoCEI introduced new innovative national curriculum standards in early 2013 for 1st to 12th grade levels with a program of rigorous teacher training and provision of standardized teaching and learning materials, which has sparked controversies between the government and a wide range of community groups (Kompas, 2013a). The critics claimed that the quality of the teachers, not a curriculum change, mattered the most (Kompas, 2013b). In my opinion, both are probably needed.

As PISA also gathers a vast amount of background information on student, parent, and school, I argue that an in-depth understanding of how a student performs at the subscale level in PISA and how such performance is related to background information is crucial. This understanding can provide insights on factors that impact the development of student's cognitive abilities and their implication for national education policy improvement. Incorporating Rasch models, having student ability and PISA item difficulty consistently placed on a common scale, allows a useful comparison and thorough investigation of specific items indicative of which content area the students are lacking. In addition, relating both person and item measurement values simultaneously to important background variables in each content area can generate better explanations on how different contextual conditions have impacted student performance while accounting for multidimensional aspects of the assessed skills.

Mathematics has always been a compulsory subject taught at school from the early years and assessed heavily. In Indonesia, mathematics also becomes one of the main subjects tested in the standardized NEs at 6th grade, 9th grade, and 12th grade level. Hence, it is recognised as a must-have skill to survive the schooling years. Competency in this subject is a significant factor for gaining a successful academic and professional life. In the fifth round, PISA 2012 focused on mathematics literacy and so administered more mathematics items than those for other subjects. Using data on Indonesian students, I examine the internal structure of PISA 2012 in the mathematics test corresponding to the subscales for the four overarching PISA mathematics' content areas: *space and shape*, *change and relationship*, *quantity*, and *uncertainty and data* (OECD, 2013b, 2013c) and evaluate the criterion-related validity of the test by linking it to the 2012 9th grade NE responses. Thus, I have the following research questions, all of which are dealt with in the context of the Indonesian sample:

[1] Starting in the last 2014-2015 academic year, the stakes of these tests were reduced, as the national exit examination results were no longer considered as the sole factor for graduating students. School-owned tests have now been included in the assessment (Sindo, 2014; MoCEI, 2016).

(1) To what extent does multidimensionality exist in the student performance on each of the PISA 2012 mathematics content areas?
(2) How do several background variables such as gender, grade levels, school types and locations correlate with the student performance across the mathematics content areas while accounting for the students' socio-economic status?
(3) To what extent are the outcomes for the PISA 2012 mathematics test comparable with those for the 9th grade national examination in mathematics?

In this study, I used the multidimensional random coefficients multinomial logit model (MRCMLM) to examine the students' use of score categories within and across different content areas and calibrate the item parameters and estimate the students' ability. MRCMLM is a generalized Rasch item response model that utilizes a scoring function and a design matrix to accommodate the applications of IRT models used in this study, such as the partial credit model, the facet model, and multidimensional versions of these models (Adams, Wilson, & Wang, 1997). In addition, a multidimensional latent regression model with socio-economic status (SES), gender, grade level, school type and school locations as covariates were applied to investigate their differential effects on the student ability estimates for each content area. Furthermore, I also examine the technical question of the effect of the randomly assigned test-booklets on student estimates. Meanwhile, the link between PISA 2012 and the NE scores in junior secondary school mathematics was investigated by using the item responses of the Indonesian 9th grade students who took both PISA test and the NE test in 2012 that was administered in a close time interval.

1.2. Background

1.2.1. Brief description of PISA 2012

The Program for International Student Assessment (PISA) is run every three years by the Organization for Economic Co-operation and Development (OECD) and administered across its country members and partner countries around the globe. It assesses 15-year-olds because in most of the participating countries, at this age students have nearly completed the end of compulsory education; some are expected to continue to higher education and/or be ready to join the labor-market under various circumstances (OECD, 2013). Student performance at this stage can also reflect the effectiveness of their learning and academic attainment in the preceding levels. Although the PISA survey measures three subjects: mathematics, reading, and science on every round, it has an alternating primary focus for each round. The fifth round of PISA in 2012 focused on mathematics. Having a different emphasis on each round gives time benefits for the participating countries to focus policy improvement efforts only on one single subject area and allow a sufficient time-frame for monitoring trends in their student performance. Therefore, PISA outcomes are often used to influence the decision-making processes in the improvement of national educational policies (Breakspear, 2012).

The expected mathematics competencies in PISA are defined in terms of "mathematics literacy" (OECD, 2013c, p. 25) in which an individual has …

> the capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain and predict phenomena. It assists individuals to recognize the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens.

Hence, PISA puts an emphasis on the mastery of process, the understanding of concepts, and the application of those understandings in interpreting and solving day-to-day problems in various contexts. In order to measure the target knowledge domain, the mathematics assessment covers four overarching areas taught throughout the schooling years: *space and shape*, *change and relationship*, *quantity* and *uncertainty and data*. These areas can cover a broad domain of math proficiency to face the challenge of today's demand on knowledge (OECD, 2013c). The four overarching areas assessed in PISA originate from the work of Steen (1990) and Devlin (1994) as indicated in PISA 2012 Assessment Framework (OECD, 2013c). Although the definitions of the content domains differ from the common and typical definitions used in the mathematics instruction and curricular strands, their ideas cover a broad range of the expected proficiencies in mathematics taught in school.

These PISA key aspects of the mathematics content areas, often referred to as sub-scales, are described in the following. First, in *Space and Space* (SS), a student must be able to recognize, describe, identify, and interpret shapes and patterns in shape, understand the similarities and differences between them, and understand the dynamic changes of visual information. The changes can also occur in a system of interrelated objects or phenomena where the system members influence one another. An understanding of change and its relationship within a context constitutes the second content category called *Change and Relationship* (CR). This content, however, may overlap with other content areas in mathematics as it involves 'functional thinking', which is "one of the most fundamental disciplinary aims of the teaching of mathematics" (OECD, 2013c). The third content *Quantity* (QY) taps into the notion of quantitative reasoning which requires an understanding and recognition of numerical sense, patterns, and representations. Proficiency in comparisons, computations, and operations of numbers falls into this domain category. Finally, the *Uncertainty and Data* (UD) category is included to respond to the current demand for understanding of the uncertain phenomena not only met in scientific and technological contexts, but also in daily life problems. This content area covers the concepts of variation in processes as well as presentation and interpretation of data. Students must be able to acknowledge uncertainty and error in measurement, summarize, and draw inferences from given information of any sort. Indeed, the four content areas can intersect and overlap with one another as their differentiations are somewhat blurred, and often one would be required to use several of them all in solving one problem (OECD, 2013c). However, the PISA test developers attempted to design items that are closely related to real life problems and that tap into only one content area.

School, as the level of government education institution that is closest to the student, plays a very important role in shaping student learning. In an effective school, the prescribed national curriculum standards are translated into appropriate levels of details at which can be

taught to students while utilizing adequate infrastructure, human and financial resources (Bryk et al., 2010). Teachers must also have the appropriate content pedagogical knowledge to maintain their student engagement and motivation to learn, which drive the students to perform better. Accommodating the contextual conditions of the school and the idiosyncrasies of the student demographic mix can assist teachers in designing and delivering their instructional practices better. PISA's school survey provides vast and useful knowledge on how well a particular country's education system works and thus, on how such knowledge can be used to inform policy improvement. It can help not only in investigating differences in curricular implementations, but findings from the school survey can also help one examine the association of the school's vertical stratification, types and locations on student performance (OECD, 2013d).

In this study, vertical stratification refers to the notion of grade repetition where students cannot progress to the next grade level unless they pass some sort of assessment at the end of each grade. In PISA, the 15-year-old students can be in different grades due to either the different entry age to the school system or grade repetition/acceleration. Although the grade repetition practice aims to create a more homogeneous learning environment by having students learn at the same level and pace so that teachers can teach better, PISA has found otherwise. It has found that the school systems in participating countries with a high degree of grade repetition have lower student performance. Thus, grade repetition seems to have an adverse impact on student learning outcomes, especially because it is highly correlated with the socio-economic and demographic background of the students and the schools (Hattie, 2009; OECD, 2013e).

Using a sophisticated rotated design to extend the coverage of content areas, the PISA 2012 assessment items were grouped into different clusters before assigning them to different booklets. These booklets were categorized into standard booklets, easy booklets, and UH booklet (i.e. Une Heure booklet, one-hour booklet dedicated to students with special needs). Unless a participating country took the offer of taking the easy booklets[2], most countries that participated in PISA 2012 administered the standard booklets. As shown in Table 1.1, each standard paper-based booklet contained 4 clusters comprised of 12 – 13 items of one assessed subject in each, which were grouped in clusters labelled as PM for the seven mathematics clusters (numbered 1 to 7), PR (numbered 1 to 3) for the three reading clusters, and PS (numbered 1 to 3) for the three science clusters. Thus, each student might have a range of 12 – 37 mathematics items depending on which booklet they were randomly assigned to (OECD, 2014). Although it has been claimed that no booklet effect impacted the scaled item parameter estimates, the distribution of student ability estimates would still be affected by the random allocation of items and their associated domains in these booklets.

[2] Some countries, which showed low performance in the past PISA administration or in the field trial test, were offered to take the easy booklet (OECD, 2014).

Table 1.1
Cluster rotation design of standard test booklets in PISA 2012 as adapted from OECD (2014)

Booklet ID	Cluster			
B1	PM5	PS3	PM6A	PS2
B2	PS3	PR3	PM7A	PR2
B3	PR3	PM6A	PS1	PM3
B4	PM6A	PM7A	PR1	PM4
B5	PM7A	PS1	PM1	PM5
B6	PM1	PM2	PR2	PM6A
B7	PM2	PS2	PM3	PM7A
B8	PS2	PR2	PM4	PS1
B9	PR2	PM3	PM5	PR1
B10	PM3	PM4	PS3	PM1
B11	PM4	PM5	PR3	PM2
B12	PS1	PR1	PM2	PS3
B13	PR1	PM1	PS2	PR3

Notes. PM, PS, and PR denote a cluster of math items, science items, and reading items, respectively.

1.2.2. Brief description of the national education system in Indonesia

Having a population of around 245 million people (Trading Economics, 2012), Indonesia is the fourth largest country in terms of population in the world, covering about 17,000 islands with more than 200 ethnic groups and at least 300 different languages/dialects. With a diverse pattern of population demographics, it is not surprising that there is a large socio-economic gap and income inequality throughout the country, which in turn is closely related to student academic performance differences (CIA, 2013; OECD, 2013a). The stagnantly low performance (i.e. in the ten bottom ranks) of the Indonesian students in PISA and other international student assessment programs on all tested subjects is of a significant concern and calls for a major national education reform (see Table 1.2 for Indonesian students' performance on PISA mathematics). Please note that the mean scores of the student math performance in Table 1.2 cannot be directly compared without linking the common items from the 2000 round to the 2012 round of test. However, the students' low performance in the five consecutive PISA rounds did give bad precedence for the national education rhetoric (Kompas, 2013d). Although the latest round of PISA was held in 2015, the data has yet been made publicly available.

Table 1.2
Performance of Indonesian students in PISA mathematics

Year	Mean Score	Rank (out of total number of participating countries)
2000	367	39 (out of 41) [1]
2003	360	38 (out of 40)[1]
2006	391	50 (out of 57) [2]
2009	371	61 (out of 65)[3]
2012	375	64 (out of 65) [4]

Sources: [1] Balitbang (2016), [2] OECD (2007), [3] OECD (2010), [4] OECD (2014)

The school system in Indonesia is comprises of 3 levels: primary/elementary school (1^{st} – 6^{th} grades), junior secondary school/middle school (7^{th} – 9^{th} grades), and senior secondary school/high school (10^{th} – 12^{th} grades) (MoCEI, 2013b). Previously, the compulsory and public education was only offered for the first nine years of schooling (i.e., primary to junior secondary levels). But recently, the MoCEI introduced the formal plan for a twelve–year compulsory education program that will include free public education for the three years of senior secondary level (MoCEI, 2015). Based on the of school management, there are three categories of school programs: (1) general school – under the jurisdiction of the Ministry of Education and Culture, (2) Islamic school – under the jurisdiction of the Ministry of Religion Affairs, and (3) vocational school –under the jurisdiction of the Ministry of Education and Culture. The first two categories provide education for K-12 programs, while the vocational school only caters for the equivalent of 9^{th} to 12^{th} grade programs with an emphasis on specific vocational programs for training students with specific skills in a variety of fields such as mechanics, business, home economics, tourism, handicraft, and art. In addition, for each of the school categories – depending on the financial sources, there are two major distinctive types of school: public and private. Public schools are managed and controlled by "a public education authority, government agency, or governing board appointed by government or elected by public franchise"; whereas private schools are managed by private institutions such as religious and non-religious foundations (OECD, 2013c). From 2010/2011 data (MoCEI, 2013b), there are about 78% of public schools at junior secondary level, whilst it is only 32% of them at the senior secondary level. The rests were private schools. To help serve the poor community, the Ministry of Religion Affairs builds and manages many Islamic schools at all levels in both rural and urban areas, whose funds can be provided either fully by the government or partly by particular community organizations (MoCEI, 2013b).

Globally, private school students typically outperformed those from public schools (OECD, 2013d), but for Indonesian students, this might not be true. As MoECI (2013b) indicated, students who failed in getting admission to the public schools are more likely to enter the private schools. A similar case also happens with the vocational programs. Only those who need to seek employment and earn a living after graduation would opt for

vocational schools[3]. Hence, public schools tend to have the advantage of getting good students and thus, are more likely to pay more attention to their teaching, learning, and assessment process due to strict and rigorous monitoring from the local district offices. Using NE results for primary and secondary school levels, and three rounds of the Family Life Survey during 1997 – 2000, Newhouse and Beegle (2006) found that the Indonesian students attending public schools and non-Islamic private schools had generally performed better compared to their cohorts at the private schools, after controlling for several background variables. They asserted that the high performance of public schoolers at junior high school level was most likely due to the upward selection bias of higher quality admits. Thus this finding is associated with the grade repetition/promotion practices discussed in the previous section. Their study also found that students at city schools performed worse than those located in non-urban areas.

School progression is very restricted in Indonesia since each student must pass a final year assessment developed by local schools in order to move up the ladder. Upon completion of each of the three levels of schooling, the students must take a high-stakes NE over several nationally-defined major subjects such as mathematics, language (Indonesian and English), science at the end of 6^{th}, 9^{th}, and 12^{th} grade, respectively. Results from these NEs become the major factor for not only passing the current school level, but also for school admission at the next level. Although currently the stakes of the NE has been greatly reduced, the results are still used for admission to the best[4] schools at any level (Sindo, 2014; MoCEI, 2016). Meanwhile, the passing rates on these NEs, are almost always 100% (Kompas, 2012; Metro, 2012; Kompas, 2013c; MoECI, 2014). There was only 0.04% of junior secondary students who failed in the 2012 NE in mathematics (Metro, 2012). The striking difference between the NE results and the PISA outcomes has prompted the Indonesian government to review its national curriculum and introduce new standards (MoCEI, 2012). The development and administration of the new-yet-controversial curriculum faced significant challenges, disputes, and oppositions from the wider community (Kompas, 2013a; Kompas 2013b). Quoting the poor performance in international assessments, the Indonesian Minister of Education and Culture re-evaluated the new curriculum standards and practices, and targeted a 20% performance increase in the next round of the PISA tests (Baswedan, 2014). Having only been implemented for six months in almost all public schools and some private schools throughout Indonesia, the MoCEI halted and revisited the implementation of the new curriculum in December 2014.

Analysis of the relationship of test scores to variables external to the test would provide another important source of validity evidence (AERA, APA, & NCME, 2014). If the external variable measures hypothetically a similar construct, correlating such measures would produce a convergent evidence confirming the usefulness of the models for their defined purposes. Hence, this study also includes a common persons linking using the item-

[3] However, the number of enrollment to the vocational schools has grown fast since 2004 due to a special endorsement from the MoCEI, specifically in three major provinces in Indonesia (DKI Jakarta, central Java, and DI Yogyakarta) in which more students enrolled in the vocational schools than those did in general senior secondary education (MoECI, 2013b).

[4] In this case, the best school refers to a school at which students with high national examination results competed to get in.

level responses of the 9th grade students' mandatory NE and of their PISA mathematics test as approximately 40% of the participating students in PISA were still in 9th grade. The NE test was administered in May 2012, only few weeks after the PISA 2012 test was held (MoCEI, personal communication, January 21, 2015). Although PISA does not test school mathematics (OECD, 2013b), knowing the relationship between the student performance in the low-stakes PISA and in the high-stakes NE may provide powerful insights on how content knowledge assessed in PISA differs from the Indonesian standards and whether the low performance of the Indonesian students in PISA could be attributed to the test-taking motivation. This comparison would provide criterion-related validity evidence.

1.3. Data Sample

1.3.1. Description of PISA 2012 items

In PISA 2012, each participating student from 65 countries was randomly assigned to one paper-based test booklet and given 2 hours to complete it, with a short break in the middle. Afterward, the individual student filled in a 30-minute background questionnaire (using the same sampling technique as the cognitive test) which collects information about the students themselves, their parental background, and their home environments. The school principal was given a 30-minute school questionnaire covering the school system and its teaching and learning environment. There was also an option of taking a computer-based assessment in mathematics and reading, but Indonesia only participated in the paper-based test (OECD, 2014).

For the purpose of this study, item level scores of Indonesian students were used for analysis, giving a sample size of 5622 from 209 schools. The not-reached (NR) items were treated as missing response during the item parameter calibration process, but scored as incorrect when being used for estimating the student proficiency and further analysis, as PISA 2012 Technical Report suggests (OECD, 2014). The rationale for this different treatment may be due to some practical and fairness reasons. In many testing conditions, the NR items are all of the consecutive missing values clustered at the end of the test book as the test-taker has time and/or ability limitations to attempt them. An item becomes not-reached mostly because of its location or the order in which it appears on a particular test. Such an item can be reached easily if placed early in the sequence, or vice-versa. Hence, treating the NR item as missing during the calibration process will not weigh down the parameter estimate and thus can give a better estimate of item difficulty, irrespective of its location on a particular test. On the other hand, scoring it as an incorrect response when estimating the test-taker's ability is often a policy decision. As it can be difficult to speculate the real reason for not reaching the item, giving a NR item a zero score is then deemed fair to all test-takers. By doing so, all information available about the test-taker's latent ability is also utilized.

By design, each mathematics item is intended to tap into one single content area, which can be considered as a latent construct or dimension. The 84 mathematics items are distributed equally across each construct, and so each construct is represented by 21 items. Most items are dichotomously scored (1 is correct, 0 is incorrect), whilst eight items are polytomous (scored as 0 to 2 with 2 as the most correct answers). The mathematics item classification by content was provided in the PISA 2012 Technical Report (OECD, 2014) and summarized in Table A.2.1.

1.3.2. Description of selected background variables

The students' item level scores were obtained along with information about their gender, grade level, assigned booklet, socio-economic status (SES), school type (public vs private) and school location (village, small town, city, and large city) from publicly available datasets. Table 1.3 shows the random proportional distribution of booklet assignment to the sampled students, while the distributions of the sampled students across gender and grade levels were presented in Table 1.4. The majority of students were in 9^{th} grade (40%) and 10^{th} grade (46%). Meanwhile, about ten percent of the students were still in 7^{th} and 8^{th} grade, whereas a little more than four percent were above 10^{th} grade. Interestingly, nineteen students were already in 12^{th} grade, who might have either started school at an early age or skipped class due to their special talent. As the 10^{th} grade is the grade level at which the number of participating students is the largest, it is considered the modal grade. At the international level, 10^{th} grade is also the modal grade.

Table 1.3
Frequency distribution of Indonesian students across standard booklets in PISA 2012

Booklet	Frequency
1	424
2	425
3	417
4	431
5	422
6	431
7	434
8	441
9	432
10	449
11	445
12	438
13	433
Total	5622

Table 1.4
The distribution of Indonesian students across gender and grade levels in PISA 2012. Grade 10 is the modal grade (bolded) for both Indonesia and International data

Gender	Grade Level						Total
	7	**8**	**9**	**10**	**11**	**12**	
Female	39	169	1,104	1,411	128	9	2860 (50.9%)
Male	61	268	1,150	1,186	87	10	2762 (49.1%)
Total (count)	100	437	2254	**2597**	215	19	5622
(%)	1.8	7.8	40.1	**46.2**	3.8	.3	100

As presented in Table 1.5, the school location variable indicates the community context in which the school is located. It has five categories including village (serving less than 3000 inhabitants), small town (3000-15,000 inhabitants), town (15,000 – 100,000 inhabitants), city (100,000 – 1 million inhabitants), and large city (> 1 million inhabitants). The distribution of sampled students across the different school categories is also delineated in Table 1.5. In PISA 2012, almost two third of the sampled students (66%) attended public schools located in a rural community of less than 15,000 inhabitants. On average, these students came from low income families compared to those who attended schools in bigger communities (see Figure 1.1). There were also higher percentages of sampled students attending public schools than private ones with more students coming from the general junior secondary programs (about 40%). Almost half of the sampled private schools were Islamic religious schools and vocational schools located in non-urban areas. As previously discussed, these schools were more likely to be serving students from poor communities (OECD-ADB, 2015).

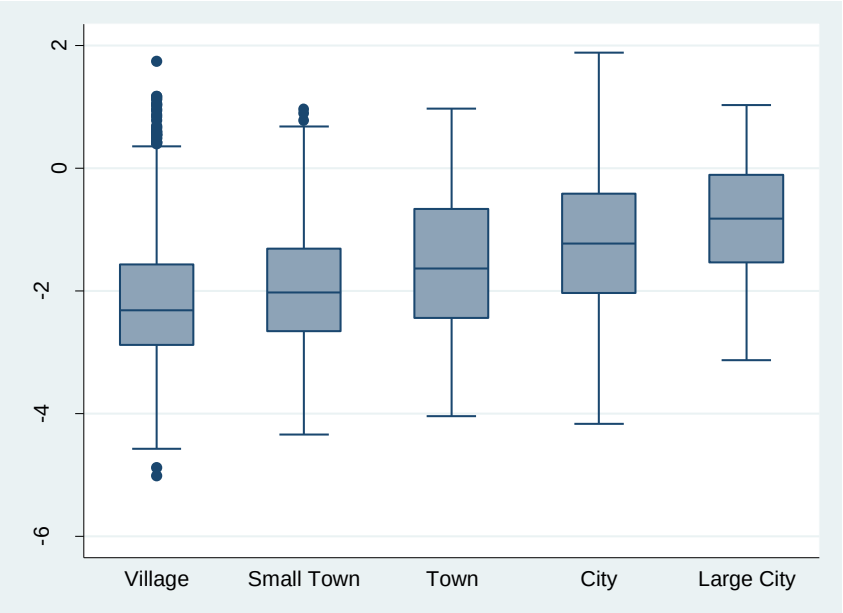
Table 1.5

The percentage of sampled student distribution across secondary school types, programs, and locations in PISA 2012, calculated with respect to the whole country sample

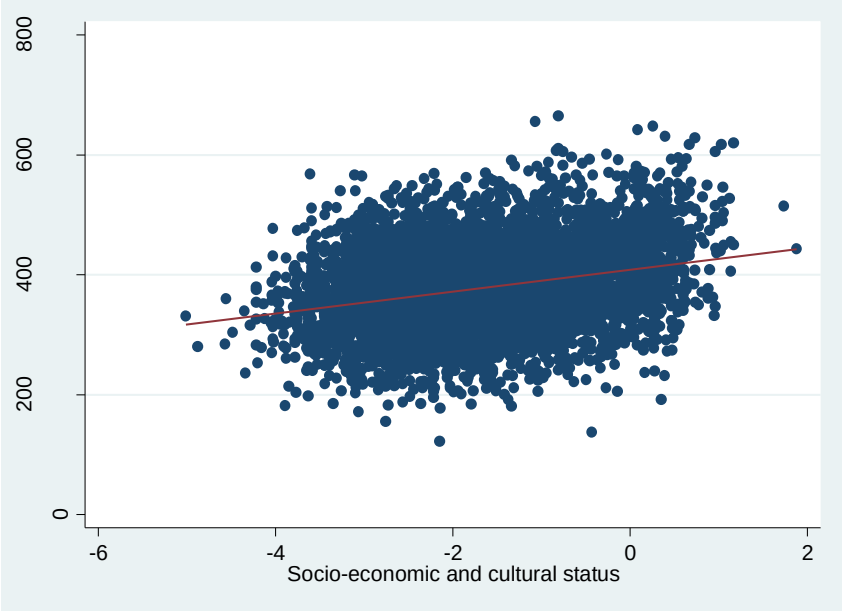
School location	General Junior (SMP)[1]		Islamic Junior (MTs)[2]		General Senior (SMA)[3]		Islamic Senior (MA)[4]		Vocational Secondary (SMK)[5]		Total Public	Total Private	Total
	Pu[6]	Pv[7]	Pu[6]	Pv[7]	Pu[6]	Pv[7]	Pu[6]	Pv[7]	Pu[6]	Pv[7]			
Village	8.7	3.6	-	6.6	2.2	0.6	.1	1.5	1.2	0.8	12.1	13.2	25.3
Small Town	15.0	3.5	.9	1.3	9.9	1.2	-	1.7	4.0	2.8	29.8	10.5	40.4
Town	3.3	0.1	-	1.2	3.6	0.9	1.1	.4	.5	3.3	8.5	6.0	14.5
City	2.7	1.8	-	-	2.9	2.3	.6	-	1.2	5.0	7.4	9.1	16.5
Large City	.5	.3	-	-	1.1	1.2	-	-	-	.2	1.6	1.7	3.3
Total	30.2	9.4	.9	9.2	19.7	6.2	1.8	3.7	6.9	12.0	59.5	40.5	100.0

Notes. [1] SMP (Sekolah Menengah Pertama) = general junior secondary school, [2] MTs (Madrasah Tsanawiyah) = Islamic junior secondary school, [3] SMA (Sekolah Menengah Atas) = general senior secondary school, [4] MA (Madrasah Aliyah) = Islamic senior secondary school, [5] SMK (Sekolah Menengah Ketrampilan) = vocational secondary school, [6] Pu stands for public schools, and [7] 'Pv' means private schools.

(a)



(b)



(c)

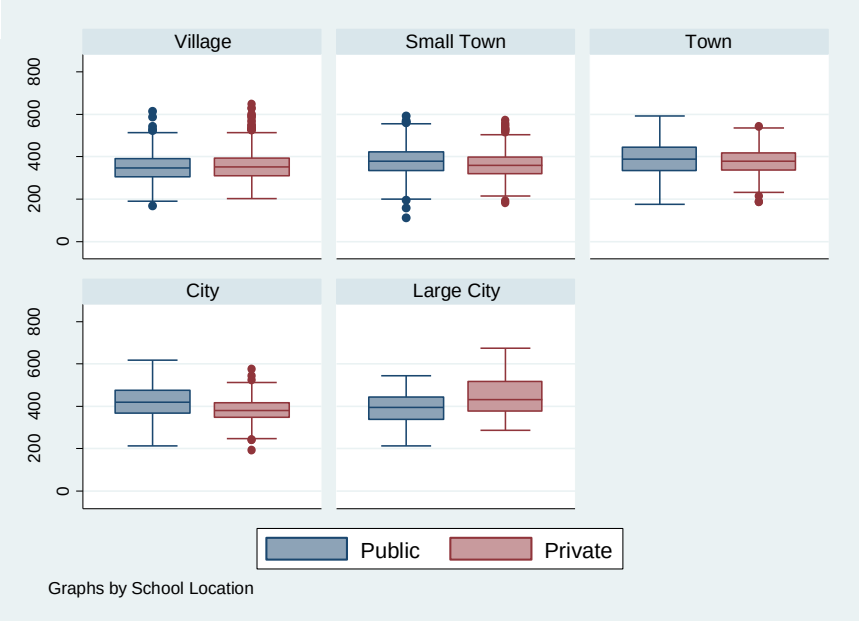


Figure 1.1. Descriptive figures of the Indonesian students participating in PISA 2012: (a) the distribution of the students' SES across different school locations[5], (b) the scatter plot of the students' math scores (from 1 plausible values) against their socio-economic and cultural status, and (c) the distribution of the students' math scores across different school type and locations.

[5] School location represents community context in which the school is located: village (serving less than 3000 inhabitants), small town (3000-15,000 inhabitants), town (15,000 – 100,000 inhabitants), city (100,000 – 1 million inhabitants), and large city (> 1 million inhabitants).

The SES indicator is a composite scale for the economic and socio-cultural status of an individual student. It is derived from the indices of highest occupational status of parents, highest educational level of parents in years of education, and home possessions using a principal component analysis (OECD, 2014). Higher values indicate relatively higher levels of SES and are standardized to a mean of zero for the population of students in OECD countries. A one-unit difference on the scale represents a difference of 1 standard deviation on the scale distribution. Most of the Indonesian sampled students have an SES index below the OECD average with a mean of -1.77 and a standard deviation of 1.09, and the minimum and maximum values of SES are -5.01 and 1.88, respectively. With large economic gaps, students with moderately high SES tend to attend schools located in the cities. In Figure 1.1(a), the mean student's SES attending schools in the less populated areas are lower than those who lived in more densely-populated cities. However, there are some anomalies as shown in Figure 1.1(a), i.e. rural students with high SES. This phenomenon might refer to students who live in small villages near natural resources or mining fields in remote areas throughout Indonesia and attend a special private school. As anticipated, there was an indication that the student performance in PISA 2012 math was positively associated with their SES levels (Figure 1.1(b)), which also varies according to their school type and locations (Figure 1.1(c)).

Hence, an investigation of the effects of gender, grade levels, socio-economic status, school types and school locations as well as the booklet assignment is necessary to understand how the students performed across the four intended content areas after controlling for these background variables. The terms "dimension" or "construct" will be used interchangeably throughout this paper to label the prescribed content area of mathematics.

1.3.3. Description of the National Examination items and data

For the purpose of linking students who took both PISA test and the NE in 2012, I obtained (1) a dataset of 9th grade students' item-level scores of NE math test (N = 497) and (2) a dataset of PISA math item-level scores for another set of 9th grade students (N = 246) from the MoECI staff (Ministry of Education and Culture personnel, personal communication, January 2015). The NE math test has 40 dichotomous items and there are no missing values. As both datasets had a unique identifier for each case (i.e. NE ID = student ID for the national examination), using NE ID as the unique key I merged both datasets to assign the corresponding NE math item-level scores to each of the 246 students whose PISA item-level scores were provided by MoECI. In order to reconfirm the accuracy of the dataset, I then compared the string of PISA math item scores obtained from MoCEI against the string of PISA math item scores obtained from the publicly available dataset (each string consists of 84 one-digit responses to all 84 PISA math items). If there were any duplicate strings, I went further on comparing the associated school type, school category, and booklet ID of the particular string (given in both datasets) and then chose the string with the same background information. The final dataset contains item-level scores of the 40 NE items and 84 PISA items from 223 ninth grade students, who took both PISA and NE math tests in 2012, nested in 16 schools.

The mathematics standards for the course of 3 academic years[6] at the junior secondary level include such topics as numbers, algebra, geometry and measurement, and statistics and probability, each of which was emphasized differently at each grade (BNSP, 2006b).For example, in the 9th grade, topics related to geometry and measurement (e.g. congruence and similarities of 2-D figures, measurement of spheres, cones, and tubes) and statistics and probabilities (e.g. measures of central tendency, data presentation, simple probability) were assigned for the first semester, whereas topics on number (e.g. integer exponents and roots, solving algebraic function with simple exponents and roots, simple calculation of arithmetic and geometric series) were dedicated to the second semester. Table A.2.2 lists the item specification of the forty 9th grade NE math items. An attempt was also made to map the NE items onto the four mathematics content areas defined in PISA 2012. This item mapping scheme can essentially show that NE has covered mostly, if not all, the same content areas of mathematics. From Table A.2.2, note that most of the NE items assessed geometry skills (43%), which were more likely to reflect the *Space and Shape* content area in PISA 2012, and only six items (15%) were related to the statistics and probability-related concept or linked to the *Uncertainty and Data* defined in PISA 2012. Meanwhile, almost the same number of items seemed to have reflected the *Change and Relationship* and *Quantity* content areas (i.e. 23% and 20%, respectively).

1.4. Methods

This study performs sub-dimension level analyses in order to examine the students' use of score categories within and across dimensions. A multidimensional random coefficient multinomial logit model (MRCMLM) is employed to calibrate the item parameters and ability estimates (Adams, Wilson, & Wang, 1997), the parameter estimation software ConQuest was used for estimating the parameters (Wu, Adams, Wilson, & Haldane, 2007; Adams, Wu, Haldane, Sun, 2012), the same as is used by the PISA itself (OECD, 2012). The MRCMLM is a generalized Rasch item response model that uses a scoring function and a design matrix to accommodate the applications of many existing IRT models used in this study such as the simple logistic model (Rasch, 1960), the partial credit model (Masters, 1982), the multi-facet model (Linacre, 1994), and the multidimensional versions of these models (Adams, Wilson, & Wang, 1997).

Master's partial credit model (PCM) is used to deal with the mixture of dichotomous and polytomous items. In multidimensional PCM, each student/person p's latent ability estimate in dimension d is estimated using a probability model where the probability (P_{ik}) of answering an item i in response category k is a function of the difference between the location of person p and the location of item i. Incorporating R number of person

[6] One academic year consists of 2 semesters and runs from mid July of a particular year to mid June of the next year.

background variables, the multidimensional latent multiple regression with PCM can be formulated as

$$\eta_{pi} = \theta_{pd} - (\delta_i + \tau_{ik} + B_l).$$

Here, $\eta_{pi} = \log\left(\frac{P_{ik}}{P_{ik-1}}\right)$ is the logit link to represent the probability model as a linear function of person latent ability θ on each dimension d, the relative item difficulty δ_i for a particular item i along with its *k*-th threshold parameter (τ_{ik}) when using PCM, and the booklet parameter B_l as facet ($l = 1, \dots, L$). The threshold parameter (τ_{ik}) is the deviation from the mean item difficulty δ_i for item i at step *k* (i.e. $k = 0, \dots, K$) and constrained such that $\sum_{k=0}^{K} \tau_{ik} = 0$. The booklet parameter is introduced to account for the order and position effect of an item in a test.

For a latent multiple regression model, $\theta_{pd} = (\sum_{r=1}^{R} \beta_r Z_{pr}) + \varepsilon_{nd}$ (Wright & Masters, 1982; De Boeck & Wilson, 2004). In this case, d indicates a specific latent dimension (i.e. $d = 1, \dots, D$); θ_{pd} represents person p's latent ability parameter on dimension/construct d; τ_{ik} is the item step difficulty parameter for item i at category k; β_r is the fixed latent regression coefficient of person covariate r (i.e. r = 1, …, R); Z_{pr} denotes the value of person p on covariate r; and ε_{pd} is the remaining person effect after the effect of the person covariates is accounted for ($\varepsilon_{pd} \sim N(0, \sigma_\varepsilon^2)$).

To address both research aims, unidimensional and multidimensional PCMs were fit to the PISA item responses of the Indonesian students. Before fitting both models on the Indonesian dataset, the mathematics item parameter estimates were obtained from using a unidimensional model on 500 randomly sampled cases from each of the 63 participating countries[7] to conform with the international calibration process defined in PISA 2012 (OECD, 2014). The selected random cases include students taking all booklet forms (Booklets 1 to 27) and having the overall 109 mathematics items. When estimating the item parameters, the booklet effect was treated as a facet[8]. However, to anticipate the effect of item locations within and between booklets, the booklet parameters for the standard booklet (Booklet 1 to 13) were calibrated separately by using the equally-weighted international pooled sample from 48 countries that took the standard booklets after excluding after excluding students with the UH booklet. More detailed discussion on the evaluation of the booklet effects is provided in the Technical Notes at the end of this paper. For the subsequent models used in this study, the item and booklet parameter estimates[9] calibrated from the international samples were used as the anchoring parameters for the model calibration with the Indonesian data. The item fit statistics of the internationally calibrated models were all within the fit tolerance bounds i.e. weighted MNSQ of 0.75 – 1.33, as recommended by Adams and Khoo (1996) in Wilson (2005).

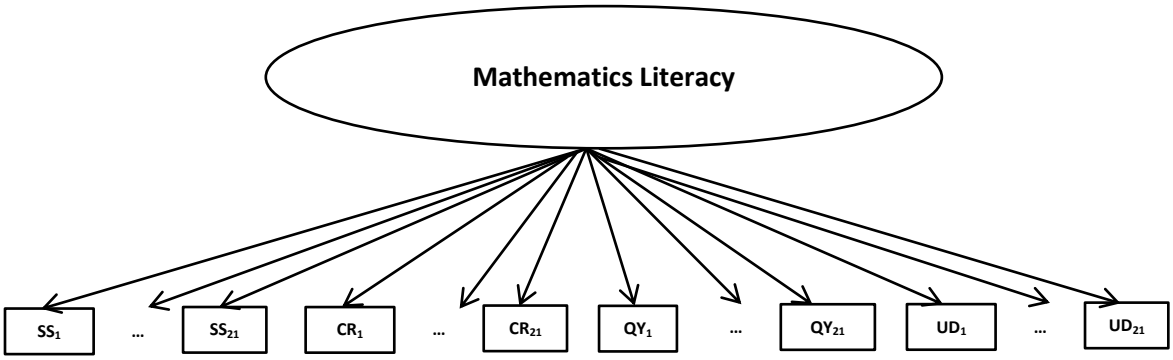
[7] Excluding Liechtenstein (country ample size is 293) and Cyprus (as there is no single authority representing both Turkey and Greek Cypriot people on the island (OECD, 2014, p. 233).

[8] Using a model statement of "item + item*step + booklet" in ConQuest.

[9] The modeled item and booklet parameters have been in a good agreement with the published figures as provided by the PISA 2012 Technical Report (OECD, 2014). Any discrepancy occurred was most likely due to the different random set of the international calibration sample used in this study.

Although students participating in the PISA test were selected randomly, the selection probabilities of these students vary and thus need to be accounted for in the parameter estimation process. This variability is mostly due to school and student sampling design such as school size and school/student non-response rates. Therefore, the given student survey weights were incorporated in all model development for analysis. The student survey weight is calculated from the school base weight, the within-school base weight, some adjustment factors to compensate for non-participation by school/students, the school base weight trimming factor, and the final student weight trimming factor (OECD, 2014).

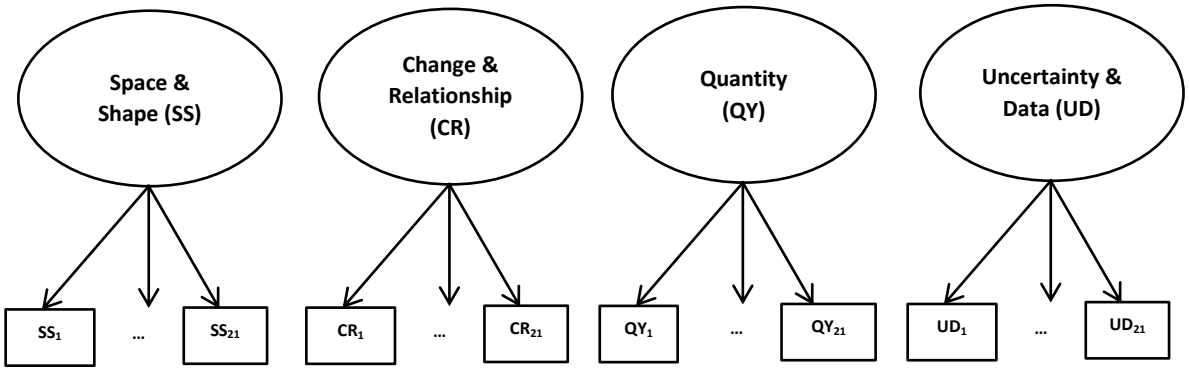
In order to examine the internal structure of the test, i.e. Research Aim 1, a unidimensional (1-D) PCM model was fit to test the assumption that the PISA mathematics domain was a single dimension (see Figure 1.2). Next, I applied the consecutive approach (Briggs and Wilson, 2003) in which each of the four mathematics content areas was modeled independently as a separate unidimensional construct (Figure 1.3). Then, I fit a between-item four-dimensional (4-D) PCM model that allows correlations between each construct as illustrated by a path diagram in Figure 1.4. The between-item 4-D model is used because, by design, each mathematics item tapped only into one particular construct. To maintain comparability, the Indonesian students' ability estimates for all of these models are estimated using the anchor item parameters. Finally, I calculated each model's Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as well as the EAP/PV reliability value to evaluate the dimensionality. The EAP/PV reliability is a measure of test reliability output by ConQuest, which is calculated by dividing the variance of the individual EAP ability estimates by the observed person variance (Adams, 2005), and is equivalent to traditional reliability measures such as Cronbach's alpha.



Legend:

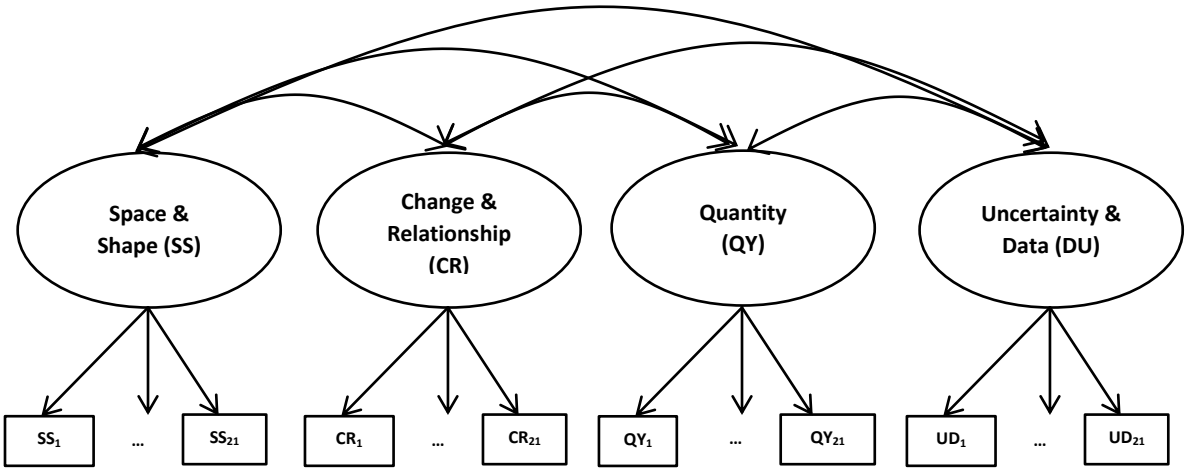
$SS_1 - SS_{21}$	: 21 items in "Space & Shape" construct
$CR_1 - CR_{21}$	: 21 items in "Change & Relationship" construct
$QY_1 - QY_{21}$	: 21 items in "Quantity" construct
$UD_1 - UD_{21}$	: 21 items in "Uncertainty and Data" construct

Figure 1.2. Illustration of unidimensional model defined for PISA 2012 mathematics domain.



Legend:
$SS_1 – SS_{21}$: 21 items in "Space & Shape" construct
$CR_1 – CR_{21}$: 21 items in "Change & Relationship" construct
$QY_1 – QY_{21}$: 21 items in "Quantity" construct
$UD_1 – UD_{21}$: 21 items in "Uncertainty and Data" construct

Figure 1.3. Illustration of the consecutive model approach defined for PISA 2012 mathematics subscales.



Legend:
$SS_1 – SS_{21}$: 21 items in "Space & Shape" construct
$CR_1 – CR_{21}$: 21 items in "Change & Relationship" construct
$QY_1 – QY_{21}$: 21 items in "Quantity" construct
$UD_1 – UD_{21}$: 21 items in "Uncertainty and Data" construct

Figure 1.4. Illustration of between-item four-dimensional model defined for PISA 2009 mathematics domain. In this model, each of the four mathematics subscales are assumed to be correlated.

To address Research Aim 2, a multidimensional Latent Multiple Regression (LMR) model was used with covariates such as gender, grade levels, school types and locations, and student's SES, to investigate effects of these background variables on the student ability estimates. Here, the booklet effect is incorporated by treating it as a facet in the model. In this case, the effect size of each covariate on a particular construct is computed by dividing the estimate of the covariate's regression coefficient by the standard deviation of the respective latent variable from the unconditional 4-D model with no covariates (Wu, Adams, Wilson, & Haldane, 2007, p. 112). I also examined the R^2 between the unconditional and conditional models to show changes in the variance explained by the model after including the covariates. All of the model parameter results are presented in Table 1.6.

When using the MRCMLM for developing a multidimensional model, the estimates of person and item location cannot be directly compared as they are estimated separately on each dimension. In other words, they are on a different scale since each dimension is separately centered at zero. Thus I used the Delta Dimensional Alignment (DDA) technique (Schwartz, 2012), which is one approach to transform the parameter estimates onto the same scale to allow direct interpretation and comparison across dimensions. Applying this alignment technique on the final multidimensional latent multiple regression model would then enable one to compare the means of the Indonesian students' latent ability estimates across the mathematics subscales and the effects of covariates on the students' ability estimates after controlling for other covariates.

For Research Aim 3, I argued that both the NE math test and the PISA 2012 math test would essentially assess the same latent construct called the "math ability", although the item emphasis of NE might not be equally distributed across the test content areas as for the PISA test (see Table A.2.2). For generating the latent estimates, I used anchoring item parameters for the NE items and the PISA items. The anchors for the NE items were generated using the bigger dataset of the NE item responses (N = 497) as previously discussed in Section 1.3.3, whereas the anchors for the PISA items were generated using the international sample as discussed in the preceding paragraph.

Table 1.6
Comparison of model runs on PISA 2012 mathematics performance of Indonesian students

No	Model	G^2	No of Par.	AIC	BIC	Mean Ability (SE)		Variance (SE)	MLE rel.	WLE rel.	EAP rel.
1	**Unidimensional model with anchored parameters (item & booklet)**	124,687	2	124,691	124,705	-2.653	(0.012)	0.856 (0.016)	0.570	0.570	0.679
2	**Unidimensional model with NO anchors, with booklet facet**	120,694	105	120,904	121,600	-1.912	(0.013)	0.854 (0.016)	0.571	0.572	0.674
3	**1-D Consecutive Approach with anchored item parameters, NO booklet facet**										
3a	Dimension 1: Space & Shape	27,191	2	27,195	27,208	-2.572	(0.014)	1.090 (0.021)			0.386
3b	Dimension 2: Change & Relationships	28,383	2	28,387	28,401	-2.750	(0.013)	0.927 (0.017)			0.368
3c	Dimension 3: Quantity	37,426	2	37,430	37,443	-2.816	(0.014)	1.108 (0.021)			0.478
3d	Dimension 4: Uncertainty & Data	35,278	2	35,282	35,295	-2.519	(0.011)	0.646 (0.012)			0.350
	Total	128,278	8	128,294	128,347						
4	**4-Dimensional unconditional model with NO anchored parameters, with booklet facet**	120,550	114	120,778	121,535						
4a	Dimension 1: Space & Shape					-2.446	(0.014)	1.044 (0.020)			0.621
4b	Dimension 2: Change & Relationships					-2.400	(0.014)	1.114 (0.021)			0.635
4c	Dimension 3: Quantity					-1.275	(0.015)	1.194 (0.023)			0.652
4d	Dimension 4: Uncertainty & Data					-1.416	(0.012)	0.756 (0.014)			0.600
5	**4-Dimensional unconditional model with anchored parameters (item & booklet)**	124,352	14	124,380	124,473						
5a	Dimension 1: Space & Shape					-2.565	(0.013)	1.001 (0.019)			0.627
5b	Dimension 2: Change & Relationships					-2.783	(0.014)	1.062 (0.020)			0.633
5c	Dimension 3: Quantity					-2.823	(0.014)	1.166 (0.022)			0.652

No	Model	G^2	No of Par.	AIC	BIC	Mean Ability (SE)		Variance (SE)	MLE rel.	WLE rel.	EAP rel.
5d	Dimension 4: Uncertainty & Data					-2.531	(0.011)	0.705 (0.013)			0.604
6	**4-Dimensional unconditional LMR model with anchored parameters (item & booklet) after DDA**	124,349	14	124,377	124,470						
6a	Dimension 1: Space & Shape					-1.865	(0.013)	1.012 (0.019)			0.627
6b	Dimension 2: Change & Relationships					-2.306	(0.014)	1.081 (0.020)			0.634
6c	Dimension 3: Quantity					-3.431	(0.014)	1.136 (0.021)			0.657
6d	Dimension 4: Uncertainty & Data					-2.815	(0.011)	0.688 (0.013)			0.614
7	**4-Dimensional conditional LMR model with anchored parameters (item & booklet)**	123,163	62	123,287	123,699						
7a	Dimension 1: Space & Shape					-2.973	(0.101)	0.937 (0.018)			0.662
7b	Dimension 2: Change & Relationships					-3.062	(0.095)	0.835 (0.016)			0.688
7c	Dimension 3: Quantity					-3.101	(0.100)	0.921 (0.017)			0.708
7d	Dimension 4: Uncertainty & Data					-2.831	(0.075)	0.521 (0.010)			0.673
8	**4-Dimensional conditional LMR model with anchored parameters (item & booklet) after DDA**	123,168	62	123,292	123,704						
8a	Dimension 1: Space & Shape					-2.271	(0.102)	0.948 (0.018)			0.657
8b	Dimension 2: Change & Relationships					-2.598	(0.097)	0.864 (0.016)			0.699
8c	Dimension 3: Quantity					-3.709	(0.100)	0.926 (0.017)			0.711
8d	Dimension 4: Uncertainty & Data					-3.116	(0.073)	0.492 (0.009)			0.686

1.5. Findings and Discussion

1.5.1. Research Aim 1 – Multidimensionality of PISA 2012 mathematics

The unidimensional (1-D) model. In this case, I assumed that all Math items together would constitute one single overarching domain called mathematics (see Figure 1.2). From Model 1 in Table 1.6, it is shown that the majority of the Indonesian students' ability estimates were located at the lower end, having a mean ability estimate of -2.65 logits (SE = 0.012) as indicated by the Wright Map in Figure 1.5(a). Many items seem to be very difficult even for the top sampled Indonesian students. The fit statistics of some items were outside the PISA tolerance bounds. Since calibrating the same 1-D model without anchoring the parameters as presented by Model 2 did not produce such misfit ($AIC_{Model\ 2}$ = 120,904 and $BIC_{Model\ 2}$ = 121,600), this suggests that the Indonesian students might demonstrate atypical performances compared to typical students throughout the participating countries. This finding can also be confirmed by Figure 1.5(b) that presents the Wright Map using the international pooled sample (from countries administering the standard booklets). In Figure 1.5(b), the items are distributed more evenly, although they also seemed to be less difficult for most students in other countries. The most difficult items perceived by the international sample, e.g. Item 79 (delta = 3.1) and 81 (delta = 4.6), were not drawn on the Wright Map in Figure 1.5(a) as none of the students tried to respond to these particular items. Items with partial credit scoring and in a constructed response format appeared to be the most difficult, i.e. Item 7, 8, 24, 39, 44, 63, 66, and 72. As shown in Table 1.6, the overall fit statistics of the 1-D model with no anchor parameters ($AIC_{Model\ 2}$ = 120,904 and BIC = 121,600), is smaller than that of the model with anchored parameters ($AIC_{Model\ 1}$ = 124,691 and $BIC_{Model\ 1}$ = 124,705), indicating that the latter model with no anchor parameters fits better. Further study is required to investigate why such phenomena occur as it is beyond the scope of our current study.

(a)

Logit Scale	Person Ability	Item scores	Booklet Difficulty
1		2 3 8 13 14 22	
		24 25 35 36 38	
		47 68 74	
		15 62 71	
		20	
		5 10 48 63	
		1 54 82 84	
		33 44	
0		16 17 32	
	X	18 30 31 42 56	
		59 64 65 76	
		9 11 27	
	X	7 37 55 80	
	X		
	X	12 43 46 58	
	XX	6 21 29	2 7 8 12
-1	X	23 28 61 75 83	1 3 4 5 6 10 11
	XX	67	9 13
	XX	49 70	
	XXX	26	
	XXXX	4	
	XXXX	52	
	XXXXXX	19 41 51 77	
	XXXXX		
-2	XXXXX		
	XXXXXXX	53	
	XXXXXXX		
	XXXXXXXX		
	XXXXXXXXXX	50 73	
	XXXXXXXX	45 ← Mean ability of -2.65 logits	
	XXXXXXXXX		
	XXXXXXXXX	34	
-3	XXXXXXXXX		
	XXXXXXXXXX		
	XXXXXXXXX		
	XXXXXXXX		
	XXXXXXX		
	XXXX		
	XXXXX		
	XXXX		
-4	XXX		
	XX		
	XX		
	XX		
	X		
	X		
	X		
	X		
-5			
-6			

(b)

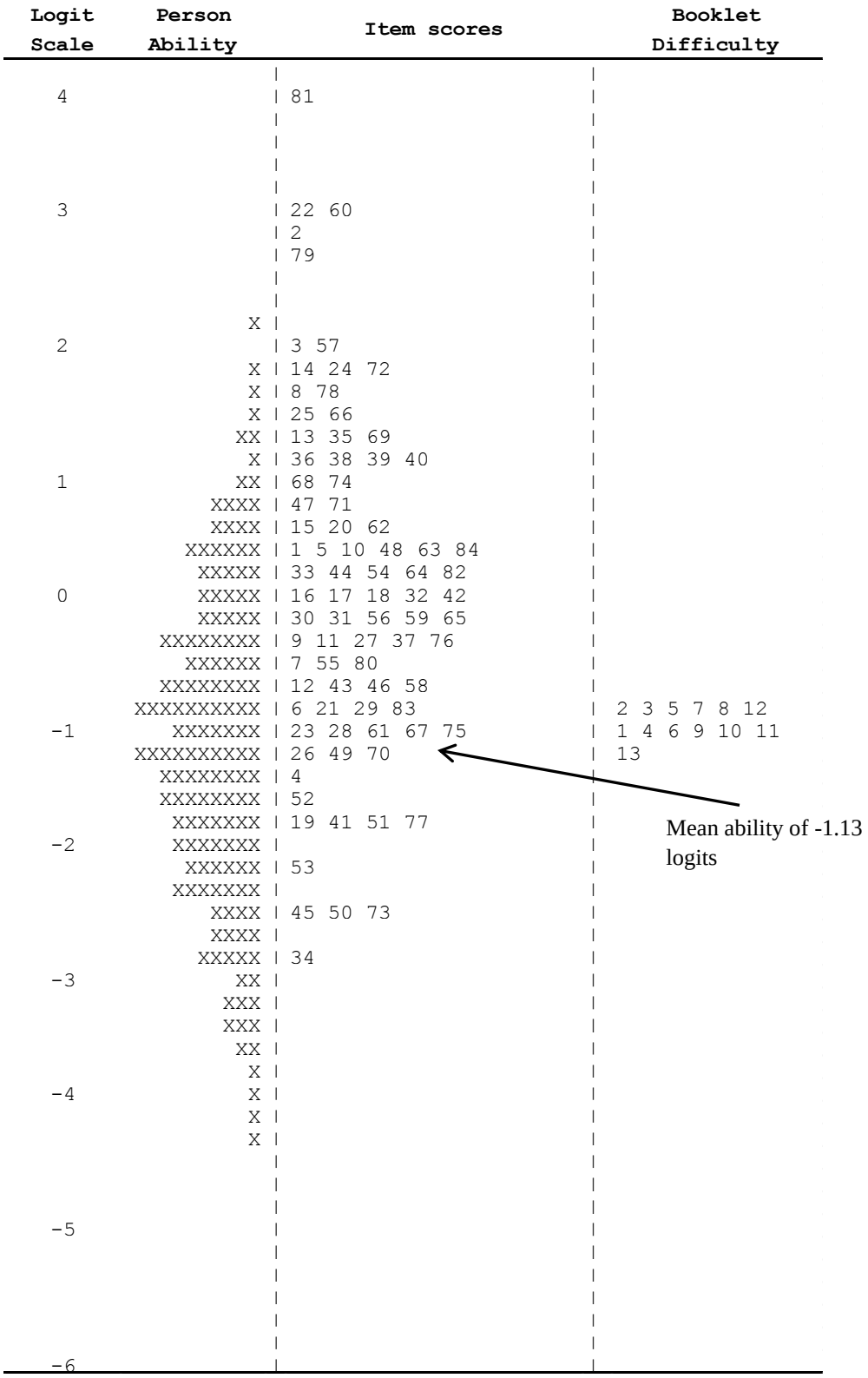

Logit Scale
Person Ability
Item scores
Booklet Difficulty
4
81
3
22 60
2
79
2
3 57
14 24 72
8 78
25 66
13 35 69
36 38 39 40
1
68 74
47 71
15 20 62
1 5 10 48 63 84
33 44 54 64 82
0
16 17 18 32 42
30 31 56 59 65
9 11 27 37 76
7 55 80
12 43 46 58
6 21 29 83
2 3 5 7 8 12
-1
23 28 61 67 75
1 4 6 9 10 11
26 49 70
13
4
52
19 41 51 77
-2
53
45 50 73
34
-3
-4
-5
-6
Mean ability of -1.13 logits

Figure 1.5. A Wright Map of the unidimensional model: (a) using the international pooled sample of countries taking the standard booklets (each 'X' represents approximately 2158 cases) and (b) using the Indonesian sample (each 'X' represents 34 cases). Numbers 1 to 84 on the "Item Score" column denotes the item number. The booklet difficulty (effect) parameters are shown in the right-most column.

The consecutive approach model. Then, I applied a consecutive-approach model to provide separate student latent ability estimates for each of the four constructs defined by PISA (Figure 1.3). In this model, each content area is considered as a single independent dimension and not correlated with other areas for estimation purposes. It can be noted that the reliability index (given by EAP person reliability) for each dimension is lower than those obtained by the 1-D model (see Model 3 in Table 1.6). This reliability value, however, was to be as anticipated because the number of items used in each dimension was also substantially reduced (since it covered only 25% of items for each construct). The worst reliability index in this case is one from the Uncertainty and Data (UD) construct by having an EAP reliability estimate of 0.35 as opposed to the 1-D reliability index of 0.678 by Model 1. In order to assess the global model fit, I summed up the AIC and BIC indices of each individual model (Model 3a – 3d) to give a value of 128,394 and 128,447, respectively, which are both greater than those of Model 1 (see Table 1.6). Since smaller AICs and BICs indicate better statistical fit, the consecutive models fit worse than the 1-D model.

The multidimensional model. Third, as a compromise between the unidimensional and consecutive approach models (Briggs & Wilson, 2003) and following the intended structure of the test, I applied a between-item four-dimensional (4-D) model, in which each item tapped into a single dimension/construct, and each dimension was allowed to correlate to each other (see Figure 1.4). This model considers the influence of other dimensions when estimating the item parameters and person ability on one dimension. Using the AIC and BIC values of the 4-D model (i.e. $AIC_{Model\ 5}$ = 124,380 and $BIC_{Model\ 5}$ = 124,473), as a basis for comparing the model fit, the multidimensional model fits the data better than both the 1-D model (Model 1) and the consecutive approach models (Model 3), as shown in Table 1.6. As the 1-D model is hierarchically related to the 4-D model, the change in the deviance (G^2) can also be used to assess the model fit. Here, the difference in G^2 between both models is 670 by using the likelihood-ratio test corrected for the boundary effects[1]. Comparing it with a χ^2(d.f. = 12), the 4-D model was indeed a better fit than the 1-D model, having a *p*-value < 0.0001. However, several item fit statistics in the 4-D model were also not within the acceptable bounds, which was potentially due to the anchor item and booklet parameters. A better model fit was also obtained in Model 4, which did not use anchor parameters (i.e. $AIC_{Model\ 4}$ = 120,778 vs $AIC_{Model\ 5}$ = 124,380).

The disattenuated (corrected) correlation is used throughout the data analysis because it has been corrected for measurement errors resulting from using imperfect assessments. The

[1] The likelihood-ratio test corrected for boundary effects: $G^2 = 2(L_1 - L_0)$, as described in Rabe-Hesketh and Skrondal (2012). Model L_1 is the full model, while model L_0 is the nested model (model with less number of parameters). Under the null hypothesis, G^2 has an asymptotic χ^2 (d.f.) null distribution. The 'd.f.' is the degree of freedom or the difference in the number of parameters between models L_1 and L_0.

disattenuated correlation among the dimensions of Model 5 (see Table 1.7) ranges from 0.86 to 0.92, which were slightly higher than those reported by Liu et al. (2008) who applied similar analysis on the American students' scores in PISA 2009 math. This suggests that the constructs are only moderately different from one another. In other words, reporting sub-scores on these different subscales can somewhat provide meaningful interpretation although students who score highly on one of the constructs would tend to score highly on the other and vice versa. The strongest pairwise correlations are between CR and QY constructs with a correlation of 0.92. Meanwhile, UD correlates about 0.86 with SS and about 0.89 with the other two constructs. This indicates that UD provides somewhat more distinct information about the specific construct.

The EAP/PV reliability estimates for each construct in this 4-D model (Model 5) are higher than those of the consecutive-approach models (Model 3a – 3d) because they also accounted for the interrelation among the constructs (see Table 1.7). However, in both the consecutive approach model (Model 3d) and the 4-D model (Model 5d), the reliability estimate for UD is the lowest.

Table 1.7

Corrected correlation matrix of the 4-dimensional (unconditional and conditional) non-DDA adjusted models

Corrected Correlation	Unconditional Model (Model 5)				Conditional LMR Model (Model 7)			
	Space & Shape	**Change & Relationships**	**Quantity**	**Uncertainty & Data**	**Space & Shape**	**Change & Relationships**	**Quantity**	**Uncertainty & Data**
Space & Shape								
Change & Relationships	0.905				0.907			
Quantity	0.906	0.924			0.896	0.916		
Uncertainty & Data	0.857	0.891	0.892		0.854	0.885	0.892	
Variance	1.001	1.062	1.166	0.705	0.937	0.835	0.921	0.521
R^2					6.39	21.37	21.01	26.10

Note. 'unconditional' refers to the basic multidimensional model before incorporating the covariates. LMR stands for latent multiple regression. DDA refers to the dimensional scale alignment approach.

Figure 1.6 presents the Wright maps of each construct modeled after the DDA adjustment; hence the students' latent ability estimates can be compared across dimensions. Consistent with the results for the 1-D model, this figure illustrates that overall items were overly difficult for most Indonesian students with the hardest items within the SS construct, i.e. Items 24, 63 and 81. Once the dimensional scales are aligned, the mean person ability estimates changes in their order. Before the DDA adjustment, Model 5 indicated that the distributions of ability estimates across dimensions were within a similar range, giving a similar mean estimate on each dimension. After applying the DDA approach, the distributions of the person ability estimates varied significantly, showing that the QY construct now seemed to be the hardest for the Indonesian students, followed by the UD construct. Furthermore, the mean latent ability estimates were low, between -3.43 and -1.87 logits with small standard errors (0.01). Almost half of the items related to the SS and CR constructs could not be readily solved by the students. Only top students could succeed in solving the least difficult item (item 49) in CR. Within the QY construct, more items appeared to be doable for the students with higher-than-average ability. Meanwhile, along the UD dimension, nearly half of the lower items spanned only the upper levels of the distribution of the student ability estimates. Item 50 and item 73 could be easily solved by the average students, but nearly 50% of the designated items were still overly difficult.

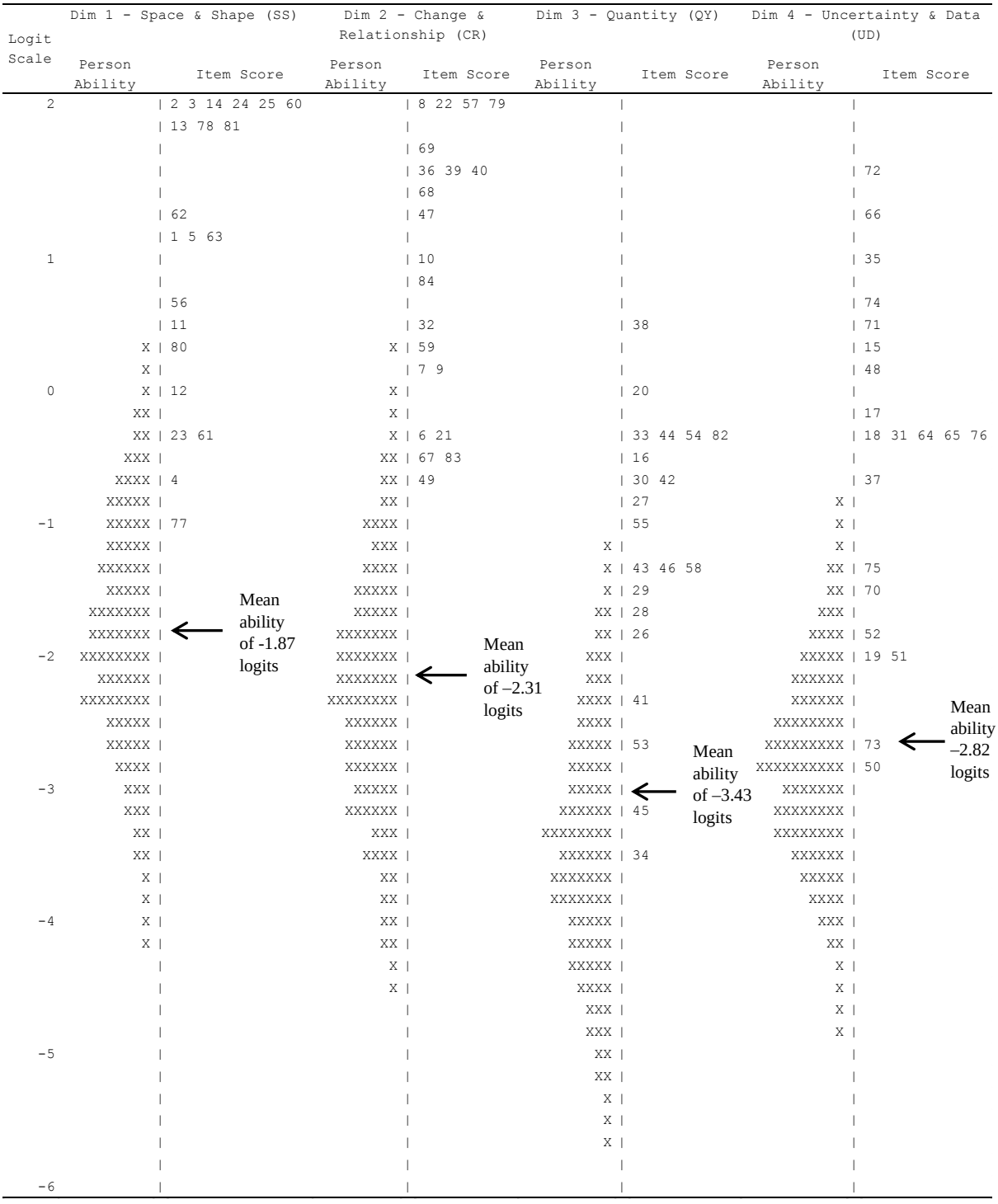


Figure 1.6. A Wright Map of the between-item 4-dimensional DDA-adjusted model using the Indonesian student scores in PISA 2012 mathematics. Each "X" in the person ability distribution represents approximately 53 cases, and numbers 1 to 84 on the item score column denote the item number. DDA refers to the delta dimensional scale alignment technique.

The Indonesian students did not seem to do well on PISA 2012 math items that invoked deeper thinking. Many of the items above the student locations are the ones with partial-credit scores and of extended constructed-response format. As in PISA 2009, the Indonesian students also had the most difficulties in solving non-multiple-choice items, i.e. short-response and open-constructed response types (Wihardini, 2012). However, there was also indication of the lack of test-taking motivation among the Indonesian participating students, which might explain the large number of not-reached items on their data scores (Indonesian Ministry of Education personnel, personal communication, January 2015). However, more detailed research would be needed to confirm this test impairment and explore how this phenomenon can influence the overall calibration and scoring process.

These findings may suggest potential problems in Indonesia's national curriculum standards and practices and perhaps also with the training of its teachers. The Wright maps of the SS and CR constructs may indicate the students' unfamiliarity with the items, potentially due to either the discrepancy of the particular parts in the school curriculum or the differences in the assessment practices of the respective constructs. The better span of items on the QY and UD constructs may indicate a better alignment of such constructs into the curriculum. The findings, however, warrant a further and more detailed investigation of the actual items to provide more justified insights on whether such problems are related to the contextual curricula, local assessment practices and/or language used in the item format.

1.5.2. Research Aim 2 – Effects of background variables using latent regression

Students' low performance in PISA can be due to all sort of things related to internal and external factors at either the personal level (e.g. self-motivation, home background), the classroom/school level (e.g. teaching and learning processes) or even the national level (e.g. curricula/standards, school finance). Hence, in addressing Aim 2, a multidimensional latent multiple regression (LMR) model (referred to as a "conditional" 4-D model) was used by incorporating the model parameters from the unconditional 4-D model as the anchor parameters. By using such approach, a more "correct" inference about the differences in the group means can be made, specifically since the group means are estimated directly from the item response data. In this study, gender, school type and location, and grade level for each student were used as covariates using dummy variables for each category of the corresponding covariates. The booklet effect has been incorporated by anchoring the booklet parameters and using them as facets in the model development. For the reference groups, I chose female for gender, public schools for school type, large city schools (serving a community of greater than 1 million inhabitants) for school location, and grade 7 students for the grade level covariates.

As shown in Table 1.6, the non-DDA adjusted conditional 4-D model (Model 7) gave a higher EAP reliability estimate for each dimension than the unconditional Model 5 did. The disattenuated correlations of the conditional Model 7 were mostly lower than the unconditional model (see Table 1.7). Fitting a model with covariates would generally produce parameter estimates with smaller standard errors as the unexplained residual can be

reduced (Wu, Adams, Wilson, & Haldan, 2007). The covariates added information to the prior distribution estimates necessary for calculating the EAP estimates. Thus, the smaller standard errors tend to lead to high reliability values. The R^2 values as given in Table 1.7 represent the additional variance that the covariates could explain for the 4-D conditional model. Thus, the conditional model variance is reduced by some 6 to 26 percent across dimensions. The largest R^2 value (about 26%) for UD suggests that there could be a contextual aspect related to this particular domain as different schools might emphasize this particular construct in their teaching and learning activities differently, e.g. at different grade levels with different quality of teaching.

Socio-economic Status (SES) Effect. Broadly from the literature, I can anticipate that SES plays an important role in students' academic success (Grisay, Gebhardt, Berezner, & Halleux-Monseor, 2007). In this study, SES gave a significantly positive effect on the student ability estimates across all constructs (see Table 1.8), after controlling for other covariates. A unit increase in the SES indicator predicts an increase of more than 0.15 logits in each of the dimensions. The inclusion of SES in the 4-D LMR as a control variable is deemed necessary as most of the Indonesian students participating in PISA 2012 attended schools in non-urban areas (see Table 1.5). For a densely populated country like Indonesia, schools located in communities inhabited by less than 15,000 people are most likely to be located in rural or remote areas, which are typically associated with low SES.

Table 1.8

Regression coefficients and effect sizes of the covariates being controlled for by the DDA-adjusted 4-D latent multiple regression model

Parameter	Dimension 1 - Space & Shape (SS)			Dimension 2 - Change & Relationships (CR)			Dimension 3 – Quantity (QY)			Dimension 4 - Uncertainty & Data (UD)		
	Coeff.	**S.E.**	**Effect Size**	**Coeff.**	**S.E.**	**Effect Size**	**Coeff.**	**S.E.**	**Effect Size**	**Coeff.**	**S.E.**	**Effect Size**
Intercept	-2.27	(0.10)		-2.60	(0.10)		-3.71	(0.10)		-3.12	(0.07)	
SES	**0.18**	**(0.01)**	**0.18**	**0.22**	**(0.01)**	**0.21**	**0.24**	**(0.01)**	**0.22**	**0.18**	**(0.01)**	**0.21**
Male	**0.30**	**(0.03)**	**0.30**	0.01	(0.03)	0.01	**0.08**	**(0.03)**	**0.07**	-0.03	(0.02)	-0.03
Private	**-0.09**	**(0.03)**	**-0.09**	-0.06	(0.03)	-0.06	**-0.11**	**(0.03)**	**-0.10**	**-0.11**	**(0.02)**	**-0.13**
Small town (3000 - 15,000 ppl.)	0.02	(0.03)	0.02	**0.11**	**(0.03)**	**0.10**	0.01	(0.03)	0.01	0.01	(0.02)	0.01
Town (15,000 - 100,000 ppl.)	**-0.11**	**(0.05)**	**-0.11**	**-0.11**	**(0.05)**	**-0.11**	-0.10	(0.05)	-0.09	-0.04	(0.03)	-0.04
City (100,000 - 1 million ppl.)	**0.07**	**(0.04)**	**0.07**	**0.24**	**(0.04)**	**0.23**	**0.18**	**(0.04)**	**0.17**	**0.17**	**(0.03)**	**0.20**
Large city (100,000 - 1 million ppl.)	**0.60**	**(0.08)**	**0.60**	**0.57**	**(0.07)**	**0.55**	**0.67**	**(0.08)**	**0.62**	**0.39**	**(0.06)**	**0.46**
Grade 8	0.14	(0.10)	0.14	0.12	(0.10)	0.11	0.08	(0.10)	0.08	**0.25**	**(0.08)**	**0.29**
Grade 9	**0.52**	**(0.10)**	**0.52**	**0.50**	**(0.09)**	**0.49**	**0.48**	**(0.10)**	**0.45**	**0.53**	**(0.07)**	**0.63**
Grade 10	**0.70**	**(0.10)**	**0.70**	**0.84**	**(0.09)**	**0.81**	**0.92**	**(0.10)**	**0.86**	**0.81**	**(0.07)**	**0.96**
Grade 11	**0.71**	**(0.12)**	**0.71**	**0.67**	**(0.11)**	**0.65**	**0.92**	**(0.12)**	**0.85**	**0.64**	**(0.08)**	**0.77**
Grade 12	**1.00**	**(0.19)**	**1.00**	**0.79**	**(0.19)**	**0.77**	**1.41**	**(0.19)**	**1.31**	**1.41**	**(0.14)**	**1.68**

Note. The significant parameter estimate at 5% level is written in a boldface font. "ppl." = people. "coeff." = regression coefficient.

The Gender Effect. After controlling for other covariates, significant gender effects appeared only for the SS and QY constructs as shown in Table 1.8, both favoring boys. These findings replicated the results of past studies that reported male students' higher ability in mathematics than girls on geometry-related construct (see Liu et al., 2008; Else-quest et al., 2010). In PISA 2009 the Indonesian male students also had a significantly better performance on the SS and QY constructs (see Wihardini, 2012). Better gender-friendly teaching and learning practices can be recommended to improve the girls' understanding and interest in learning these mathematics areas.

The Grade Effect. As anticipated, compared to 7^{th} grade students' performance, significant effect sizes across all constructs as grade level increases are apparent in Table 1.8. Fifteen year-olds who still sat in grade 7 or 8 would likely be students who repeated grade or who started school late due to various reasons – but mostly related to having a low SES background. Significantly large positive effects (ranging from 0.5 – 1.7) starting at 9^{th} grade for all constructs could potentially indicate some curriculum effect. The time period in which the PISA 2012 test was administered may support this rationale. PISA 2012 test was administered in March – April 2012 (The Indonesian Ministry of Education, personal communication, January 2015), right before the mandatory 9^{th} grade and 12^{th} grade NE were normally held (early-mid May). This may explain the large effects of being in 12^{th} grade across all dimensions since students participating in the PISA 2012 might have also been prepared well for the high-stakes exit exams that normally covered all materials from the previous three grade levels. Another evidence of a curriculum effect is the high significantly positive effect of the UD construct seen for 9^{th} grade level, at which such construct was first introduced as the Indonesian National Board of Educational Standards suggested (BNSP, 2006a). However, one should also note that the 15-year-olds who were already in 12^{th} grade at the time of the test can also be a special group of students. They could have either skipped class or sat in a special advanced class due to their academic talents.

The School Type Effect. Having a significantly negative effect for attending private schools across three constructs confirmed the superior performance of the public schooling system in Indonesia (MoECI, 2013b; OECD-ADB, 2015) , after controlling for other covariates. This finding agrees with a previous study by Newhouse and Beegle (2006) who found that public schools were the most preferable school choice in Indonesia. They argued that this superior benefit was mainly due to the high quality of inputs to public schools. The rationale of high quality of input can also explain the better performance of public school students in PISA because the 15-year olds, who in the majority were the 10^{th} and 11^{th} graders, had to pass the NE upon completion of 9^{th} grade in order to be accepted at public senior secondary schools (10^{th} – 12^{th} grade). As the NE score is a big ticket to get accepted at a majority of senior secondary schools, public schools would generally admit academically better students than private schools. This phenomenon suggests that Indonesia needs a better and more rigorous monitoring of national curriculum implementation at private schools, especially those located in non-urban areas.

The School Location Effect. School location dummies are included in the 4-D LMR model to provide some insights on how the location is related to the academic achievement

as it is also typically associated with the students' SES and the degree of the school's facility and capacity. In PISA 2012, more than half of the Indonesian sampled students attended schools located in a community that served less than 15,000 people, i.e. villages and small towns (see Table 1.5). In the context of Indonesia, these areas are normally located in low socio-economic environments with minimal educational resources (Suryadarma and Jones, 2013). Controlling for the other covariates, compared to students attending village schools, students at schools in more populated areas – often with better access to high quality teachers and educational resources – appeared to have better performances as Table 1.8 shows. The effect sizes of different school locations mostly increased at 5% statistical significance level across all constructs as the number of inhabitants the school is serving also increased, after controlling for other covariates. For instance, students attending large city schools outperformed similar students in village schools by about 0.6 logits in the SS and QY dimensions. In terms of its effect size, the difference in the performance between two school locations is about 60% of a standard deviation, after controlling for other covariates. However, there was no apparent significant influence of attending schools at small towns or towns on the student performance in the QY and UD dimensions when compared to the village schools, after controlling for other covariates. One might speculate that there was no substantial difference in the educational resources provided to nonurban schools. This finding may reinforce the need to close existing academic gaps between urban and nonurban students by ensuring an equitable distribution of financial and educational resources as well as improving poverty alleviation programs.

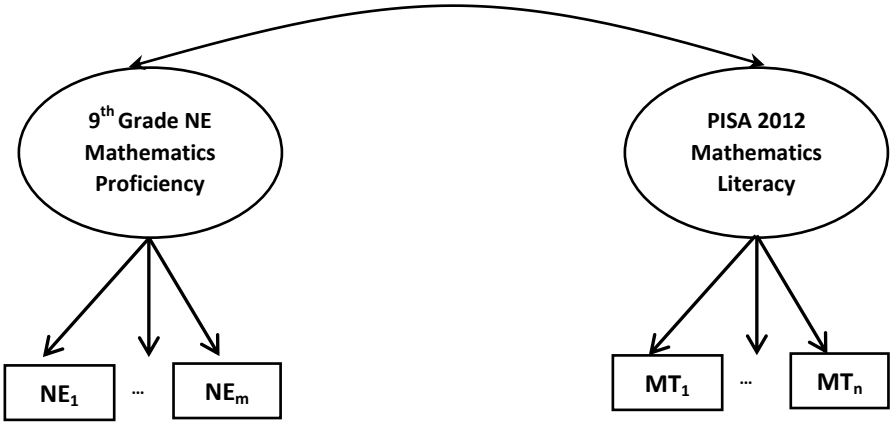
1.5.3. Research Aim 3 – Linking with the 2012 9th grade National Examination in mathematics

To estimate the latent correlation between the "math proficiency" construct of the NE math test and the "math literacy" construct of the PISA math test, I fit a between-item 2-dimensional model (see Figure 1.7), whose item parameters were anchored, using the subset of the 9th grade obtained data (N = 223). For the 2-D model, the NE's math construct tapped onto one dimension with 40 items, while the PISA's math construct tapped onto the other dimension with 122 items[1]. As a result, the disattenuated correlation of this 2-D model was about 0.08 indicating that there is almost no linear relationship between NE and PISA ($corr_{wle}$ = 0.06, $corr_{eap}$ = 0.10, and $corr_{raw}$= 0.07). This I interpreted as being due to the almost perfect passing rate of the national examination, i.e. there is almost no variability in the scores[2]. For the purpose of getting more informative insights on how students' math ability and item difficulties on both tests are comparable, I fit another 2-D model in which the items were freely calibrated using the sub-dataset, not anchored using the previously calibrated item parameters. By doing so, I aimed to apply the DDA technique in order to compare the parameter estimates of both tests just on this particular group of the 9th grade students. After

[1] Using this dataset, two PISA math items were excluded from the model estimation as none of the 9th grade students in the dataset were correct on those items.

[2] Correlation measures the extent to which the variability in one variable corresponds to the variability in another variable. As most of the students had almost 100% passing rate, the NE variable was almost equal to a constant and thus, there would be no other variable in can be correlated with.

applying the DDA technique, the Wright Map of the 2-D model as illustrated in Figure 1.8 shows that almost half of the NE math items seemed to be overly easy for the Indonesian 9th graders as students with the average math ability of -0.53 logits could endorse almost all NE items well. However, these students had less chance of getting most of the PISA math items correct. In other words, the PISA math items were seemingly overly difficult for them. Apart from a probable mismatch of curriculum being assessed or item format used in both test, another potential rationale to explain this finding is that the students did not take the low-stakes PISA test seriously as it was administered just few weeks before the scheduled high-stakes NE (Ministry of Education and Culture Staff, personal communication, Aug 20, 2015). A further investigation is warranted to justify the latter explanation.



Legend:
$NE_1 – NE_m$: *m* items in the 9th grade National Examination (NE) construct of mathematics proficiency
$MT_1 – MT_n$: *n* items in the PISA 2012 construct of mathematics literacy

Figure 1.7. Illustration of the between-item 2-dimensional model proposed to relate the 9th grade National Examination (NE) math results with the mathematics performance as defined by PISA 2012.

Although the test outcomes have such a low correlation, Figure 1.8 can be used to provide an insight and comparison on which content area that an item was perceived as difficult. For example, the sampled students with a latent ability estimate of -1.0 logits would have less than 50% chance of getting geometry-related items (Item 19, 27, 32 and 33) correct and thus, would find such items too hard, but these students got most of the PISA math items wrong. Examining Figure 1.8 also reveals that the easiest items seemed to represent the QY and UD content areas in either the NE test (i.e. Item 7 and 35 as indicated by Table A.2.2) or in the PISA math test (i.e. Item 73 and 89 or 111 as indicated by Table A.2.1) since these items were located at the bottom end of the Wright Map in Figure 1.8. Furthermore, it can be seen from Figure 1.8 that the complex multiple-choice (polytomous) items, i.e. represented by Item 54, 78, 96, 104, 117, and 119, of the PISA test seemed to be way too hard for the 9th graders as they were located at the very top of the Wright Map on the second dimension. Hence, should the dataset of the 9th graders who took both NE and PISA tests be extended (to all 9th graders participating in PISA 2012), more in-depth analysis of the item difficulties and

how these items as well as student ability estimates of the PISA test correlate with those of the NE test would be interesting.

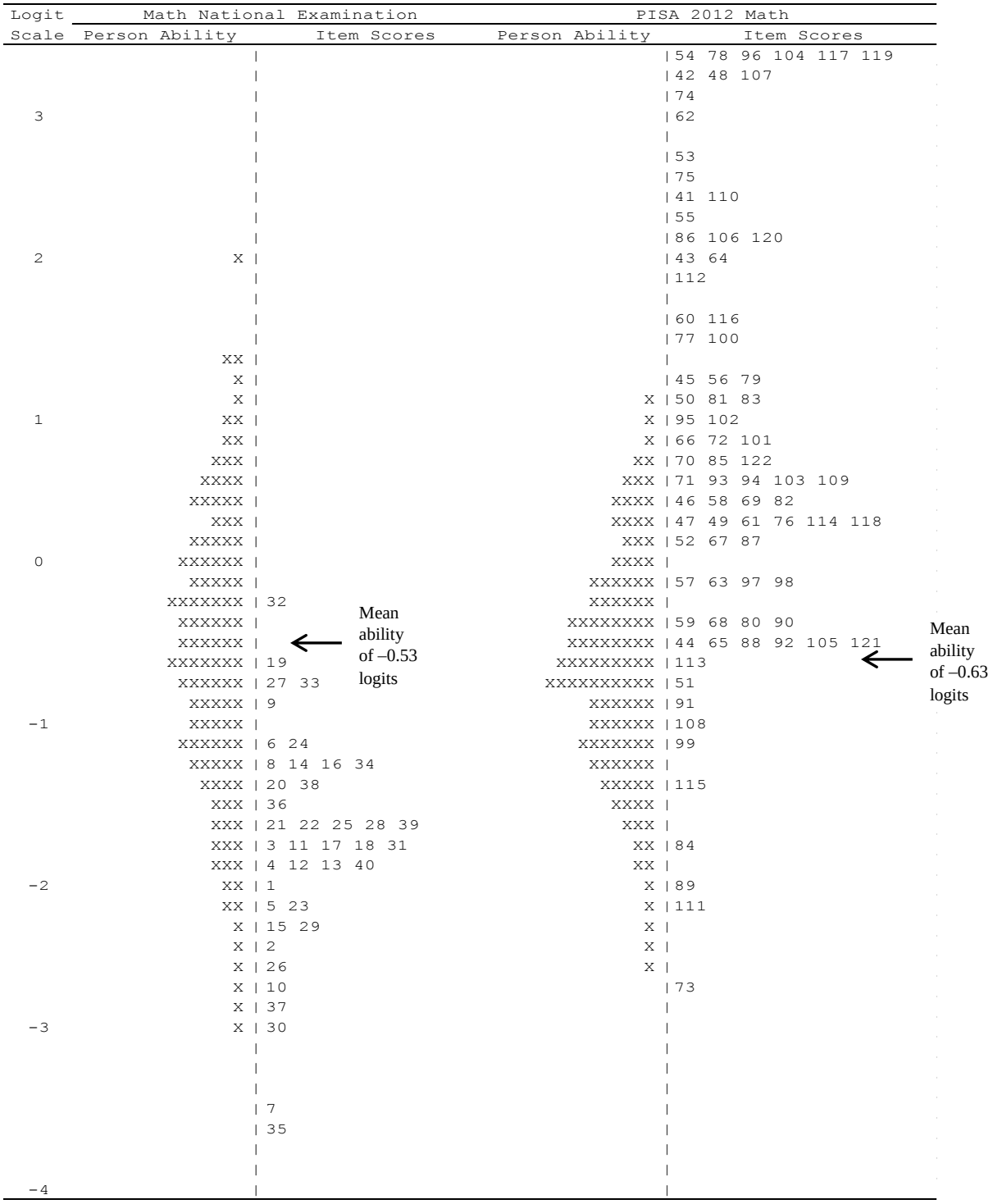


Figure 1.8. A Wright Map of the between-item 2-dimensional DDA-adjusted model using the Indonesian 9th grade student scores in the 2012 national math examination and the PISA

2012 mathematics. Each "X" in the person ability distribution represents approximately 2 cases, and numbers 1 to 122 on the item score columns denote the item number. DDA refers to the delta dimensional scale alignment technique.

1.6. Limitations and Future Directions

Findings from this study should be considered in light of several limitations. First, the person clustering as a result of the two-stage sampling has not been taken into account in the investigation of how the background variables correlated with the students' math literacy estimates across the four content areas. One would expand the multidimensional latent regression model to a three-level hierarchical linear model that takes into account the school clustering effect to better estimate these effects on the students' math literacy.

Second, anchoring the item parameters calibrated from the international sample to estimate the person ability produced a lot of model misfit for the Indonesian sample. This suggests a potential behavioral difference between the Indonesian students and students in other countries which were used as the basis for scaling. More investigations to understand how and why the models show misfit may give a better rationale behind the low scores for a particular set of test-takers.

Furthermore, as some PISA items are derived from common stimulus materials, incorporating the item-bundle method in scaling the scores could also be the next step, specifically for the overfit items. The sharing of a common stimulus can violate the strict local independence assumption of the item response model since the success of responding to one item may somewhat be influenced by the understanding of a common material or responding to another item at the same bundle.

In addition, the limited sample of students who took both PISA test and the national examination in the same year might contribute to ineffectiveness of the common person linking approach to provide acceptable criterion-related validity evidence of the PISA test on Indonesian students. Using a larger and more representative sample of these students can increase the utility of the PISA test along with the associated vast amount of background variables. Although PISA does not assess school curriculum, benchmarking any national examination results with the PISA outcomes can give essential information on how comparable the local education system is with the global system.

Finally, another future avenue of research can be proposed in investigating the common items that link several rounds of PISA tests to monitor the student performance over time. This longitudinal evaluation of the student performance would provide a better picture and evidence of the progress (or stagnancy) of the national education development. Relating this outcome with the student- or school-level background information would provide sound evidence to target improvements in the national curriculum development, teacher training, and school management.

1.7. Conclusion

In this study, I investigated three issues: (a) the sub-dimensional structure of PISA 2012 mathematics for Indonesian students, (b) the effects of the sampled students' background information such as socio-economic and cultural status (SES), gender, school types and locations, and grade levels on the latent ability estimates, and (c) the correlation between PISA and NE outcomes. First, I fitted a unidimensional model, then consecutive-approach models, and finally a four-dimensional (4-D) between-item model to assess whether multidimensionality of the mathematics content areas came about as expected. For estimating these models on the Indonesia data, I used item and booklet parameters calibrated from the international sample as anchors. As a result, I found the 4-D between-item model fitted the best, which imitated the intended PISA design. Hence, the multidimensional structure of PISA 2012 mathematics is supported. However, the high correlation between CR and QY content areas suggest that these domains may not be providing unique information about the corresponding aspects of the student's ability. The close relations between the CR domain and the other domains can be explained by the PISA 2012 assessment framework (OECD, 2013c). Solving problems related to many content domains in mathematics needs a type of operational and functional thinking that closely taps into the notion of change and relationship. At the same time, the success of solving those problems often depends upon the student's quantitative reasoning, which is the core aspect of the QY domain. It would, therefore, be interesting to see whether collapsing these two overarching domains into one dimension and pairing it with the other areas would constitute a better three-dimensional between-item model that may provide somewhat more distinct information. From the item level analysis across all constructs, the majority of items appeared to be overly difficult for the Indonesian students, and items with partial-credit scoring were the most difficult. This finding demands a further investigation on the actual item format and wordings to provide better insights for the development and improvement of the pedagogical content knowledge of such items and their corresponding constructs.

Secondly, a 4-D latent multiple regression with background variables as covariates was used to examine their effects on the students' latent ability estimates. This model was also estimated using the anchored item and booklet parameters. Here, significant effects of these covariates were found across most constructs. The following discussion presents a summary of the effects given by each covariate, after controlling for other covariates.

(1) As anticipated, SES had a strong association with the student's latent traits.
(2) The gender effect, favoring boys, appeared significant on the SS and QY domains, which repeated the finding from a past study (see Liu et al., 2008).
(3) This study also shows that public school students outperformed similar students in private schools across all constructs due to the advantage of getting high quality of inputs for public schools. Since many private schools cater to students with low SES, a better monitoring program on the implementation of the national curriculum standards and

distribution of educational funds and resources is urgently needed to improve their academic performance.

(4) School locations gave effects on the student's locations on all constructs. Students attending schools in more densely populated areas tend to perform better, suggesting the need to improve the teaching and learning processes in remote areas throughout Indonesia. Improving management and distribution of educational funds and resources, including providing good teachers, to non-urban areas may provide students with better access to high quality education.

(5) In this study, for lower grades, the student estimates on most constructs are lower. This finding could be due to late entry and/or promotion policies. These performance differences of students at different grade levels can also attest to the disadvantage of having a strict school progression system. Apart from having a late start at school or being considered incapable of progressing to the next grade level due to various reasons for the Indonesian case, these 7^{th} and 8^{th} graders could have missed the opportunity to learn basic skills required at their age and be at risk to achieve future academic success, unless accompanied by other interventions.

Finally, I found almost no correlation between the latent construct represented by PISA math and the latent construct represented by the NE math (disattenuated correlation = 0.08) from fitting a 2-D model of those constructs. However, showing outcomes in terms of the model's Wright maps can provide information on which mathematics content areas that the sampled students were lacking and hence needed to be emphasized and improved in the teaching and learning strategies.

Findings from this study have three major implications for the evaluation and implementation of the new curriculum, specifically in Indonesia. *First*, the results can be seen as contributing to development of the national mathematics curriculum at all grade levels, as successful outcomes on the PISA's four overarching mathematics constructs would depend upon the strength of their foundational knowledge taught in earlier years. The multidimensional analysis at item level can provide substantial insights on the difference in the student performance across constructs. In PISA 2012, the Indonesian students apparently performed the least well on the QY construct, followed by the UD construct. Although items in the SS and CR constructs appeared to be more difficult, the Indonesian students' ability estimates were higher than those in other constructs. These relative differences in the student performance across the constructs can then be utilized to improve the curriculum structure and/or the content pedagogical knowledge on the seemingly more difficult concepts. In-depth descriptions of student performance on each assessed construct may influence the assessment framework and inform changes in the instructional and pedagogical content of the teachers. *Second*, by analyzing the effect of background variables on student performance, the diverse socio-economic gaps and vast differences in local communities, they can be seen as important hurdles for the implementation of the new curriculum for students of all backgrounds. *Third*, utilizing PISA outcomes and relating them with the national examination outcomes at 9^{th} grade, or even 12^{th} grade level, can become a sound policy developmental tactic as it provides criterion-related validity evidence for justifying the cost and benefit of participating in the international large-scale assessment and validating the needs for rigorous education reforms. Furthermore, many critics of the new curriculum

change might welcome more information on studies about the benchmarks upon which the new curriculum is based. This study provides sound information on how the Indonesian students actually performed in the assessment framework laid out by PISA mathematics and how such information can be used to leverage and target evidence-based policy decisions.

Appendix A.1. Technical Notes on Booklet Effect

Corresponding to the PISA 2012 Technical Report (OECD, 2014), the booklet effect was modeled using a facet model to prevent confounding item difficulties and booklet design, and defined in a similar way as item parameters. Thus, a booklet parameter reflects the booklet difficulty, which can be added to the ability estimate of students who responded to the respective booklet. As PISA used the internationally estimated booklet effects in correcting all students' score, the steps described in the PISA 2012 Technical Report to investigate the booklet effects were implemented (OECD, 2014; ACER, personal communication, February 2015). As previously discussed in the earlier section, the booklet effects were calculated from using the equally-weighted international pooled sample of students in 48 countries that took the standard booklets after excluding students with the UH booklet (i.e. Une Heure booklet, one-hour booklet dedicated to students with special needs). Incorporating the booklet effect as facet in ConQuest gave booklet difficulty parameters, each of which was then adjusted by subtracting it from the booklet grand mean to give the same value as booklet effects published in the Technical Report. It can be seen from Table 1.9 that the booklet effects were all significant. A Wright map showing the booklet difficulty locations for the 1-D model from the international sample has been provided in Figure 1.5, which also shows the expected grouping of booklets around the mean ability of -1.13 logits.

Table A.1.1
The standard booklet parameter estimates in comparison with the published values in the PISA 2012 Technical Report (OECD, 2014)

Booklet	Estimate (SE)		Estimates using deviation contrast Coding	Published booklet effects
1	-1.009	(0.003)	-0.050	-0.054
2	-0.85	(0.004)	0.109	0.095
3	-0.911	(0.003)	0.048	0.059
4	-1.021	(0.002)	-0.062	-0.064
5	-0.952	(0.002)	0.007	0.001
6	-1.005	(0.002)	-0.046	-0.055
7	-0.883	(0.003)	0.076	0.076
8	-0.846	(0.004)	0.113	0.127
9	-1.051	(0.003)	-0.092	-0.076
10	-0.977	(0.003)	-0.018	-0.005
11	-0.996	(0.003)	-0.037	-0.038
12	-0.882	(0.005)	0.077	0.058
13	-1.079	(0.004)	-0.120	-0.123

Note. The published booklet effects were calculated as conditioning variables using deviation contrast coding (OECD, 2014, pp. 157 & 242).

Appendix A.2. Item Classifications

Table A.2.1

Item classification for PISA 2012 math items included in the standard booklets, as modified from OECD (2014). The bolded items are polytomous items, while the rest is dichotomous items

Item Index[1] (original)	Item Index[2] (standard)	Item Index[3] (NE-PISA)	Item Code	Item Name	Dim1 - Space and Shape (SS)	Dim2 - Change and Relation-ship (CR)	Dim3 - Quantity (QY)	Dim4 - Uncertain ty and Data (UD)
1	1	41	PM00FQ01	P2012 Apartment Purchase Q1	x			
2	2	42	PM00GQ01	P2012 An Advertising Column Q1	x			
3	3	43	PM00KQ02	P2012 Wheelchair Basketball Q2	x			
4	4	44	PM033Q01	P2000 A View with a Room Q1	x			
5	5	45	PM034Q01T	P2000 Bricks Q1	x			
6	6	46	PM155Q01	P2000 Pop Pyramids Q1		x		
7	**7**	**47**	**PM155Q02D**	**P2000 Pop Pyramids Q2**		**x**		
8	**8**	**48**	**PM155Q03D**	**P2000 Pop Pyramids Q3**		**x**		
9	9	49	PM155Q04T	P2000 Pop Pyramids Q4		x		
10	10	50	PM192Q01T	P2000 Containers Q1		x		
11	11	51	PM273Q01T	P2000 Pipelines Q1	x			
12	12	52	PM305Q01	P2000 Map Q1	x			
13	13	53	PM406Q01	P2003 Running Tracks Q1	x			
14	14	54	PM406Q02	P2003 Running Tracks Q2	x			
15	15	55	PM408Q01T	P2003 Lotteries Q1				x
16	16	56	PM411Q01	P2003 Diving Q1			x	
17	17	57	PM411Q02	P2003 Diving Q2				x
18	18	58	PM420Q01T	P2003 Transport Q1				x
19	19	59	PM423Q01	P2003 Tossing Coins Q1				x
20	20	60	PM442Q02	P2003 Braille Q2			x	
21	21	61	PM446Q01	P2003 The Thermometer Cricket Q1		x		
22	22	62	PM446Q02	P2003 The Thermometer Cricket Q2		x		

Item Index[1] (original)	Item Index[2] (standard)	Item Index[3] (NE-PISA)	Item Code	Item Name	Dim1 - Space and Shape (SS)	Dim2 - Change and Relation-ship (CR)	Dim3 - Quantity (QY)	Dim4 - Uncertain ty and Data (UD)
23	23	63	PM447Q01	P2003 Tile Arrangement Q1	x			
24	**24**		**PM462Q01D**	**P2003 The Third Side Q1**	**x**			
25	25	64	PM464Q01T	P2003 The Fence Q1	x			
26	26	65	PM474Q01	P2003 Running Time Q1			x	
27	27	66	PM496Q01T	P2003 Cash Withdrawal Q1			x	
28	28	67	PM496Q02	P2003 Cash Withdrawal Q2			x	
29	29	68	PM559Q01	P2003 Telephone Rates Q1			x	
30	30	69	PM564Q01	P2003 Chair Lift Q1			x	
31	31	70	PM564Q02	P2003 Chair Lift Q2				x
32	32	71	PM571Q01	P2003 Stop the Car Q1		x		
33	33	72	PM603Q01T	P2003 Number Check Q1			x	
34	34	73	PM800Q01	P2003 Computer Game Q1			x	
35	35	74	PM803Q01T	P2003 Labels Q1				x
36	36	75	PM828Q01	P2003 Carbon Dioxide Q1		x		
37	37	76	PM828Q02	P2003 Carbon Dioxide Q2				x
38	38	77	PM828Q03	P2003 Carbon Dioxide Q3			x	
39	**39**	**78**	**PM903Q01**	**P2012 Drip Rate Q1**		**x**		
40	40	79	PM903Q03	P2012 Drip Rate Q3		x		
41	41	80	PM905Q01T	P2012 Tennis Balls Q1			x	
42	42	81	PM905Q02	P2012 Tennis Balls Q2			x	
43	43	82	PM906Q01	P2012 Crazy Ants Q1			x	
44	**44**	**83**	**PM906Q02**	**P2012 Crazy Ants Q2**			**x**	
45	45	84	PM909Q01	P2012 Speeding Fines Q1			x	
46	46	85	PM909Q02	P2012 Speeding Fines Q2			x	
47	47	86	PM909Q03	P2012 Speeding Fines Q3		x		
48	48	87	PM915Q01	P2012 Carbon Tax Q1				x
49	49	88	PM915Q02	P2012 Carbon Tax Q2		x		
50	50	89	PM918Q01	P2012 Charts Q1				x
51	51	90	PM918Q02	P2012 Charts Q2				x
52	52	91	PM918Q05	P2012 Charts Q5				x

Item Index[1] (original)	Item Index[2] (standard)	Item Index[3] (NE-PISA)	Item Code	Item Name	Dim1 - Space and Shape (SS)	Dim2 - Change and Relation-ship (CR)	Dim3 - Quantity (QY)	Dim4 - Uncertain ty and Data (UD)
53	53	92	PM919Q01	P2012 Zs Fan Merchandise Q1			x	
54	54	93	PM919Q02	P2012 Zs Fan Merchandise Q2			x	
55	55	94	PM923Q01	P2012 Sailing Ships Q1			x	
56	56	95	PM923Q03	P2012 Sailing Ships Q3	x			
57	57	96	PM923Q04	P2012 Sailing Ships Q4		x		
58	58	97	PM924Q02	P2012 Sauce Q2			x	
68	59	98	PM943Q01	MATH - P2012 Arches Q1		x		
69	60		PM943Q02	MATH - P2012 Arches Q2	x			
73	61	99	PM949Q01T	MATH - P2012 Roof Truss Design Q1	x			
74	62	100	PM949Q02T	MATH - P2012 Roof Truss Design Q2	x			
75	**63**	**101**	**PM949Q03**	**MATH - P2012 Roof Truss Design Q3**	**x**			
76	64	102	PM953Q02	MATH - P2012 Flu Test Q2				x
77	65	103	PM953Q03	MATH - P2012 Flu Test Q3				x
78	**66**	**104**	**PM953Q04D**	**MATH - P2012 Flu Test Q4**				**x**
79	67	105	PM954Q01	MATH - P2012 Medicine Doses Q1		x		
80	68	106	PM954Q02	MATH - P2012 Medicine Doses Q2		x		
81	69	107	PM954Q04	MATH - P2012 Medicine Doses Q4		x		
82	70	108	PM955Q01	MATH - P2012 Migration Q1				x
83	71	109	PM955Q02	MATH - P2012 Migration Q2				x
84	**72**	**110**	**PM955Q03**	**MATH - P2012 Migration Q3**				**x**
93	73	111	PM982Q01	MATH - P2012 Employment Data Q1				x
94	74	112	PM982Q02	MATH - P2012 Employment Data Q2				x
95	75	113	PM982Q03T	MATH - P2012 Employment Data Q3				x
96	76	114	PM982Q04	MATH - P2012 Employment Data Q4				x
102	77	115	PM992Q01	MATH - P2012 Spacers Q1	x			
103	78	116	PM992Q02	MATH - P2012 Spacers Q2	x			
104	79	117	PM992Q03	MATH - P2012 Spacers Q3		x		
105	80	118	PM995Q01	MATH - P2012 Revolving Door Q1	x			
106	81	119	PM995Q02	MATH - P2012 Revolving Door Q2	x			

Item Index[1] (original)	Item Index[2] (standard)	Item Index[3] (NE-PISA)	Item Code	Item Name	Dim1 - Space and Shape (SS)	Dim2 - Change and Relation-ship (CR)	Dim3 - Quantity (QY)	Dim4 - Uncertainty and Data (UD)
107	82	120	PM995Q03	MATH - P2012 Revolving Door Q3			x	
108	83	121	PM998Q02	MATH - P2012 Bike Rental Q2		x		
109	84	122	PM998Q04T	MATH - P2012 Bike Rental Q4		x		
Number of items per content area/dimension for the uni-/ multi-dimensional math models					21	21	21	21
Number of items per content area/dimension for the NE - PISA model					19	21	21	21

Notes. [1] Item number as reflected in the original item classification table in the PISA 2012 Technical Report (OECD, 2014, p. 408-409), [2] Index of item as presented on the Wright maps provided for the unidimensional and multidimensional math models illustrated in Figure 1.5 and 1.6, [3] Index of item as presented on the Wright map provided for the 2-dimensional model of the national examination and PISA 2012 math illustrated in Figure 1.8.

Table A.2.2
Potential item mapping of the 9th grade national examination in mathematics across the PISA 2012 math content areas

Item No	Item description (Indonesian)	Item Description (English)	General concept	PISA Content Area			
				Space & Shape (SS)	Change & Relationships (CR)	Quantity (QY)	Uncertainty & Data (UD)
1	Menyelesaikan masalah yang berkaitan dengan operasi tambah, kurang, kali, atau bagi pada bilangan.	Solve a problem related to addition, subtraction, multiplication, or division of numbers	Number, arithmetics			x	
2	Menyelesaikan masalah yang berkaitan dengan operasi tambah, kurang, kali, atau bagi pada bilangan.	Solve a problem related to addition, subtraction, multiplication, or division of numbers	Number, arithmetics			x	
3	Menyelesaikan masalah yang berkaitan dengan perbandingan.	Solve a problem related to comparison	Number, arithmetics		x		
4	Menyelesaikan masalah yang berkaitan dengan bilangan berpangkat atau bentuk akar.	Solve a problem related to integer exponents or roots	Number			x	
5	Menyelesaikan masalah yang berkaitan dengan bilangan berpangkat atau bentuk akar.	Solve a problem related to integer exponents or roots	Number			x	
6	Menyelesaikan masalah yang berkaitan dengan perbankan atau koperasi dalam aritmetika sosial sederhana.	Solve a problem related to financial problems in a simple social arithmetics	Number, arithmetics			x	
7	Menyelesaikan masalah yang berkaitan dengan barisan bilangan dan deret.	Solve a problem related to arithmetic and geometric series	Number			x	
8	Menyelesaikan masalah yang berkaitan dengan barisan bilangan dan deret.	Solve a problem related to arithmetic and geometric series	Number			x	
9	Menyelesaikan masalah yang berkaitan dengan barisan bilangan dan deret.	Solve a problem related to arithmetic and geometric series	Number			x	
10	Menentukan pemfaktoran bentuk aljabar.	Determine factoring in an algebraic function	Algebra, functions		x		

Item No	Item description (Indonesian)	Item Description (English)	General concept	PISA Content Area			
				Space & Shape (SS)	Change & Relationships (CR)	Quantity (QY)	Uncertainty & Data (UD)
11	Menyelesaikan masalah yang berkaitan dengan persamaan linier atau pertidaksamaan linier satu variabel.	Solve a problem related to one-variable linear equation or inequality	Algebra		x		
12	Menyelesaikan masalah yang berkaitan dengan persamaan linier atau pertidaksamaan linier satu variabel.	Solve a problem related to one-variable linear equation or inequality	Algebra		x		
13	Menyelesaikan masalah yang berkaitan dengan himpunan.	Solve a problem related to sets	Algebra, sets		x		
14	Menyelesaikan masalah yang berkaitan dengan fungsi.	Solve a problem related to functions	Algebra, functions		x		
15	Menyelesaikan masalah yang berkaitan dengan fungsi.	Solve a problem related to functions	Algebra, functions		x		
16	Menentukan gradien, persamaan garis, atau grafiknya.	Determine a linear function, its gradient and graph	Algebra, graphs		x		
17	Menyelesaikan masalah yang berkaitan dengan sistem persamaan linier dua variabel.	Solve a problem related to two-variable linear equations or inequalities	Algebra, functions		x		
18	Menyelesaikan masalah menggunakan teorema Pythagoras.	Solve a problem using the Pythagoras theorem	geometry	x			
19	Menyelesaikan masalah yang berkaitan dengan luas bangun datar.	Solve a problem related to the area of two-dimensional figures	geometry	x			
20	Menyelesaikan masalah yang berkaitan dengan keliling bangun datar.	Solve a problem related to the perimeter of two-dimensional figures	geometry	x			
21	Menyelesaikan masalah yg berkaitan dgn hubungan dua garis, besar & jenis sudut, serta sifat sudut yg terbtk dari dua garis yg di potong garis lain	Solve a problem related to the relationship between two lines, and the magnitude, type, and characteristics of angles formed by the two lines intersected by another line	geometry	x			

Item No	Item description (Indonesian)	Item Description (English)	General concept	PISA Content Area			
				Space & Shape (SS)	Change & Relationships (CR)	Quantity (QY)	Uncertainty & Data (UD)
22	Menyelesaikan masalah yang berkaitan dengan garis-garis istimewa pada segitiga.	Solve a problem related to special sides of triangles	geometry	x			
23	Menyelesaikan masalah yang berkaitan dengan unsur-unsur/bagian-bagian lingkaran atau hubungan dua lingkaran.	Solve a problem related to the elements/parts of circles, or the relationship between two circles	geometry	x			
24	Menyelesaikan masalah yang berkaitan dengan unsur-unsur/bagian-bagian lingkaran atau hubungan dua lingkaran.	Solve a problem related to the elements/parts of circles, or the relationship between two circles	geometry	x			
25	Menyelesaikan masalah yang berkaitan dengan kesebangunan atau kongruensi.	Solve a problem related to congruence or similarities (of shapes)	geometry	x			
26	Menyelesaikan masalah yang berkaitan dengan kesebangunan atau kongruensi.	Solve a problem related to congruence or similarities (of shapes)	geometry	x			
27	Menyelesaikan masalah yang berkaitan dengan kesebangunan atau kongruensi.	Solve a problem related to congruence or similarities (of shapes)	geometry	x			
28	Menentukan unsur-unsur pada bangun ruang.	Determine elements of three-dimensional figures	geometry	x			
29	Menentukan unsur-unsur pada bangun ruang.	Determine elements of three-dimensional figures	geometry	x			
30	Menyelesaikan masalah yang berkaitan dengan kerangka atau jaring-jaring bangun ruang.	Solve a problem related to the frame of three-dimensional figures	geometry	x			
31	Menyelesaikan masalah yang berkaitan dengan volume bangun ruang.	Solve a problem related to the volume of three-dimensional figures	geometry	x			
32	Menyelesaikan masalah yang berkaitan dengan volume bangun ruang.	Solve a problem related to the volume of three-dimensional figures	geometry	x			

Item No	Item description (Indonesian)	Item Description (English)	General concept	PISA Content Area			
				Space & Shape (SS)	Change & Relationships (CR)	Quantity (QY)	Uncertainty & Data (UD)
33	Menyelesaikan masalah yang berkaitan dengan luas permukaan bangun ruang.	Solve a problem related to the surface area of three-dimensional figures	geometry	x			
34	Menyelesaikan masalah yang berkaitan dengan luas permukaan bangun ruang.	Solve a problem related to the surface area of three-dimensional figures	geometry	x			
35	Menentukan ukuran pemusatan atau menggunakannya dalam menyelesaikan masalah sehari-hari.	Determine the central tendency or apply it to solve an every-day problem	statistics				x
36	Menentukan ukuran pemusatan atau menggunakannya dalam menyelesaikan masalah sehari-hari.	Determine the central tendency or apply it to solve an every-day problem	statistics				x
37	Menyelesaikan masalah yang berkaitan dengan penyajian atau penafsiran data.	Solve a problem related to the presentation or interpretation of data	statistics				x
38	Menyelesaikan masalah yang berkaitan dengan penyajian atau penafsiran data.	Solve a problem related to the presentation or interpretation of data	statistics				x
39	Menyelesaikan masalah yang berkaitan dengan peluang suatu kejadian.	Solve a problem related to the probability of an event	probability				x
40	Menyelesaikan masalah yang berkaitan dengan peluang suatu kejadian.	Solve a problem related to the probability of an event	probability				x
		Number of NE items likely to tap on a particular PISA content area		17	9	8	6
		Percentage of NE items likely to tap on a particular PISA content area		.43	.23	.20	.15

Chapter 2

Unpacking the Opportunity-to-learn (OTL) Measures in PISA 2012: The Case of Indonesian Students

2.1. Introduction

Understanding the variability of student academic performance in an international large-scale assessment (ILSA) is crucial in order to identify the rationale and source of such variability, and determine the fairness of the assessment use. In most ILSA programs, such performance is compared across the participating countries and utilized to provide informed knowledge on the extent to which the national education system has worked and is compared to the global system. To make a sound comparison, factors affecting the student performance need to be investigated and measured to predict a better future performance. As past studies have indicated, students are more likely to succeed academically if they are given equitable opportunity to learn, experience, and be exposed to the specific topics and instructional quality (Floden, 2002; Pullin & Haertel, 2008; Schmidt & Maier, 2009). Hence, the notion of opportunity-to-learn (OTL) should encompass a set of circumstances that would enable students to succeed academically.

Using information provided by ILSA, this chapter aims to define possible measures of OTL that can better explain the variability of student performance, especially since it becomes the necessary condition for a successful education process. In line with the call for effective and equitable learning environments as stated by one of the United Nations' newest Sustainable Development Program goals (United Nations, 2016), the Organisation for Economic Co-operation and Development (OECD) also collects information on classroom learning environment along with its cognitive test given to 15-year-old students from over 60 countries in its Program for International Student Assessment (PISA) in 2012. As PISA results are often used to inform national policies in education (Breakspear, 2012), understanding how the classroom-learning factors form measures of OTL that can impact student performance will potentially provide significant insights for each participating country to develop and target education policy reforms. Therefore, the investigation of the OTL measures and how they relate to the student performance may also increase the utility of PISA outcomes for policy development purposes.

School plays a very important role in shaping student learning (Schmidt, Zoido, & Cogan, 2013), and an evaluation of what happens in school may help explain student achievement gaps (Flores, 2007), particularly in mathematics. Compared to reading, mathematics skill is often perceived as a being mainly a product of schooling. The variability in student math achievement may not only be caused by the student's innate ability, but also

affected by having access to positive learning opportunities such as good curriculum, quality teachers, and a supportive learning environment, provided by the school (Flores, 2007). In an effective school, the prescribed curriculum standards are translated into appropriate instruction for students while utilizing adequate infrastructure, human and financial resources. However, due to many factors, the intended curriculum may not be implemented fully and attained appropriately in classroom. Apart from being contingent on the school resources, the implementation of curriculum also depends on teacher capacity to interpret, teach, and assess the prescribed standards. Albeit in the same classroom with the same teacher, each individual student may also benefit from classroom activities at a varying degree (Schmidt & Maier, 2009; Connor et al., 2009). Hence, it is important to find out about students' classroom learning experience as well as teachers' or school principal's views on the provision of such opportunity particularly in the classroom or at school in general, respectively. The opportunity-to-learn (OTL) term used throughout this chapter embraces factors influencing both achievement gaps and opportunity gaps.

As PISA assessment focuses on the overarching cumulative knowledge that needs to be attained by almost leaving-school-age students, sound analysis of PISA outcomes can help in the development of evidence-based reforms in education at national level. The explanatory item response models, introduced in Chapter 1, can be used to deconstruct the cognitive mathematics test scores as well as the non-cognitive OTL-related survey scores while simultaneously controlling for student and school contextual variables (see Adams, Wilson, & Wang, 1997). Among the low performers of PISA, Indonesia – the fourth largest populated country with roughly 250 million inhabitants in 13,000 islands on 700 dialects – is always ranked in the bottom five nations (as discussed in Chapter 1). This chapter describes how the participating Indonesian students in PISA 2012 mathematics respond to each item across different OTL aspects under diverse contexts, and how these responses provide insights on the students' performance. Based on the OTL-related items defined in PISA 2012, I'm proposing a new definition of the OTL measure that can be utilised to provide insights on the classroom-learning factors that may have impacted the students' performance in mathematics. As a result, findings of this study can inform targeted improvements of academic and learning environment for Indonesian students from diverse backgrounds.

2.2. Definitions of Opportunity-to-learn (OTL)

2.2.1. Importance of Opportunity-to-learn

The opportunity-to-learn (OTL) concept was introduced in the 1960s as one of the many external factors that may explain the variability of the student academic performance. Carroll (1963) introduced an early measure of learning by defining the degree of learning as a function of student characteristics (aptitude, ability, perseverance) and school factors (time allowed for learning (defined as OTL), quality of instruction). Along with the concept of mastery learning, Benjamin Bloom explored and utilized this learning function, especially

the OTL aspect, to measure student learning environment factors in the IEA[1]'s First International Mathematics Study (FIMS) (Cogan & Schmidt, 2015). In FIMS, OTL was defined as the student's "opportunity to study a particular topic or learn how to solve a particular type of problem presented by the test" (Husén, 1967a, p. 163). Since then, OTL has been considered as an important policy relevant variable for interpreting student performance in subsequent international comparison studies.

During the No-Child-Left-Behind era, this concept regained its popularity when many local and international studies attempted to emphasize, redefine, and measure the OTL as a significant factor in explaining the variability of students' academic achievement (Stevens & Grymer, 1993; Guiton & Oakes, 1995; MacDonnell, 1995; Porter, 1995b; Herman & Klein, 1997; Wang, 1998; Floden, 2002; Flores, 2007; Schmidt & Maier, 2009). As Stevens and Grymes (1993) indicated the following four main potential OTL-related factors emerging from past studies when explaining differences in the student performance:

1. Content coverage, indicating the extent to which the core curriculum and its contents were covered properly in the learning activities and tests for a particular subject matter and grade level;
2. Content exposure, indicating the extent to which the allocated-time and depth of teaching (time-on-task) is supplied for students to adequately learn and cover the curriculum contents;
3. Content emphasis, indicating the extent to which the selection of topics covers the expected curriculum;
4. Quality of instructional delivery, indicating the extent to which teaching strategies in the classroom are applied to satisfy the educational needs of all students.

This categorization has seemed to cultivate further studies that tend to focus only on one or two particular aspect of OTL. Many of the past studies described in their report used the OTL term to cover concepts related to the first three variables, whilst the quality of teaching defined as the last variable was often considered as a separate entity. Similarly, Schmidt (2009, 2013, 2015) has also conceptualized OTL in terms of content coverage throughout his work. He focused on the content exposure aspect of OTL because it is "the most important in terms of learning the content" (Schmidt & Maier, 2009) and associated it to the student's math literacy as assessed in PISA 2012. When discussing the international comparison of OTL and math performance, Schmidt and his colleagues (2013) found that differences in the content exposure across schools within a country had statistically significant and strong relationships to student performance.

OTL plays important roles in education as either a research indicator, an education indicator, a policy instrument, or even a combination of them (McDonnell, 1995; Guiton & Oakes, 1995). As a research indicator, the OTL concept is used to indicate whether or not the student has the opportunity to study and learn how to solve a problem on a topic being considered. It aims to explain the differences in the international/national benchmarking

[1] IEA stands for the International Association for the Evaluation of Educational Achievement. It is an international, independent organization that collaborates with cross-national research institutions and government bodies to conduct large-scale comparative studies in education (IEA, 2011).

studies that relate curriculum coverage and pedagogical differences to gaps in academic achievement. Next, the concept of OTL becoming an education indicator stems from classroom processes and practices that enable learning for individuals or students with diverse background. The results of OTL measures can provide interim information on what needs to be done to improve the teaching and learning process (Herman, Klein, & Abedi, 2000). Having supportive teachers and a conducive learning environment promotes student engagement that should maximize learning and thus, achievement (Stevens & Grymes, 1993; Flores, 2007). Lastly, the OTL concept has also become an attractive policy tool to equalize educational opportunities in managing the gaps between the advantaged and disadvantaged students (Floden, 2002). It became widely-discussed during the Clinton era when standards-based accountability was introduced, and further emphasized in the next decades. Since then, schools have been held accountable for their students' performance against the defined learning standards. However, when achievement gaps have seemed not to diminish, the advocates of equitable education suggested that the variability in the student achievement might be due to the lack of OTL for some students (Guiton & Oakes, 1995; Floden, 2002; Flores, 2007; Kurz & Elliott, 2014).

A greater emphasis on fairness as a validity issue in the new 2014 standards for educational and psychological testing has further highlighted the importance of OTL in evaluating student learning performance. Defined as "the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test" (AERA, APA, & NCME, 2014, p. 56), OTL plays an essential role in the interpretation and the inference-making of the test scores. Students should not be held accountable for their academic performance if they have not had the opportunity to learn the content/knowledge expected and receive appropriate instructions about them (Porter, 1995b; Wang, 2008; Pullin & Haertel, 2008). It is argued that the government should be responsible for making sure that every school is able to provide good quality of OTL before asking students to perform well in a test (McDonnell, 1995; Schmidt, Zoido, & Cogan, 2013). Hence, OTL bears a significant policy implication, as without considering it, inappropriate interpretations of test results and thus, policies may occur (Porter, 1995a; McDonnell, 1995; Floden, 2002).

Although the definition of the OTL concept may sound trivial, it is not clear how to measure it because of several reasons: (1) the common standards of OTL have yet been defined, (2) the operationalization of OTL aspects depends on their definitions, and (3) data gathering on OTL aspects can be costly and complex (McDonnell 1995; Porter, 1995b; Guiton & Oakes, 1995; Herman, Klein, & Abedi, 2000; Floden, 2002). O'Day and Smith (1993) have advocated a systemic reform for school improvement by developing school standards for accountability purposes. These school standards should contain three parts: (1) *Resource* – the provision of all of the necessary means (e.g. teachers, school materials, curriculum) to allow all students have the opportunity to learn the curriculum contents to a high performance level, (2) *Practice* – the implementation of school programs or activities that provides such opportunities, and (3) *Performance* – the achievement of high performance goals. These definitions of school standards promote aspects of OTL that need to be included and measured in the attempt of leveling out the playing field before any accountability assessment based on test scores is imposed on students, teachers, or school.

However, difficulties in achieving and maintaining a consensus on the acceptable standards have arisen as the sufficiency of the OTL-related standards may depend on the individual characteristics of the students and their interactions with teachers and peers at school (McDonnell, 1995, Floden, 2002). OTL is not just defined by the curriculum exposure and the time allocated for it; how the content was presented, who delivered it, and when such observation/measure was obtained would influence the interpretation of its measurement results (Porter, 1995b; Herman, Klein, & Abedi, 2000; Long, 2014). In addition, ideological perspectives of the researchers or policymakers could also drive the analysis and decision-making related to how OTL has been and will be provided equally and equitably (Guiton & Oakes, 1995). To avoid such conflicting perspectives, the use of multiple measures of OTL obtained at different time frames is strongly recommended; but in doing so, the cost for gathering OTL-related information would be much greater (Porter, 1995a). To sum up, the strength of the OTL measure(s) and its relationship with student performance will depend on how it is defined and operationalized (Floden, 2002; Schmidt & Maier, 2009).

Nevertheless, it is still crucial to obtain empirical information on the OTL aspects because if the OTL indicators are not considered, their effects can be mistakenly attributed to some other factors among so many different factors influencing academic achievement (Floden, 2002). As a curriculum and/or policy variable, OTL indicator(s) can be utilized for explaining the variability in student achievement and identifying effective strategies for educational policy reforms. Without considering OTL as a factor, differences in the academic achievement among some groups of students have often been attributed to differences in intelligence, language ability, or family background (see Flores, 2007). Particularly in mathematics education, inspecting what students have experienced and being exposed to inside the classroom may contribute to the understanding that differential OTL gaps actually exist and have impact on the achievement gaps (NRC, 2001; Akiba, LeTendre, & Scribner, 2007; NCTM, 2012). Hence, a careful identification and measurement of the related OTL aspects warrants an appropriate attention so that tailored interventions can be orchestrated according to the specific needs of the students and/or schools. How the information on these gaps of OTL is translated to policy decisions at either local or national level would determine the success of attempts in closing the student achievement gaps.

2.2.2. Measures of Opportunity-to-learn

Due to the importance of understanding school and classroom learning-related factors toward student achievement, many studies have attempted to define, operationalize, and measure aspects of OTL (see Stevens & Grymes, 1993; Floden, 2002; Schmidt & Maier, 2009; Kurz, 2011). Most of these studies focused on one or two of the curriculum aspects of OTL (i.e. content coverage, exposure, or emphasis), whereas only a few studies have assessed both curriculum and instructional aspects. Depending on how OTL is defined and operationalized, some, if not all, aspects of students' OTL have been found to affect their academic achievement (Muthén et al., 1995; Gamora, Porter, Smithson, & White, 1997; Wang, 1998; Lee, 2004; Tornroos, 2005; Albano & Rodriguez, 2013; Allen et al., 2013; Mo, 2013; Schmidt, Ziodi, & Cogan, 2013). Some of the early studies on the OTL aspects are

described in further detail below as they have provided a basis for understanding the development of dimensions of OTL in this dissertation.

The first IEA's comparative study in mathematics, FIMS, pioneered the inclusion of the OTL aspects in the policy discussion on academic achievement since 1964. In FIMS, where class became the sampling unit (all students in the class would participate in the test is their class was sampled), the class teachers were asked to rate whether the topic that each math test item dealt with had been covered by their students. In this case, FIMS focused on the content coverage/exposure aspect of OTL. It hypothesized that the more topics the test-takers have covered, the higher the students' OTL level in the content aspect would be; which should then give the students more capability and a higher degree of success in answering such items. However, only a modest correlation was found between OTL and student achievement in FIMS. The actual item wording requested the teacher to indicate the proportion of his/her group of students (at least 75%, 25-75%, or under 25%) had had an opportunity to learn the particular *type* of problem (Floden, 2002; Schmidt & Maier, 2009). Some complications might have occurred in interpreting such question. The teacher could misinterpret the wording "this type of problem" to represent either the topic that this item tapped into, or the particular way such a math problem was set up or written, or even the expectation of getting the item response correct. This potential ambiguity might have produced the modest correlation between OTL and achievement, which led to the conclusion that OTL had a limited use in curriculum planning (Husén, 1967b).

In the Second International Mathematics Study (SIMS), a three-level model of curriculum was introduced (Floden, 2002; Schmidt & Maier, 2009). The first level, the *intended* curriculum, denotes the national or local standards set by the government or authority bodies. The second level, the *implemented* curriculum, indicates the curriculum that the teacher teaches in the classroom. While the last level, the *attained* curriculum, refers to the body of knowledge the students have actually learned through their performance on a test. SIMS improved the OTL measure by making it specific to the implemented curriculum: for each item, it asked teachers to predict the percentage of students in the target class would get each item correct without guessing and indicate whether the mathematics needed to answer the item correctly was taught or reviewed during the same school year of the test, had been taught previously, would be taught later, or even was not included in the curriculum (Schmidt & Maier, 1990). These two questions introduced another vagueness as they confounded the students' OTL in learning such math content related to the item with their familiarity with the item. It might also suggest that students should perceive a math problem easy if the teacher had provided adequate OTL by covering the materials needed to answer such item. Although no large effects of OTL on achievement were reported for SIMS, considerable differences in OTL between and within (some) countries were found (Schmidt & Maier, 2009). In this study, OTL started to be recognized as a policy relevant curriculum variable (Floden, 2002).

Countering to the inadequacy of the OTL measures in FIMS and SIMS, Schmidt and his colleagues refined the OTL measures in the next round of IEA's study called the Third International Mathematics and Science Study (TIMSS). TIMSS included items assessing OTL at each level of the previously-described curriculum model (Schmidt & Maier, 2009).

At the intended curriculum level, the national curriculum officer at each participating country was requested to detail the actual curriculum content coverage for each grade-level (K-12). These details were assessed and classified into a list of general topics specific to a grade level. For measuring OTL at the implemented curriculum level, the participating teacher of a particular grade then indicated the amount of instructional time (number of periods) allocated to cover each of the specific topics being tested. Meanwhile, the OTL measure for the achieved/attained curriculum level was illustrated by aligning each test item with a specific topic and providing a sub-score on the respective topic. Furthermore, TIMSS introduced an additional level in the curriculum model called the *potentially-implemented* curriculum, in which textbooks became a measure of OTL as textbook contents often dictated topics being taught in many classroom could explain the variability in the student achievement (Tornroos, 2005; Schmidt & Maier, 2009).

Other examples of OTL-related studies that used large-scale assessment data include two studies using 1992 NAEP math by Muthén et al. (1995) and Lee (2004), and the study of schooling effects in PISA 2012 by Schmidt, Ziodo, and Cogan (2012). Focusing specifically on the aspects of content coverage and emphasis, Muthén and his colleagues investigated the effect of OTL on student achievement (Muthén et al., 1995). Their study analyzed the teacher-reported content emphases on the 1992 NAEP math topics in 4^{th}, 8^{th}, and 12^{th} grade levels and correlated them with the student performance in mathematics while accounting for some school and student background variables. On the other hand, Lee (2004) centered his research on the instructional aspect of OTL by examining the relationship between instructional resources and practices and thus, their associations with the school performance. He used the so-called "progressive instruction" items as the operationalization of OTL, some of which asked the math teachers to indicate the extent to which they put emphasis on reasoning/analysis and communicating math ideas, how often their students worked in small groups, wrote reports/did projects, wrote about problem-solving, discussed math with others, or worked real-life problems, and how often they assessed their students with written responses/projects/portfolios (Lee, 2004, p. 179). These items were used to infer teaching and learning strategies that promoted higher-order cognitive skills. Meanwhile, Schmidt, Zoido, and Cogan (2013) focused on the content exposure aspect of OTL because they considered it as the most important in terms of learning the content (Schmidt & Maier, 2009) when analyzing the PISA 2012 data. They used students' self-report data on their experience, familiarity, and exposure of some math concepts and tasks, and then related these aspects to student performance. Further detail on the OTL measurement in PISA 2012 is described in the next section.

Covering broader aspects of OTL, the Institute for Research on Teaching at Michigan State University introduced the use of multi-dimensional content matrices and teaching logs to assess the alignment between content emphasis, coverage and instructional practices of the curriculum (Porter, 1995a; Floden, 2002; Schmidt & Maier, 2009). The first content matrix lists main topics of a subject matter (e.g. mathematics or science) at a particular grade level, each of which is then broken down into specific sub-topics and their corresponding modes of instruction and levels of expected knowledge/skills. For example in mathematics, the first dimension (Dimension A) contains categories of taught topics, e.g. number and number relations, arithmetic, measurement, statistics, etc. Each of these topics are then broken down

into sub-topics and listed in the second dimension (Dimension B). Within the statistics topic, for instance, Dimension B can encompass such subtopics as types of distributions, measures of central tendency, variability, and regression. Next, the third dimension (Dimension C) lists the modes of instruction (e.g. exposition, pictorial or concrete models, equations/formula, graphs, and so on), whereas the last dimension (Dimension D) identifies the expected knowledge/skill level that the student must attain "as a result of instruction" (e.g. from the level of memorizing facts to building and revising theories/proofs) as described by Porter (1995a, p. 53). Apart from these matrices, teachers were also asked to fill in daily teaching logs at the end of the day and describe instructional time, content, and practices as well as their students' classroom activities. These OTL measures, albeit comprehensively describing the enacted curriculum and how it is related to the intended curriculum, were arduous, costly, and very time intensive.

In the early 2000s the Study of Instructional Improvement (SII) at the Michigan State University focused on the content coverage and the instructional aspect of OTL on mathematics and literacy teaching (Rowan, Harrison, & Hayes, 2004; Rowan & Correnti, 2009). Teachers we asked to select a group of eight target students per class and then report on a teaching log about all contents and instructions given to each of these particular students at different times over one or two year-period. The teaching logs collected information on whether or not specific topics and student tasks became a major/minor focus of one particular day's learning, and on what and how students were asked to work on specific topics (see Rowan, Harrison, & Hayes, 2004). By using data from longitudinal and focused observations, this study asserted that the use of teaching logs was more effective than administering annual surveys as the study outcomes could better explain the central tendency and variability of content coverage and teaching practices over time.

Attempting to measure all four aspects of OTL as defined by Stevens & Grymes (1993), Wang (1998) investigated the teaching and learning practices in 21 eight-grade science classrooms taught by 6 science teachers and examined how these practices associated with the students' performance in specific tests. For the content coverage aspect, each teacher estimated the percentage of their students who would do well in the tests, while the percentage of class periods dedicated to specific science topics constituted the OTL aspect of content exposure. For content emphasis, teaching materials were assessed to yield the percentage of pages in the materials that discussed particular science topics. Lastly, several measures were obtained to operationalize the instructional quality factor such as teacher preparation, integration of concepts, the adequacy of teaching materials, equipment use, and availability of textbooks. Despite the fact that the actual teacher's instructional practice was not assessed, this variable was the most significant predictor of the student post-test score, after controlling for other variables. Content exposure was the most significant predictor of the written test-scores.

Next, Kurz (2011) proposed a new focus in measuring OTL that assessed the time allocated for covering contents and teaching them. Using an online teacher log system, Kurz, Elliott, Kettler, and Yel (2014) defined OTL as the degree to which three enacted curriculum dimensions co-occur during instruction (see Table 2.1) and produces five indicators of the intended curriculum: (1) time on content, (2) content coverage, (3) cognitive processes, (4)

instructional practices, and (5) grouping formats. For all of these indicators, teachers reported the planned and enacted instructional time for covering the prescribed content standards as well as time allocations for covering each of the defined learning topics across different levels of cognitive process as prescribed by the Modified Bloom's Taxonomy, each of the given instructional practices, and each of the grouping formats. The instructional practices delineate a set of teaching and student learning strategies such as direct instruction, visual representation, question-asking, think-aloud, independent practice, guided feedback, reinforcement, and approaches for student knowledge assessment (e.g. quizzes, tests, or other forms). Meanwhile, the grouping format includes an option of individual, small group, or whole class. Hence, teachers would indicate how much time is allocated to each instructional practice within the three grouping formats. Although Kurz and his colleagues focused on the assessment of OTL for special education, their definition of OTL indicators have incorporated past OTL studies and thus, can be applied to the general mainstream education (Kurz, 2011).

Table 2.1
Kurz's opportunity-to-learn (OTL) indices, as adapted from Kurz et al. (2014)

Enacted Curriculum Dimension	*OTL Index*	*Index Definition*
Time	Instructional time	Instructional time dedicated to teaching the general curriculum standards and, if applicable, any custom objectives.
Content	Content coverage	Content coverage of the general curriculum standards and, if applicable, any custom objectives.
Quality	Cognitive processes	Emphasis of cognitive process expectations along a range from lower order to higher order thinking skills.
	Instructional practices	Emphasis of instructional practice along a range from generic to empirically supported practices.
	Grouping formats	Emphasis of grouping formats along a range from individual to whole class instruction.

The Classroom Learning Assessment Scoring System (CLASS) is another instrument that can particularly be used for measuring the instructional quality dimension of OTL (Pianta & Hamre, 2009; Kurz, 2011; Stuhlman, Hamre, Downer, & Pianta, 2015). This classroom-observation instrument was aimed at assessing teacher-student interactions in K-12 classrooms, which then became the underlying framework for the teaching quality part of the measurement of the classroom learning environment in PISA 2012 (OECD, 2014). For an observational study conducted by Allen and his colleagues for secondary school level (Allen et al., 2013), the CLASS framework was revised to be developmentally appropriate for describing secondary school students' interaction with their teachers, consisting of three overarching domains as shown in Figure 2.1 and described in the following paragraphs.

The first domain, *Emotional Support* (ES), encompasses three dimensions*: Classroom Climate*, *Teacher Sensitivity*, and *Regards for Student Perspective*. The *Classroom Climate* refers to how a sense of positive (e.g. warmth and connectedness, communication and affective, respect, relationship) and negative (expressed negativity, punitive control, disrespect) environment occurred in the classroom setting. How teachers responded to student academic or emotional needs and provided supportive classroom atmosphere is reflected in the *Teacher Sensitivity* dimension, while the teachers' ability to recognize and capitalize on students' autonomy, connections, ideas, and interactions among peers as well as the teacher's flexibility in accommodating student perspectives is depicted through the *Regards for Student Perspective* dimension.

The second domain, *Classroom Organization* (CO), reflects the extent to which the teacher is able to (1) use effective methods to encourage positive student behaviors and prevent/redirect misbehaviors (*Behavior Management* dimension), (2) maximize learning time and routines as well as transition time (*Productivity* dimension), and (3) define learning targets and use a variety of effective learning activities for improving student engagement (*Instructional Learning Formats* dimension).

Meanwhile, the *Instructional Support* (IS) domain comprises all of the instructional strategies that foster content understanding, analysis and problem solving, and quality of feedback, each of which taps onto its own dimension. The *Content Understanding* dimension includes strategies that foster the transmissions of content knowledge, comprehension of key ideas, and understanding of facts, concepts, and principles as well as build on background knowledge and correct misconceptions. In the *Analysis and Problem Solving* dimension, teachers are expected to use instructional strategies that can provide opportunities for their students to use and attain higher-order thinking skills (e.g. reasoning, integration, experimentation, problem solving, reflection, and metacognition) during their learning activities and assessments. Under *Quality of Feedback* dimension, the teachers would also be required to allow routine exchanges of feedback with their students, use an appropriate scaffolding approach when delivering feedback or comments on student work, and give encouragement and/or affirmation to students for their accomplishment.

When using this CLASS-S framework to analyze video recordings of classroom interactions in 37 classrooms across 11 schools, positive associations[2] were found between the student achievement scores and the indices of all of the three overarching domains (see Allen et al., 2013). Similarly, the correlations between most of the dimensions and student achievement were also found to be statistically significant[3] and positive, except for the *Negative Climate* dimension, which was negatively correlated with other dimensions as anticipated. Therefore, it can be argued that having more positive indicators in the domains of emotional support, classroom organization, and instructional support in a classroom would point to a higher degree of OTL. As a result, this formulation of OTL domains and dimensions has seemed to be the basis of the development of OTL-related items in PISA 2012 (OECD, 2014).

[2] As indicated by having a positive coefficient when regressing each of the three domains to predict the end-of-year achievement test.

[3] Correlations are statistically significant at $p < .05$ with values ranging from .41 to .86 among the dimensions.

2.2.3. Opportunity-to-learn in PISA 2012

Held every three years, PISA is an international assessment that measures the readiness of 15-year-old students to meet future challenges by testing their proficiencies in mathematics, science, and reading. Participating countries often use PISA results to inform changes in their national educational policies. Investigating the indicators of effective classroom learning environment as defined in PISA can contribute to the narrative of the fairness of the use of the PISA scores to gain a deeper understanding of the student performance. A detailed investigation on how student' classroom learning experience may influence the academic performance will benefit the policy makers to orient the education policy reforms.

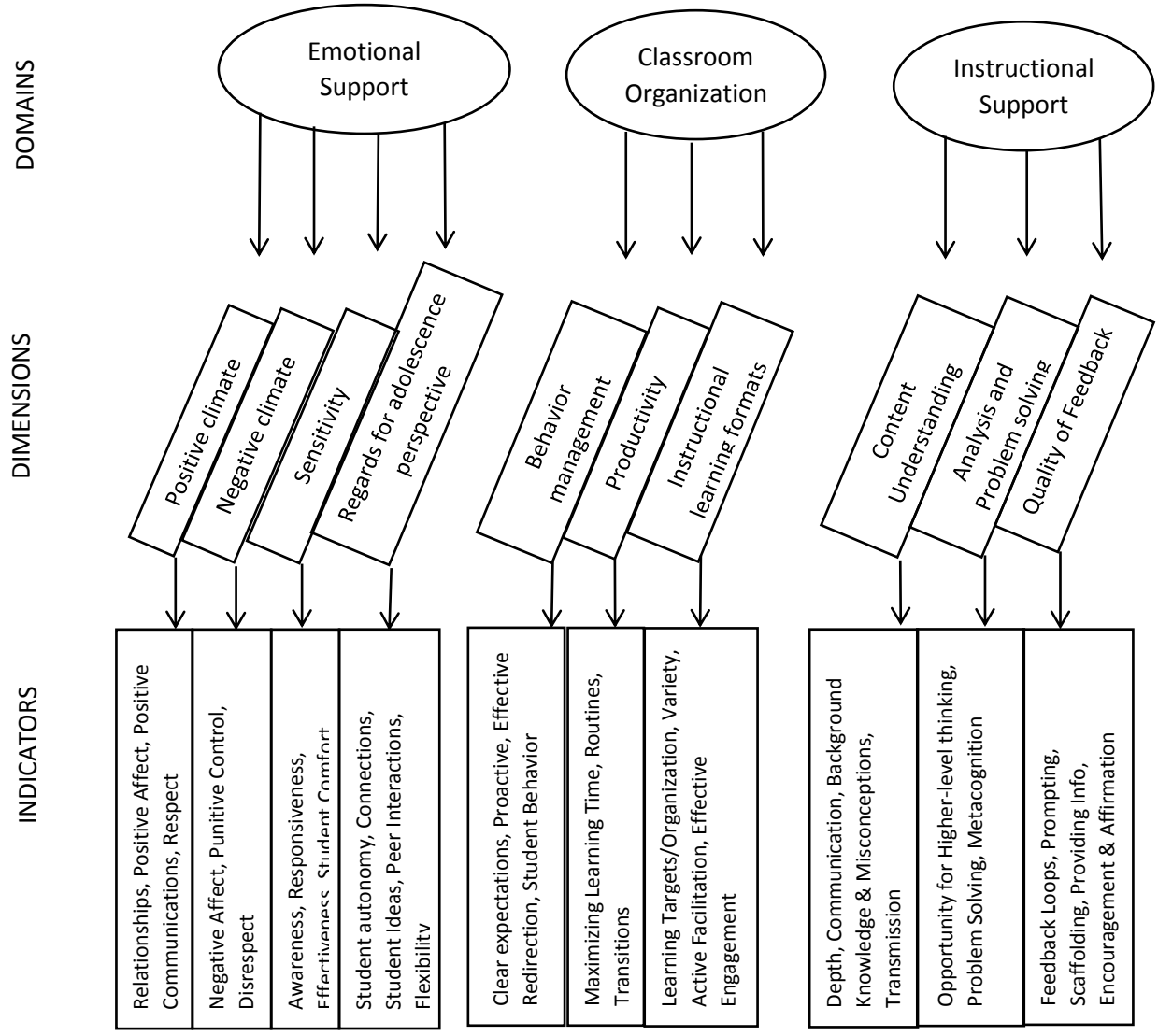


Figure 2.1. The framework of Classroom Learning Assessment Scoring System – Secondary (CLASS-S) instrument, as adapted from Allen et al. (2013, p. 77).

Following the general principles of effective teaching promulgated by Good, Wiley, and Flores (2009), the factors of successful learning described by Stevens and Grymes (1993), and the concept of teaching quality by Pianta and Hamre (2009), PISA 2012 collected information on the aspects of the classroom learning environment and classified the related items into three aspects of OTL, namely *content exposure*, *teaching practices*, and *teaching quality* from their student self-report questionnaire (OECD, 2014). PISA 2012 categorized each of these aspects further into three sub-aspects as shown in Table 2.2, which are described in the following subsections. As mathematics was the focus of the PISA 2012 round, most of the items are concentrated on mathematics learning. All of the items were administered in the student background questionnaire.

Table 2.2
Definition of opportunity-to-learn (OTL) aspects in PISA 2012

Opportunity To Learn (OTL)		
Content Exposure (CE)	**Teaching Practices (TP)**	**Teaching Quality (TQ)**
Experiences with Math Tasks (ST61)	Teacher-directed Instruction (ST79)	Teacher Support (ST77)
Familiarity with Math Concepts (ST62)	Student Orientation (ST79)	Disciplinary Climate (ST81)
Exposure to Types of Math Tasks (ST73-76)	Formative Assessment (ST79)	Cognitive Activation (ST80)

Note. The corresponding OTL-related itemset code administered in PISA 2012 is given in brackets. An itemset represents one survey question that consists of several sub-items.

Content Exposure

The first aspect – *Content Exposure* (CE) – asks students to rate the degree of (1) their familiarity with certain formal mathematics contents, (2) their experiences with various types of math problems during their schooling, and (3) the frequency of being taught to solve specific math problems involving formal or applied mathematics. Table B.1 – B.3 in Appendix B list items corresponding to each of these three CE sub-aspects (thirty items in total). These sub-aspects also reflect factors pertaining to the content coverage and exposure variables of Stevens and Grymes (1993) and the content emphasis of Schmidt (2009). To capture time-on-task or how the students were engaged in learning the curriculum content, time-related scales are utilized: both the familiarity with math concepts sub-aspect (Table B.1) and the exposure to types of math tasks sub-aspect (Table B.3) used a 4-point rating scale with "frequently–sometimes–rarely– never" options. Meanwhile, the sub-aspect of experiences with math tasks (Table B.2) used a 5-point rating scale with "never heard of it – heard of it once or twice – heard of it a few times – heard it often – know it well, understand the concept" response categories. The item set ST76 Example 1 and 2 listed in Table B.3 include two sample PISA math items, which were items from Unit 9 Robberies and Unit 46

Heartbeat, respectively (OECD 2009a; Cogan & Schmidt, 2015). For this CE aspect, the more frequently a student responded to a particular item, the higher the degree of content exposure the particular student is assumed to have on mathematics.

Teaching Practices

The *Teaching Practices* (TP) aspect are depicted through student self-report items on how often the teacher directs instruction (*teacher directed instruction* sub-aspect), orients student participation (*student orientation* sub-aspect), and provides feedback to students on their assessment and learning (*formative assessment* sub-aspect). This aspect denotes the specific practices/strategies of effective teaching, which were adapted from the OECD's Teaching and Learning International Survey (TALIS) (OECD, 2009b). Meanwhile, each of the sub-aspects was grouped and scaled individually to give an index of teacher behavior in the PISA 2012 data analysis for OTL measures. Thirteen items related to the TP aspect/sub-aspects are listed in Table B.4 with a 4-point time rating scale: "every lesson–most lessons–some lessons–never or hardly ever". Highly effective teaching can be demonstrated by having students indicate high frequency of such practices occurred in their math lessons.

Teaching Quality

The aspect of *Teaching Quality* (TQ) embraces three loosely defined characteristics of a supportive learning environment: (1) teacher's socio-emotional support, (2) classroom organization and management, and (3) instructional support/cognitive activation (Pianta & Hamre, 2009; Stuhlman, Hamre, Downer, & Pianta, 2015). First, the sub-aspect of *Teaching Support* (TS), includes questions on how often the teacher shows interest in the student learning, gives extra help when needed, helps students with their learning, continues teaching until the students understand, and allows students to express opinions in the math lessons (see Table B.5). Second, the *Disciplinary Climate* (DC) sub-aspect illustrates several disciplinary-related classroom learning situations that may occur during math lessons such as students not listening to what the teacher say, noise and disorder, the need for the teacher to wait a long time for students to quiet down, the difficulty for students to work well, and a delayed work start-time after the lesson begins (see Table B.6). The TS items were adapted from the PISA 2000 and 2003 student questionnaire forms (OECD, 2002, 2005), while the DC items also appeared in the PISA 2000, 2003, and 2009 rounds (OECD, 2002, 2005, 2012). Both the TS and DC items have a 4-point time scale with "every lesson–most lessons–some lessons–never or hardly ever" response categories. Lastly, the instructional support sub-aspect, referred to as *Cognitive Activation* (CA), asks students to think about the mathematics teacher that taught their last class and indicate how often a series of teaching strategies has happened. As the nine items listed Table B.7 show, these teaching strategies aim to cognitively activate students to develop their mathematical literacy skills by encouraging the students to be able to solve problems in different contexts with multiple solutions (Baumert et al., 2010; OECD, 2014; Burge, Lenkeit, & Sizmur, 2015). These CA items were modified from instruments developed by Kunter and Baumert (2006) as the German extension to the 2003 PISA assessment. Students rate whether each strategy always

or almost always, often, sometimes, or never/rarely happens in their last mathematics class. Hence, a high CA teaching approach is denoted by a high frequency of having such strategies in the classroom.

One significant study on the OTL aspects of PISA 2012 outcomes was performed by Schmidt, Zoido, and Logan (2013) who used only a subset of items under the CE aspect as this aspect was considered to matter the most (Schmidt & Maier, 2009). Following the nature of data sampling, they fit a 3-level model (in ascending order: student-school-country) regressing the students' scores in PISA 2012 math with the derived OTL indices: exposure to formal mathematics (algebra and geometry topics), frequency of encountering word problems, and frequency of encountering applied math problems. These indices were obtained from calculating the average of mean student responses to parts of item set ST62 related to their familiarity of algebra and geometry contents and the sum of dichotomously-scored responses on item set ST74 for the *formal mathematics* indicator, the mean responses to item set ST73 for the *world problems* indicator, and the average of mean student responses to item set ST75 and ST76. Schmidt and his colleagues hypothesized that the variability in OTL was largely due to the differences in the math content exposure the individual students had been having up to the time of survey. These differences might stem from the different types of math course, teachers, or even school tracking system as students were sampled randomly at each selected school in PISA. Using those oversimplified indices, however, this study found significant relationships between the three indices of OTL and the math scores at all three levels. Having more exposure in formal mathematics was shown to give significantly stronger association with student performance in the math test. In addition, there was a significant relationship between the frequency of exposure to applied math problems and the student performance even after controlling for the amount of exposure to formal mathematics. This might suggest that students given more frequent opportunities to do applied math problems in school tended to do better on the math literacy assessment than those who didn't have such opportunities. However, this study also found that in some countries, more frequent exposure to applied math problems did not give additional benefit beyond a certain point (Schmidt, Zoido, & Cogan, 2013). Further research was warranted to examine this intriguing finding.

No study has been carried out using all three overarching dimensions of OTL as defined in PISA 2012. It would also be interesting to see how each of the OTL aspects as originally designed would come together for explaining the variability of the math performance within a country. Hence, this dissertation aims to develop a single overarching and multidimensional OTL measure that can help each participating country utilizing the PISA outcomes in providing evidence for the narratives of national education reforms.

2.3. Research Objectives

From the literature review of OTL-related studies described in the preceding sections, few studies were found that comprehensively measured all four OTL aspects: content

coverage, content exposure, content emphasis, and quality of instruction. Wang (1998) has attempted to do so, but her operationalization of the quality of instructional delivery was not sufficient to investigate classroom dynamics that influenced individual student's opportunity to learn. On the other hand, the CLASS-S framework has only provided a comprehensive operationalization of the effective classroom learning related to the instructional quality aspect of OTL. The Kurz' model of enacted curriculum looks promising as it has defined a list of effective teaching strategies under the instructional practice aspect, but having time (in minutes) as the unit of measurement can be cumbersome and time-consuming; yet its generalizability to a larger population can be somewhat difficult (Kurz, 2011). Hence, this dissertation study proposes an overarching latent measure of multidimensional OTL that embraces all four of Stevens and Grymes' successful classroom learning factors as operationalized by PISA 2012 student questionnaire items.

As described previously, Indonesia has performed poorly in PISA since 2000 (see Table 1.2). Although poor curriculum and low quality of teaching have always been named as the culprit (Kompas, 2013a; Kompas, 2013b, Detik, 2013), there is limited, if any, research publications about which aspect of curriculum and teaching quality need to be improved in Indonesia's context. A new and revolutionary integrative curriculum – Kurikulum 2013 – with innovative teaching strategies for 1^{st} – 12^{th} grade levels was hastily implemented in mid 2013, which was then halted by end of the year with the change of government at the time, due to poor implementation (Kompas, 2014). After being re-evaluated for about 1.5 year, the Ministry of Education and Culture of Indonesia released the national integrative curriculum standards in July 2016 with an intention to better prepare the younger generation for global competition. In light of these new reforms in Indonesia's education system, this dissertation asserts that the investigation of the OTL aspects are crucial to the implementation of the new standards to cater for student demographic patterns.

Using the PISA 2012 data on Indonesian students, in this chapter I investigate the validity of the OTL measures and examine how the aspects of content exposure, teaching practices, and teaching quality relate to the student performance. To do so, I have the following research questions:

1. To what extent can the OTL aspects as defined in PISA 2012 be measured using the related OTL items provided?
    a) How are the OTL measures related to the student math performance in PISA 2012?
2. Is there an alternative model of OTL measures using related items in PISA 2012 that can better explain the student math performance?

The first research question aims to investigate the internal structure of the OTL measures using the prescribed definition by PISA 2012 and how these measures could explain the variability of the student performance in PISA 2012 mathematics. Whereas, the second research question seeks some alternative models of the OTL measures, which still utilize the provided items, examines if and how such alternative models could better explain the variability of the student performance, and finally propose one model that represents the best OTL measures using PISA 2012 items.

2.4. Data Sample

This dissertation uses item responses to the 62 OTL-related items (obtained from the student background questionnaire, see Tables B.1 – B.6) and 84 item level scores of the mathematics test from the Indonesian dataset. The dataset consists of all 5622 participating students from 209 schools. The item parameter estimates used for the math items in this chapter's data analyses are those obtained from the international and national calibration processes performed in Chapter 1.

As PISA uses a two-stage stratified sampling technique in selecting students to participate in the test, the selection probabilities of these students vary and thus need to be accounted for in the parameter estimation process. This variability is mostly due to the school and student sampling design using factors such as school size and school/student non-response rates. Therefore, the given student survey weights are incorporated in all model development and analyses on the Indonesian students, unless otherwise stated. The student survey weight (provided by PISA) is calculated from the school base weight, the within-school base weight, some adjustment factors to compensate for non-participation by school/students, the school base weight trimming factor, and the final student weight trimming factor. Further details about each of these weight components are described in the PISA 2012 Technical Report (OECD, 2014). However, analyses with the randomly sampled international dataset did not use weights. Thus, each nation counts equally to the calibration.

In this PISA round, the student questionnaire also followed a rotational design in which each participating student did not receive the same amount of items as others. The three forms of the questionnaire were distributed randomly to the students (OECD, 2014). Similar with the matrix sampling of the cognitive tests, this approach was taken in order to increase the content coverage of surveyed topics without increasing the response time for the respondent as the administration time was only 30 minutes. Each of these forms has a common part and a rotated part. The common part included items about student demographics, whereas the rotated parts dealt with the attitudinal, perception, and other non-cognitive constructs. The OTL related items were distributed as an intact set across the three different forms in such a way that the missing values can be assumed as random.

2.5. Methods

As one of the modern measurement approaches, item response modeling can deconstruct student responses to the survey items in terms of the assessed content areas. With this modelling approach, I can explain how students responded to each item across different OTL aspects. While accounting for any measurement error and upholding the validity and

reliability standards, this explanatory function contributes to inherent advantages of item response models, particularly Rasch models, for a wide variety of evaluations. One advantage is the ability of validly identifying how particular student responses are aligned with the level of each of the latent constructs being assessed on the same scale, which in this case, is the student's level of a particular aspect of OTL.

For the main data analysis method, I applied unidimensional and multidimensional Rasch models (Rasch, 1960) to examine the student scores in the OTL-related items and the cognitive mathematics test items. With this modelling approach, the given scores can explain how the students' OTL are accommodated and how students perform on each item and across different subdomains and under diverse contexts. This explanatory function contributes to inherent advantages of item response models for a wide variety of evaluations (De Boeck and Wilson, 2004).

A multidimensional random coefficient multinomial logit model (MRCMLM) was employed to calibrate the item parameters and ability estimates (Adams, Wilson, & Wang, 1997). The parameter estimation software ConQuest version 4.5 was used for estimating the parameters (Adam, Wu, & Wilson, 2015). The MRCMLM is a generalized Rasch item response model that uses a scoring function and a design matrix to accommodate the application of many existing IRT models used in this study such as the simple logistic model (Rasch, 1960), the partial credit model (Masters, 1982), the multi-facet model (Linacre, 1994), and the multidimensional versions of these models (Adams, Wilson, & Wang, 1997). Master's partial credit model (PCM) is used to deal with the mixture of dichotomous and polytomous items.

Assume that each item maps to just one of the multiple dimensions (i.e., the model is a so-called between-item model). The multidimensional partial-credit model can be formulated as

$$\eta_{pi} = \theta_{pd} - (\delta_i + \tau_{ik}). \qquad (2.1)$$

wherein $\eta_{pi} = \log\left(\frac{P_{ik}}{P_{ik-1}}\right)$ is the logit link based on the probability (P_{ik}) of answering an item i in response category k, modeled as a linear function of the person latent ability θ on dimension d and the relative item difficulty δ for a particular item i along with its k-th threshold parameter (τ_{ik}). The threshold parameter (τ_{ik}) is the deviation from the mean item difficulty δ_i for item i at step k (i.e. $k = 0, \dots, K$) and constrained such that $\sum_{k=0}^{K} \tau_{ik} = 0$.

I used the MRCMLM model to examine the students' responses within and across different OTL aspects, and calibrate the item parameters and ability estimates. When using the MRCMLM for developing a multidimensional model, however, the estimates of person and item location cannot be directly compared across dimensions because they are on different scales since each dimension is separately centered at zero. Thus, where direct comparison is needed, I used the Delta Dimensional Alignment (DDA) technique (Schwartz, 2012), which is one approach to transform the parameter estimates of a multidimensional model onto the same metric to allow direct interpretation and comparison across dimensions. To obtain the DDA-adjusted item parameters, in the first step, I fit unidimensional model

item parameters to obtain the set of mean item locations ($\mu_{d(uni)}$) and standard deviation ($\sigma_{d(uni)}$) for all the items together. The assumption in using this technique is that there is a latent unidimensional construct underlying all dimensions. In other words, I postulate that, albeit not exactly or necessarily true, all of the dimensions are expected to be "moderately to strongly correlated" (Schwartz, 2012, p. 54). Second, I run the multidimensional model and also calculated the mean ($\mu_{d(multi)}$) and standard deviation ($\sigma_{d(multi)}$) for each subset of items per dimension. Due to the identification constraint used in ConQuest, the mean of each dimension in the multidimensional model is zero. Using the parameter estimates obtained in the two preceding steps, the multidimensional item parameter estimate for each item i in dimension d are then transformed using the following formula:

$$\delta_{id(transformed)} = \delta_{id(multi)} \left(\frac{\sigma_{d(uni)}}{\sigma_{d(multi)}} \right) + \mu_{d(uni)} , \tag{2.3}$$

whereas for the step parameter k:

$$\tau_{ikd(transformed)} = \tau_{ikd(multi)} \left(\frac{\sigma_{d(uni)}}{\sigma_{d(multi)}} \right) . \tag{2.4}$$

These transformed item parameters are then anchored on the next round of model calibration, and hence the resulting model would allow direct interpretation and comparison across dimensions as the dimensions are all on the same scale. Further explanation about the DDA technique can be found in Schwartz (2012).

In order to define an appropriate model for the OTL measures, I performed both qualitative and quantitative investigations on the OTL-related items provided by PISA 2012. The qualitative investigation was carried out by scrutinizing the item wordings and critiquing them using the hypothesized construct of each of the OTL aspects as defined in PISA, as well as other literature. Then, for a quantitative analysis, I investigated the internal structure of such measures by fitting unidimensional and multidimensional partial credit models to test how the OTL-related items supported the underlying constructs. To compare such models and their dimensionality, I calculated each model's Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as well as the EAP/PV reliability value. The EAP/PV reliability is a measure of test reliability output by ConQuest, which is calculated by dividing the variance of the individual EAP ability estimates by the observed person variance (Adams, 2005), and is the IRT equivalent to traditional reliability measures such as Cronbach's alpha. Next, I fitted a structural path model using the 2-stage least squares method (ACER, 2016) to relate the OTL constructs to the student math performance in the PISA 2012 test. In this case, the OTL aspects became the exogenous latent variables, which were assumed to predict the student performance. In order to maintain comparability, the OTL dimensions used anchor parameters (DDA-adjusted) obtained from the preceding model calibration, while the Math dimension used anchor parameters obtained from using the international sample as discussed in Chapter 1. Since the main objective of this study is to define OTL measures that can explain student achievement, I utilized a pragmatic-realist perspective (Maul, Wilson, & Irribarra, 2013) to reach the decision in selecting which model to be used and hence, put a greater emphasis on the extent to which the selected model can explain the variability in the student achievement than on model fit (Floden, 2002; Schmidt & Maier, 2009).

2.6. Findings and Discussion

This section delineates the calibration processes of the OTL measures by applying the Rasch modeling approach for the Indonesian dataset and developing multidimensional models to describe the OTL measures' internal structure, i.e. using the partial credit model (PCM) and investigating the OTL levels. The model fit characteristics of each of the models were calculated, presented, and compared. In this case, the term "dimension" indicates an aspect of OTL being measured, e.g. content exposure or teaching practices, and/or even the sub-aspects of OTL such cognitive activation, disciplinary climates, or a newly proposed aspect as further described in the later section. It is argued that the increasing opportunity to learn with respect to aspects of instruction and knowledge exposure is associated with positive performance (Wang, 1998; Lee, 2004; Allen et al., 2013; Schmidt, Zoido, & Cogan, 2013; Kurz & Elliott, 2014). Hence, for all of the models developed in this chapter, the latent construct θ represents the notion of opportunity to learn of each participating student. In the unidimensional model, θ denotes a single overarching OTL level; whilst in the multidimensional model, there will be several θ-s for each student, each of which represents the latent OTL construct of one dimension.

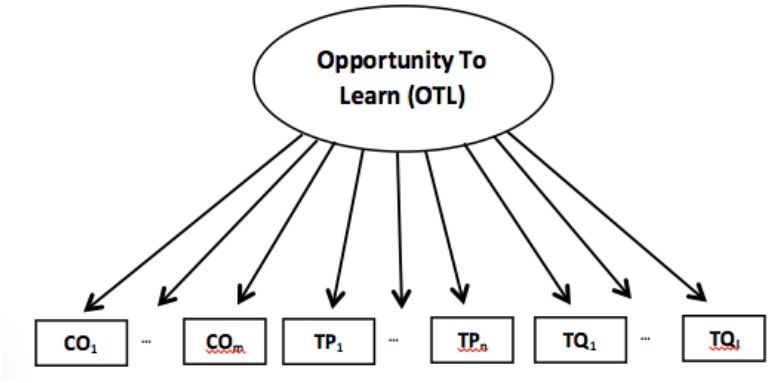
The PCM is used as the items are polytomous or have multiple time-related response categories, ranging from never/rarely/hardly ever to always/almost always/every lesson (see Tables B.1 – Table B.7). In the actual survey, most of these response categories were in the reverse order. Hence, the item responses of those particular items were reversed during model calibration so that higher item difficulty corresponds to higher frequency. Specifically, the PCM models the probability of going from category level j to $j + 1$, such as 2 to 3, given the student has completed the previous step for each item. For the OTL measure(s), there are four or five levels, and thus three or four step parameters, respectively, to be estimated for each item—step 1 represents the difficulty of going from a category of 1 to 2, step 2 is the difficulty of going from a category of 2 to 3 given the student has completed the step from score 1 to 2, and step 3 is the difficulty of going from a score of 3 to 4 given the student has completed the previous steps, and so on.

2.6.1. Research Aim 1 – PISA-defined measures of opportunity-to-learn

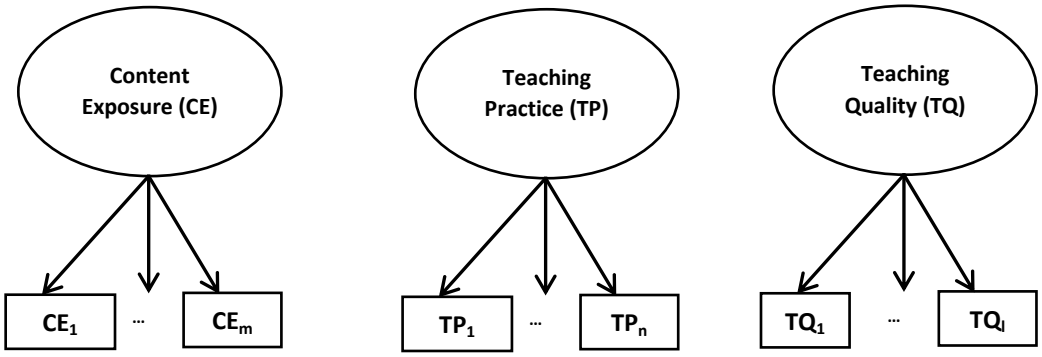
To address the 1st research question, I explored the dimensionality of the OTL measures based on the structure of the concept as defined in PISA 2012 (see Table 2.2) and investigated how these measures would relate to the student performance in PISA 2012 math. I fitted uni- and multi-dimensional models to examine the internal structure of the proposed OTL measure(s) using 62 items as listed in Table B.1 – Table B.7. Figure 2.2 illustrates the models being calibrated with the latent construct depicted in a circle. Given multidimensional models illustrated in Figure 2.2(c) and (d), I allowed correlations between the dimensions as Wang (1998) suggested that the different dimensions of opportunity to learn should be

measured simultaneously. For the investigation of the relationship between the OTL aspects and the student math performance, the latent math ability or literacy is considered as an additional separate dimension. By doing so, the correlations between the math literacy skill and each of the OTL dimensions can be provided.

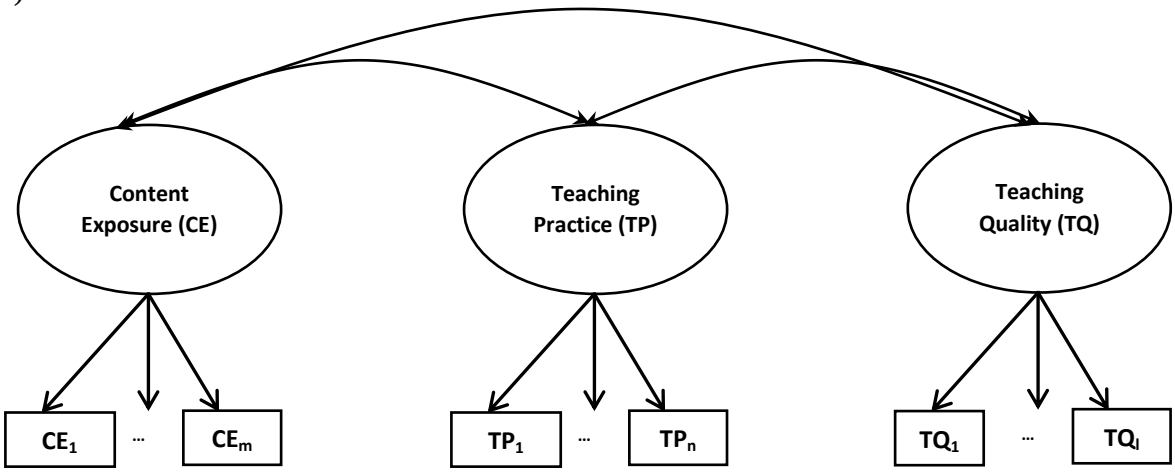
(a)

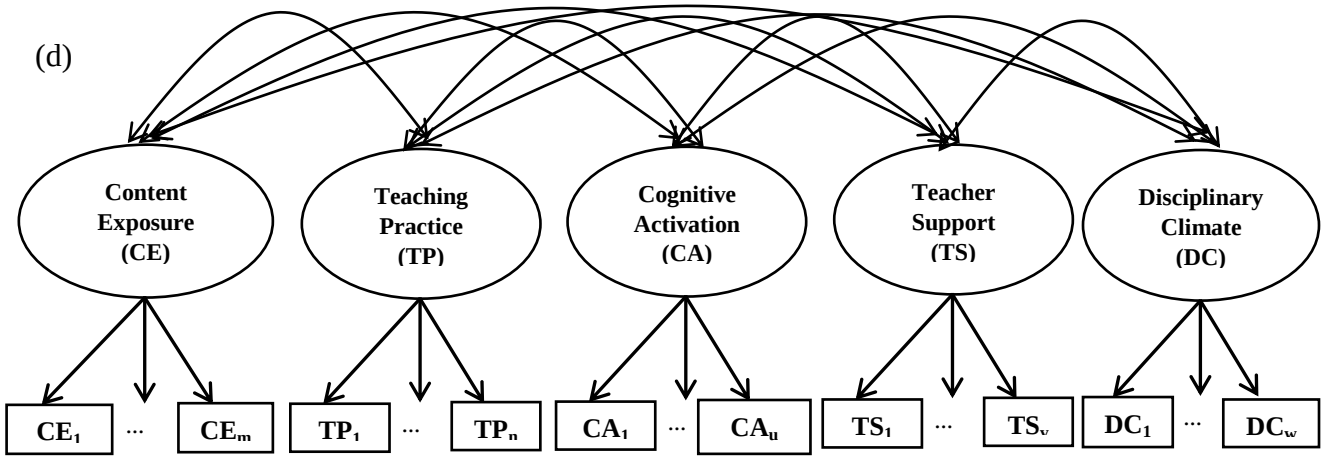


(b)



(c)





Legend:

$CE_1 - CE_m$	: *m* items in "Content Exposure" construct
$TP_1 - TP_n$	: *n* items in "Teaching Practice" construct
$TQ_1 - TQ_l$	: *l* items in "Teaching Quality" construct
$CA_1 - CA_u$	: *u* items in "Cognitive Activation" construct
$TS_1 - TS_v$	: *v* items in "Teacher Support" construct
$DC_1 - DC_w$	: *w* items in "Disciplinary Climate" construct

Figure 2.2. Illustration of (a) the unidimensional, (b) consecutive models, (c) between-item 3-dimensional model, and (d) between-item 5-dimensional model for measuring the opportunity-to-learn in an effective classroom learning environment as defined in PISA 2012.

To decide which models would fit best, I compared the model fit of the unidimensional model of the overarching OTL concept, one-dimensional models resulting from the consecutive modeling approach (Briggs & Wilson, 2003), and the multidimensional models in terms of their AIC[4] and BIC[5] values. From Table 2.3, the unidimensional model appeared to fit worse than the others, while the five-dimensional model fitted the best since it had the lowest AIC and BIC, i.e. $AIC_{1\text{-}D}$ = 527,630 and $BIC_{1\text{-}D}$ = 528,956 vs. $AIC_{5\text{-}D}$ = 512,200 and $BIC_{5\text{-}D}$ = 514,046. These results suggested that the unidimensionality of the PISA-defined indicators of OTL was not supported, but instead, they confirmed the multidimensionality of the OTL concept as found in the past studies (Wang, 1998; Floden, 2002).

First, the consecutive models, as illustrated in Figure 2.2(b), were calibrated by having the OTL aspects of *Content Exposure* (CE), *Teaching Practices* (TP), and *Teaching Quality* (TQ) tapped onto a separate dimension, each of which was not correlated to one another. Next, the three-dimensional model (Figure 2.2(c)) has the three main dimensions of OTL – CE, TP, and TQ – correlated to one another. Finally, the five-dimensional model (Figure 2.2(d)) was derived from the sub-aspect of TQ, i.e. *Cognitive Activation* (CA),

[4] AIC (Akaike Information Criterion) is calculated as "-2 (log likelihood) + 2(number of parameters)".
[5] BIC (Bayesian Information Criterion) is calculated as "-2(log likelihood) + (number of parameters)(ln(sample size)".

Teacher Support (TS), and *Disciplinary Climate* (DC), tapping onto its own individual dimension, in addition to the previous CE and TP dimensions.

Table 2.3
Comparison of models using 62 opportunity-to-learn related items as defined in PISA 2012

No	Model	G^2	AIC	BIC	EAP Reliability
1	Unidimensional (Figure 2.2(a))	527,230	527,630	528,956	.85
2	Consecutive (Figure 2.2(b))	522,011	522,415	523,672	CE = .58
					TP = .55
					TQ = .47
3	Three-dimensional (Figure 2.2(c))	520,963	521,373	522,732	CE = .66
					TP = .64
					TQ = .58
4	Five-dimensional (Figure 2.2(d))	512,200	512,628	514,046	CE = .64
					TP = .60
					CA = .55
					TS = .46
					DC = .51

Note. G^2 is the model deviance or -2(log likelihood), AIC (Akaike Information Criterion) is calculated as "G^2 + 2(number of parameters)", and BIC (Bayesian Information Criterion) is calculated as "G^2 + (number of parameters)(ln(sample size)".

As Model 4 in Table 2.3 shows, the disattenuated correlations – corrected for measurement error – among the OTL aspects in this 5-D model ranged from -.3 to .76 with almost zero correlations for DC-TP and DC-CA as shown in Table 2.4 (i.e. r = -.01 and r = -.08, respectively). Confirming past studies (Schmidt & Maier, 2009; Schmidt, Zoido, & Cogan; 2013), CE has some positive moderate correlations with TP, CA, and TS, but is less correlated with DC. Moderately high correlation is observed for TP-CA as anticipated, particularly since both of these two aspects contain items regarding the math teacher's instructional strategies. This model had good item fit with all infit MNSQ values ranging from .89 to 1.17, which is still within the fit tolerance bounds (i.e. weighted MNSQ of 0.75 – 1.33 (Wilson (2005)).

Although the last 5-D model of OTL fitted the best, it did not seem to be able to explain the student performance well. Adding the *Math Literacy* (MA) dimension to this 5-D model has also given almost zero correlations between MA and TP (r = -.03), MA and TS (r = .05), and MA and DC (r = .06), as shown in Table 2.5. The variance of the DC aspect (σ^2= .89) was higher than the variance of other aspects. Having no strong correlations between these aspects as well as between DC and some other aspects in the previous 5-D model (see Table 2.4) and high within-variance of DC prompted a further qualitative investigation on the DC-related items.

Table 2.4
Correlation (disattenuated) matrix of the 5-dimensional opportunity-to-learn model

Dimension	Dim-1	Dim-2	Dim-3	Dim-4	Dim-5
Dimension 1: Content Exposure					
Dimension 2: Teaching Practice (TP)	.35				
Dimension 3: Cognitive Activation (CA)	.20	.72			
Dimension 4: Teacher Support (TS)	.33	.51	.34		
Dimension 5: Disciplinary Climate (DC)	.14	-.01	-.08	.14	
Variance	.35	.61	.88	.58	1.81
(Std. Err.)	(.007)	(.012)	(.017)	(.011)	(.034)

Note. Number of items = 62, N = 5582.

Table 2.5
Correlation (disattenuated) matrix of the DDA-adjusted 5-dimensional opportunity-to-learn model with the student math literacy as one single dimension

Dimension	Dim-1	Dim-2	Dim-3	Dim-4	Dim-5	Dim-6
Dimension 1: Content Exposure						
Dimension 2: Teaching Practice (TP)	.35					
Dimension 3: Cognitive Activation (CA)	.28	.76				
Dimension 4: Teacher Support (TS)	.35	.49	.34			
Dimension 5: Disciplinary Climate (DC)	.16	-.03	-.11	.18		
Dimension 6: Math Literacy (MA)	.30	-.03	.17	.05	.06	
Variance	.32	.46	.57	.46	.89	.82
(Std. Err.)	(.006)	(.009)	(.011)	(.009)	(.017)	(.016)

Note. No. of items = 62 OTL-related items + 84 math-related items, N = 5622. For model calibration, the OTL-related items use parameters obtained from the DDA-adjusted model as anchors, while the math-related items used the international parameters as obtained and described in Chapter 1.

Besides CE-related items that focused the student's attention on content exposure, most of the instructional strategy-related items in TP, CA, and TS aspects centered their inquiries regarding the student perspective about their math teacher's instructional capacity. DC items, however, seemed vague and ambiguous as they asked students to make judgments about their peers and classroom environments (see Table B.6). For example, students were asked to indicate how frequent the students (probably including oneself) did not listen to what the teacher said, whether s/he and peers could not work well, and if there was noise and disorder. Whereas in the literature, the operationalization of the classroom climate, management, and organization aspects of OTL would include strategies that the classroom teacher would take in dealing with such aspects (Kurz, 2011; Allen et al., 2013). In addition, there was a high variability in the student perception of DC aspect, i.e. variance = 1.81 logits

(see Table 2.4) indicating that the students might have had some misunderstanding in interpreting the items. Because of these reasons, I decided to exclude the DC aspect in my proposed OTL measure(s).

In calibrating the last 6-D model, the OTL-related items were anchored with the parameters obtained from the DDA-adjusted 5-D model, while the math items were anchored with the parameters obtained from the international calibration in math as described in Chapter 1. Having done so, I was able to make a direct comparison only across the OTL dimensions, but not with the MA dimension. To further evaluate how the 5-D OTL aspects affected the math performance, I further ran a structural path model using two-stage least square function in ConQuest by having the performance in math literacy (MA) as the effect variable (see Figure 2.3). The resulting coefficients are presented in Table 2.6 that indicates only 19% of the variance in MA can be explained by the five aspects of OTL as defined by the 6-D model. Having a constant value of -1.94 means that, when each of the OTL measures is 0.0, the mean math literacy skill is -1.94 logits. The CE aspect seemed to have a statistically and significantly positive association with MA, and so did CA, indicating that the higher frequency of students receiving exposure to content and cognitively activated instructional strategies, the student's math literacy tended also to increase. Interestingly, the coefficient for TP is statistically and significantly negative, which could mean that the better performer in PISA 2012 tended not to have higher perception in the frequency of positive teaching practices such as having clear goals for learning set, having to present their thinking or reasoning, having worked in small groups, and so on. Meanwhile, the TS and DC aspects did not seem to give significant effects toward MA by having almost zero coefficients. Since the sensitivity of OTL indicators depends upon their operationalization as manifested in the items (Muthén et al., 1995; Floden, 2002), these results have prompted further qualitative investigation of the OTL-related items and their groupings.

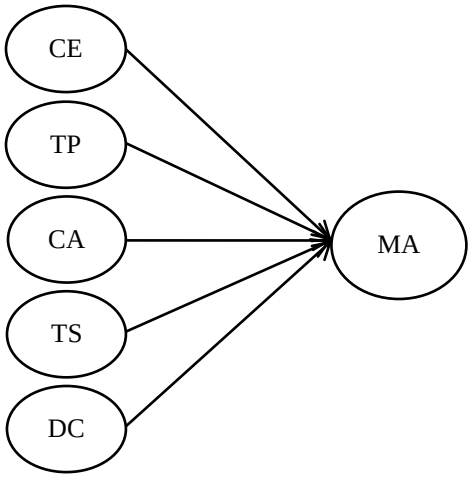
Table 2.6
Structural path model for the 5-dimensional opportunity-to-learn (OTL) model using two-stage least squares having the Indonesian students' math performance in PISA 2012 as the effect variable

Exogeneous Variable	Coefficient	S.E.
Constant	-1.94	.016
Dimension 1: Content Exposure	.53*	.021
Dimension 2: Teaching Practice (TP)	-.64*	.027
Dimension 3: Cognitive Activation (CA)	.53*	.022
Dimension 4: Teacher Support (TS)	.02	.019
Dimension 5: Disciplinary Climate (DC)	.04*	.012
R-Squared	.19	

Note. All models were calibrated using DDA-adjusted parameters for OTL-related items and international item parameters for math items (calibrated in Chapter 1) as anchors. S.E. is the standard error of the coefficient.
* Coefficient is statistically significant with $p < 0.05$.

2.6.2. Research Aim 2 – Proposed alternative models

Apart from the literature review of previous OTL-related measures, I also performed a qualitative investigation of all of the remaining 57 items after excluding the DC items. Then, I developed and proposed three potential models for measuring OTL to better evaluate their impacts on the student math performance. The description of each of these alternative models is given as follows.



Legend:

CE = *Content Exposure* aspect
TP = *Teaching Practices* aspect
CA = *Cognitive Activation* aspect
TS = *Teacher Support* aspect
DC = *Disciplinary Climate* Aspect
MA = *Math Literacy* dimension

Figure 2.3. Illustration of the simple structural path model of the PISA-defined aspects of OTL having the Indonesian students' math performance in PISA 2012 as the effect variable.

2.6.2.1. Alternative Model 1

The first alternative model is based on the original classification of the OTL-related indicators in PISA 2012, but was then slightly modified by having the CA and TS sub-aspects teased out from the TQ aspect to tap onto its own dimension after excluding DC, as illustrated in Figure 2.4. Apart from the CE aspect of OTL, this model focused on the teaching related issues to operationalize the instructional delivery aspect of OTL. Hence, this model is four-dimensional consisting of CE, TP, CA, and TS dimensions. Hypothetically, the higher the degree to which students were exposed to content, positive teaching practices, high cognitively activated strategies, and good teacher support, the higher opportunity to learn such a student would have.

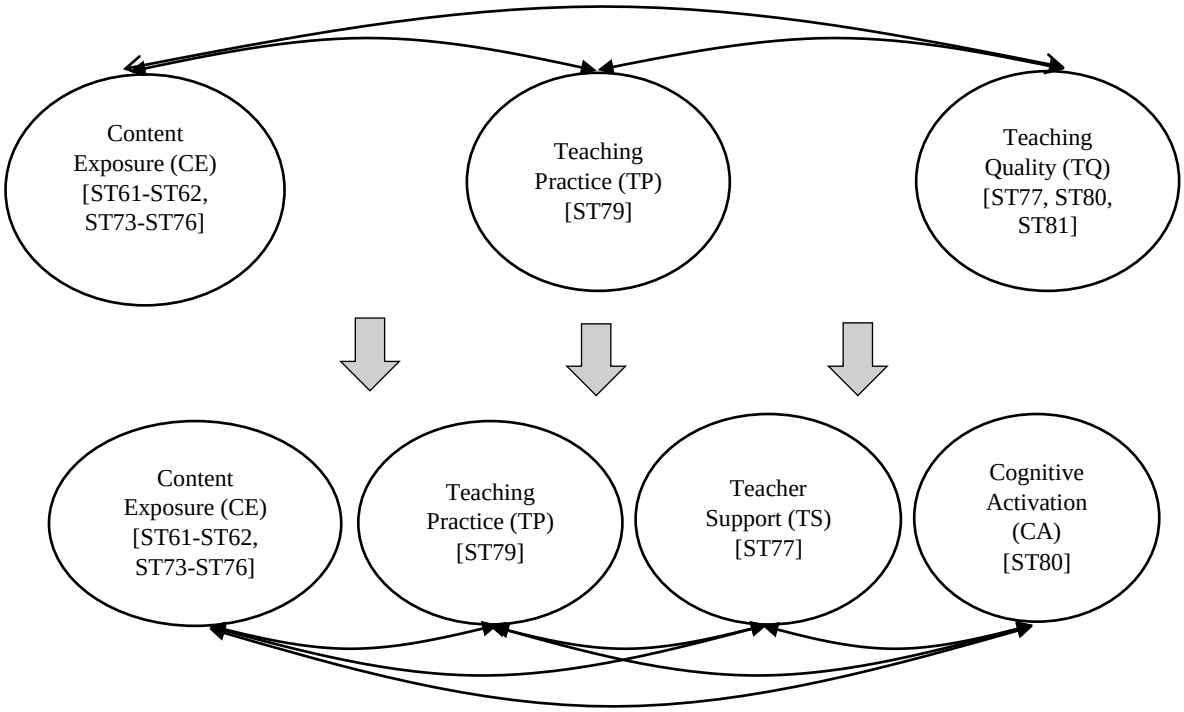


Figure 2.4. Illustration of between-item 4-dimensional Alternative Model 1 proposed for measuring OTL in an effective classroom learning environment.

2.6.2.2. Alternative Model 2

For the second alternative model, I inspected the item wordings of the preceding TP, CA, and TS aspects and reclassified them to fit with the CLASS-S framework (Allen et al., 2013), as introduced in Section 2.2.2. The reclassification of items into the aspects of CLASS-S followed the preliminary definition of the Teaching Quality aspect in PISA 2012. Figure 2.5 illustrates the derivation from the Alternative Model 1 to the Alternative Model 2, in which items from the TP, TS, and CA aspects are regrouped into different dimensions.

1. *Content Exposure (CE)*

This dimension is the same as previously defined representing OTL-related items that indicate the degree to which students were exposed to the specific mathematics contents and tasks (see Table B.1 – Table B.3). For this CE aspect, hypothetically the more frequent a student responded to being exposed to a particular topic/task in math, the higher the degree of content exposure and the more OTL the particular student is assumed to have on mathematics.

2. *Emotional Support* (ES)

There are six items related to the concept of teacher sensitivity and regard for student perceptive as defined in CLASS-S framework. These items can portray not only how teachers respond to and address students' concerns and needs, but also value their

perspectives (see Table 2.7). Having students indicated high frequency of such emotional support given by their teacher in their math lessons would hypothetically show a high degree of teacher support, which in turn would enable students to have a higher OTL.

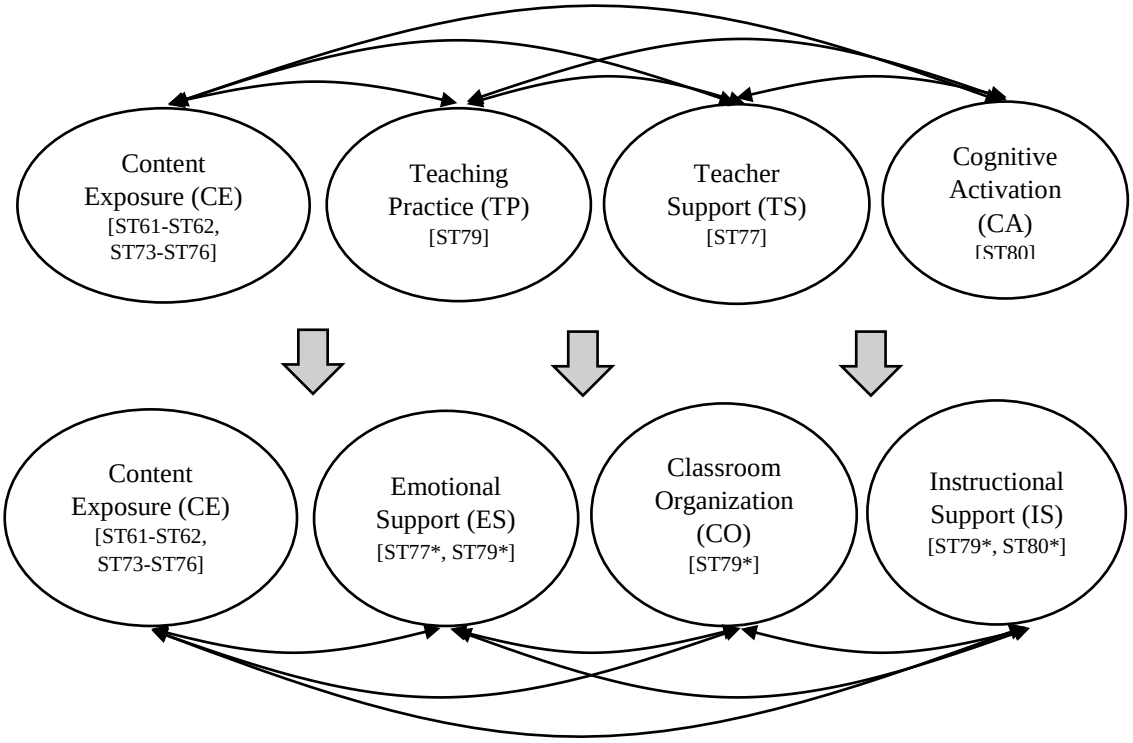


Figure 2.5. Illustration of between-item 4-dimensional Alternative Model 2 proposed for measuring OTL in an effective classroom learning environment.

Table 2.7
Items related to the Teacher's Emotional Support (ES) dimension

Item No	Item code	Item Wording
31	ST77Q01	a) The teacher shows an interest in every student's learning
32	ST77Q02	b) The teacher gives extra help when students need it
33	ST77Q04	c) The teacher helps students with their learning
34	ST77Q05	d) The teacher continues teaching until the students understand
35	ST77Q06	e) The teacher gives students an opportunity to express opinions
41	ST79Q06	f) The teacher asks questions to check whether we have understood what was taught

Note. Item prompt: "How often do these things happen in your mathematics lessons?" Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.

3. *Instructional Support* (IS)

This dimension represents instructional strategies that foster content knowledge, procedural, analysis, and reasoning skills as well as provision of quality feedback to student work. Therefore, 14 items teased out from the TP aspect and all items from the CA sub-aspect are included in this dimension (see Table 2.8). Highly effective and intellectually challenging teaching can be hypothesized by having students indicated high frequency of such practices occurred in their math lessons, which in turn would enable students to have a higher OTL.

Table 2.8
Items related to the Instructional Support (IS) dimension

Item No	Item code	Item Wording
37	ST79Q02[1]	b) The teacher asks me or my classmates to present our thinking or reasoning at some length
39	ST79Q04[1]	d) The teacher assigns projects that require at least one week to complete
40	ST79Q05[1]	e) The teacher tells me about how well I am doing in my mathematics class
45	ST79Q11[1]	j) The teacher gives me feedback on my strengths and weaknesses in mathematics
48	ST79Q17[1]	m) The teacher tells me what I need to do to become better in mathematics
49	ST80Q01[2]	a) The teacher asks questions that make us reflect on the problem
50	ST80Q04[2]	b) The teacher gives problems that require us to think for an extended time
51	ST80Q05[2]	c) The teacher asks us to decide on our own procedures for solving complex problems
52	ST80Q06[2]	d) The teacher presents problems for which there is no immediately obvious method of solution
53	ST80Q07[2]	e) The teacher presents problems in different contexts so that students know whether they have understood the concepts
54	ST80Q08[2]	f) The teacher helps us to learn from mistakes we have made
55	ST80Q09[2]	g) The teacher asks us to explain how we have solved a problem
56	ST80Q10[2]	h) The teacher presents problems that require students to apply what they have learned to new contexts
57	ST80Q11[2]	i) The teacher gives problems that can be solved in several different ways

Note. [1] Item prompt: "How often do these things happen in your mathematics lessons?" Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: "Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen?" Scale category: Always or almost always-Often-Sometimes-Never or rarely.

4. *Classroom Organization* (CO)

Having no items related to behavior management, this dimension only contains items related to productive classroom and strategies for student engagement. These strategies include, but not limited to, development of learning goals and teaching plan, and student grouping for task assignment. Table 2.9 lists the seven related items, derived from several items defined in the TP aspect before. An organized classroom would be hypothetically indicated by having a high frequency of such teaching practice occurred in the math classroom, which in turn would enable students to have a higher OTL.

Table 2.9
Items related to the Classroom Organization (CO) dimension

Item No	Item code	Item Wording
36	ST79Q01	a) The teacher sets clear goals for our learning
38	ST79Q03	c) The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster
42	ST79Q07	g) The teacher has us work in small groups to come up with joint solutions to a problem or task
43	ST79Q08	h) At the beginning of a lesson, the teacher presents a short summary of the previous lesson
44	ST79Q10	i) The teacher asks us to help plan classroom activities or topics
46	ST79Q12	k) The teacher tells us what is expected of us when we get a test, quiz or assignment
47	ST79Q15	l) The teacher tells us what we have to learn

Note. Item prompt: "How often do these things happen in your mathematics lessons?" Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.

2.6.2.3. Alternative Model 3

Besides the original classification of PISA OTL-related indicators as described in Section 2.2.3, the third alternative model was developed after considering factors of effective instructional strategies described by Hattie (2009), the Kurz' OTL standards, and the CLASS-S framework. As illustrated in Figure 2.6, the Alternative Model 3 is derived from the Alternative Model 1 in which items related to TP, CA, and TS aspects were highly inspected, teased out and regrouped to have the following dimensions:

1. *Content Exposure* (CE)

This dimension is the same as previously defined to represent OTL-related items that indicate the degree to which students were exposed to the specific mathematics contents and tasks (see Table B.1 – Table B.3). For this CE aspect, hypothetically the more frequent a student responded to being exposed to a particular topic/task in math, the higher the degree of content exposure and the more OTL the particular student is assumed to have on mathematics.

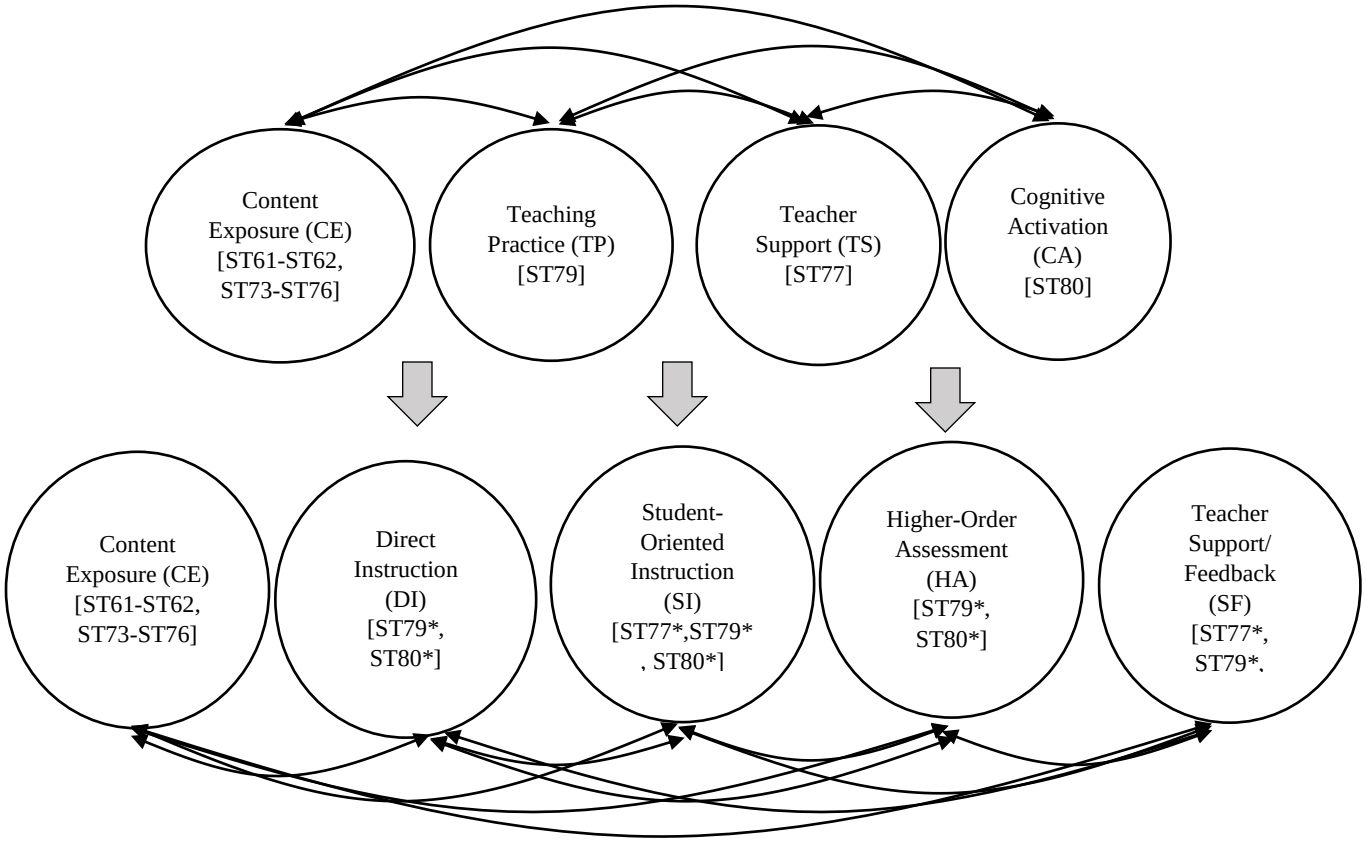


Figure 2.6. Illustration of between-item 5-dimensional Alternative Model 3 proposed for measuring OTL in an effective classroom learning environment.

2. *Direct Instruction* approach (DI)

This dimension encompasses direct teaching practices that state learning goals, inform success criteria, build commitment and engagement, provide guided lessons and practices, provide summary/conclusions, and allow independent practice after direct instruction (Hattie, 2009; Kurz & Elliott, 2014; Stuhlman et al., 2015). The related items were teased out from the TP aspect (Table B.4) and CA aspect (Table B.7), and recompiled in Table 2.10. This classification was similar with the grouping of the *teacher directed instruction*-related items as originally defined in the TP aspect, but it included two extra items: how the teacher sets the expected outcomes from a test, quiz, or assignment (ST79Q12) and how the teacher asks for an explanation of a problem solving approach (ST80Q09). These two extra items seemed to represent the sub-aspect of *teacher directed instruction* more than the *formative assessment* sub-aspect of the TP's original classification. It is hypothesized that having more frequent instructional strategies in which the teacher holds the main authority and becomes dominant would represent a higher degree of direct instruction aspect of teaching. How this aspect would influence a student's OTL may depend on the cultural context of a specific

teaching and learning environment. In Indonesia, for example, the teacher-centered practices were more predominant than the student-centered approaches (Worldbank, 2010).

Table 2.10
Items related to the Direct Instruction (DI) dimension

Item No	Item code	Item Wording
36	ST79Q01[1]	a) The teacher sets clear goals for our learning
37	ST79Q02[1]	b) The teacher asks me or my classmates to present our thinking or reasoning at some length
41	ST79Q06[1]	f) The teacher asks questions to check whether we have understood what was taught
43	ST79Q08[1]	h) At the beginning of a lesson, the teacher presents a short summary of the previous lesson
46	ST79Q12[1]	k) The teacher tells us what is expected of us when we get a test, quiz or assignment
47	ST79Q15[1]	l) The teacher tells us what we have to learn
55	ST80Q09[2]	g) The teacher asks us to explain how we have solved a problem

Note. [1] Item prompt: How often do these things happen in your mathematics lessons? Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen? Scale category: Always or almost always-Often-Sometimes-Never or rarely.

3. *Student-oriented Instruction* approach (SI)

In this student-centered approach, teachers tailor their instruction or assessment according to the student needs such as grouping and allowing student's self selected approach to problem-solving (Hattie, 2009; Allen, et al., 2013). Table 2.11 presents the related items, which were teased out from the TS, TP, and CA aspects. This group contains three items out of five items that operationalized the *student orientation* sub-aspect in the PISA's original definition of TP aspect. The two additional items refer to how the teacher would orient their teaching to suit the student's need (see ST77Q05) and allow some flexibility and autonomy for students in problem-solving (see ST80Q05). Hypothetically, when teachers weigh in student factors for their instructional strategies more frequently, the notion of student-oriented instruction increases. The direction of this student-oriented instruction's effect (SI) on student achievement may be on the contrary of the effect of the direct instruction (DI) approach. Depending on the context, when the DI aspect positively influences the academic achievement, such environment may have a negative effect of the student-oriented approach.

Table 2.11
Items related to the Student-oriented Instruction (SI) dimension

Item No	Item code	Item Wording
34	ST77Q05[1]	d) The teacher continues teaching until the students understand
38	ST79Q03[1]	c) The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster
42	ST79Q07[1]	g) The teacher has us work in small groups to come up with joint solutions to a problem or task
44	ST79Q10[1]	i) The teacher asks us to help plan classroom activities or topics
51	ST80Q05[3]	c) The teacher asks us to decide on our own procedures for solving complex problems

Note. [1] Item prompt: How often do these things happen in your mathematics lessons? Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen? Scale category: Always or almost always-Often-Sometimes-Never or rarely.

4. *Higher-order assessment* (HA)

Based on the cognitive activation concept described by Kunter and Baumert (2006) and Baumerst et al. (2010), this dimension represents the provision of intellectually challenging tasks to students that would require higher-order thinking or cognitive process. It is hypothesized to have a higher degree of higher-order assessment/task when the students have their math teachers assigned them projects that require more than a week to complete, problems with no immediately obvious solutions or many different approaches for solutions, problems that require to think for an extended time, and so on (see Table 2.12). Most of the items were derived from the CA aspect, except one item that asked if the teacher assigned projects with at least a week to be due (ST79Q04). Having taken a week to complete, a project could be assumed to have a considerable degree of complexity. From past studies, being exposed to progressive instructional strategies that promote high-order cognitive skills/thinking can improve students' OTL, which in turn would be positively associated with higher academic performance (Lee, 2006; Kunter & Baumert, 2006; Allen et al., 2013).

5. Quality of *Teacher Support & Feedback* (SF)

The teacher's emotional support and sensitivity (Stuhlman et al., 2015), sense of immediacy (Hattie, 2009, p. 183), and feedback system (Hattie, 2009, p. 173) are captured in this *teacher support and feedback* (SF) dimension. Table 2.13 delineates the corresponding items, which were teased out from the TS, TP and CA aspects. Similarly, this dimension also contains four items related to the original definition of the TS aspect (ST77Q01-06). The additional three items were moved from the *formative assessment* sub-aspect (ST79Q05-17) because they reflected more on how the teacher provides feedback for improvement. Lastly, one item from previously the CA aspect was moved to this dimension as it asked how the teacher helped students to learn from their own mistakes (ST80Q08). When the math teacher shows interest in every student's learning, gives help with student learning and even extra help when

needed, while giving feedback about students' strengths and weaknesses as well as guidance and praises, hypothetically such student would have a high opportunity to learn in the teacher's support and feedback aspect.

Table 2.12
Items related to the Higher-order Assessment (HA) dimension

Item No	Item code	Item Wording
39	ST79Q04[1]	d) The teacher assigns projects that require at least one week to complete
49	ST80Q01[2]	a) The teacher asks questions that make us reflect on the problem
50	ST80Q04[2]	b) The teacher gives problems that require us to think for an extended time
52	ST80Q06[2]	d) The teacher presents problems for which there is no immediately obvious method of solution
53	ST80Q07[2]	e) The teacher presents problems in different contexts so that students know whether they have understood the concepts
56	ST80Q10[2]	h) The teacher presents problems that require students to apply what they have learned to new contexts
57	ST80Q11[2]	i) The teacher gives problems that can be solved in several different ways

Note. [1] Item prompt: How often do these things happen in your mathematics lessons? Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen? Scale category: Always or almost always-Often-Sometimes-Never or rarely.

Table 2.13
Items related to the quality of teacher's support/feedback (SF) dimension

Item No	Item code	Item Wording
31	ST77Q01[1]	a) The teacher shows an interest in every student's learning
32	ST77Q02[1]	b) The teacher gives extra help when students need it
33	ST77Q04[1]	c) The teacher helps students with their learning
35	ST77Q06[1]	e) The teacher gives students an opportunity to express opinions
40	ST79Q05[1]	e) The teacher tells me about how well I am doing in my mathematics class
45	ST79Q11[1]	j) The teacher gives me feedback on my strengths and weaknesses in mathematics
48	ST79Q17[1]	m) The teacher tells me what I need to do to become better in mathematics
54	ST80Q08[2]	f) The teacher helps us to learn from mistakes we have made

Note. [1] Item prompt: How often do these things happen in your mathematics lessons? Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen? Scale category: Always or almost always-Often-Sometimes-Never or rarely.

2.6.2.4. Results from the Alternative Models

For model calibration, I applied the DDA technique to the three alternative models of OTL and used the international math item parameters as anchors for the item parameters in the MA dimension. To find the best-fit model for explaining the effect of OTL aspects on the students' mathematics literacy as assessed in PISA 2012, first I compared the model fit of the three alternative models. As shown in Table 2.14, the AIC and BIC of the Alternative Model 1 are the lowest ($AIC_{Model\ 1}$= 475081, $BIC_{Model\ 1}$= 477006), while the Alternative Model 3 has the highest AIC and BIC ($AIC_{Model\ 3}$= 477404, $BIC_{Model\ 3}$= 478722). However, the differences of the AIC and BIC values among the three models appear to be negligible. Next, I added the MA dimension along with each of these three alternative models to explore how the proposed OTL aspects correlated with the student achievement as represented by the student performance in MA. The results are described as follows.

Table 2.14
Comparison of the alternative models using 57 opportunity-to-learn related items as defined in PISA 2012

No	Model	G^2	AIC	BIC
1	4-dimensional alternative model 1 (Figure 2.4)	474,693	475,081	476,366
2	4-dimensional alternative model 2 (Figure 2.5)	476,644	477,032	478,318
3	5-dimensional alternative model 3 (Figure 2.6)	477,006	477,404	478,722
4	Unidimensional model	483,088	483,458	484.684

Note. G^2 is the model deviance or -2(log likelihood), AIC (Akaike Information Criterion) is calculated as "G^2 + 2(number of parameters)", and BIC (Bayesian Information Criterion) is calculated as "G^2 + (number of parameters)(ln(sample size)".

With the Alternative Model 1, there were two almost-zero correlations between MA and TP (r = -.06), and between MA and TS (r = -.04) as delineated in Table 2.15. Whereas, Alternative Model 2 also failed to give better correlations as MA had almost no correlations with ES, IS, and CO, i.e. r = .09, .07, -.03, respectively (see Table 2.16). Meanwhile, Alternative Model 3 gave better correlational indices among the defined aspects with the MA dimension (see Figure 2.7), and had only one almost zero correlation between MA and SF as demonstrated in Table 2.17. Since the sensitivity of OTL-related items on student achievement can influence results and the success of OTL measures depends on their

operationalization (Muthén et al., 1995; Floden, 2002, Schmidt & Maier, 2009, Cogan & Schmidt, 2015), Alternative Model 3 is found to be the best model representing the OTL measures. Although Alternate Model 3 has the highest AIC and BIC compared to the other two models, I selected it as, consistent with the pragmatic-realist perspective, it would be able to provide more information about the effects of the prescribed OTL aspects toward academic achievement. In addition, all of the prescribed aspects of OTL in Alternative Model 3 gave statistically significant linear associations with MA (Table 2.18). Confirming the previous prediction, the direction of the association between MA-DI (coeff. = .37) and MA-SI (coeff. = -2.14) are of the opposite sign, suggesting that direct instructional practices are more effective than the student-oriented instructions for the Indonesian students. The OTL aspects defined by this model explained about 45% in the variability of MA, whereas the OTL aspects defined by Alternative Model 1 and Model 2 only explained about 17% and 14% of the variability of MA, respectively. When comparing this alternative model's R^2 with that of the previous 5-D model using the PISA-defined aspects (i.e. 19%), one can clearly see that the proposed model can explain the variability of the students' mathematics performance (MA) better, which is mostly due to a better operationalization of the OTL aspects. Apart from the SI aspect, the other OTL aspects did give positive effects on MA.

Table 2.15
Correlation (disattenuated) matrix of the 5-dimensional opportunity-to-learn model developed from the Alternative Model 1 and Math

Dimension	Dim-1	Dim-2	Dim-3	Dim-4	Dim-5
Dimension 1: Content Exposure (CE)					
Dimension 2: Teaching Practice (TP)	.35				
Dimension 3: Cognitive Activation (CA)	.33	.73			
Dimension 4: Teacher Support (TS)	.36	.50	.36		
Dimension 5: Mathematics Literacy (MA)	.29	-.06	.13	.04	
Variance	.34	.53	.61	.53	.85
(Std. Err.)	(.006)	(.010)	(.012)	(.010)	(.016)

Note. No. of items = 57 OTL-related items + 84 math-related items, N = 5622. For model calibration, the OTL-related items used parameters obtained from the DDA-adjusted model as anchors, while the math-related items used the international parameters as obtained and described in Chapter 1.

Table 2.16

Correlation (disattenuated) matrix of the 5-dimensional opportunity-to-learn model developed from the Alternative Model 2 and Math

Dimension	Dim-1	Dim-2	Dim-3	Dim-4	Dim-5
Dimension 1: Content Exposure (CE)		.14	.16	.17	.16
Dimension 2: Emotional Support (ES)	.34		.25	.30	.06
Dimension 3: Instructional Support (IS)	.38	.51		.45	.05
Dimension 4: Classroom Organization (CO)	.39	.59	.85		-.02
Dimension 5: Mathematics Literacy (MA)	.29	.09	.07	-.03	
Variance	.35	.46	.53	.54	.88
(Std. Err.)	(.007)	(.009)	(.010)	(.010)	(.016)

Note. No. of items = 57 OTL-related items + 84 math-related items, N = 5622. For model calibration, the OTL-related items used parameters obtained from the DDA-adjusted model as anchors, while the math-related items used the international parameters as obtained and described in Chapter 1.

Table 2.17

Correlation (disattenuated) matrix of the 6-dimensional opportunity-to-learn model developed from the Alternative Model 3 and Math

Dimension	Dim-1	Dim-2	Dim-3	Dim-4	Dim-5	Dim-6
Dimension 1: Content Exposure (CE)						
Dimension 2: Direct Instruction (DI)	.41					
Dimension 3: Student-oriented Instruction (SI)	.22	.81				
Dimension 4: Higher-order Assessment (HA)	.27	.75	.84			
Dimension 5: Teacher Support/Feedback (SF)	.41	.87	.87	.73		
Dimension 6: Math Literacy (MA)	.29	.13	-.16	.16	.01	
Variance	.34	.68	.41	.56	.35	.86
(Std. Err.)	(.006)	(.013)	(.008)	(.010)	(.007)	(.016)

Note. No. of items = 57 OTL-related items + 84 math-related items, N = 5622. For model calibration, the OTL-related items used parameters obtained from the DDA-adjusted model as anchors, while the math-related items used the international parameters as obtained and described in Chapter 1.

Table 2.18

Regression output of the structural path model for the Alternative Model 3 having the Indonesian students' math performance in PISA 2012 as the effect variable

Exogeneous Variable	Coefficient	S.E.
Constant	-1.71	.013
Dimension 1: Content Exposure (CE)	.16*	.019
Dimension 2: Direct Instruction (DI)	.37*	.025
Dimension 3: Student-oriented Instruction (SI)	-2.14*	.040
Dimension 4: Higher-order Assessment (HA)	1.11*	.024
Dimension 5: Teacher Support/Feedback (SF)	.50*	.042
R-Squared	.45	

Note. All models were calibrated using DDA-adjusted parameters for OTL-related items and international item parameters for math items (calibrated in Chapter 1) as anchors. S.E. is the standard error of the coefficient.

* Coefficient is statistically significant with $p < 0.05$.

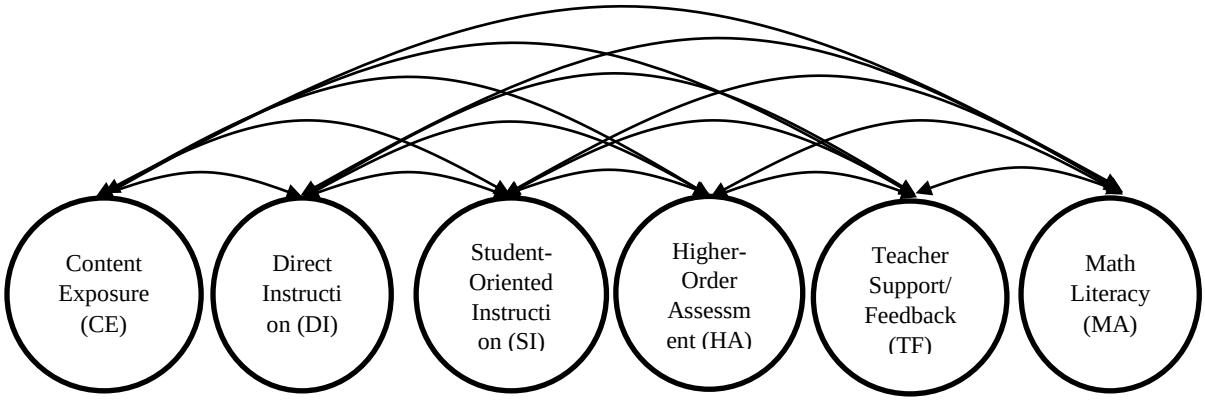


Figure 2.7. Illustration of between-item 6-dimensional model, consisting of the 5-dimensional Alternative Model 3 of OTL and unidimensional Math model, proposed for relating OTL in an effective classroom learning environment to the student's math literacy skills assessed in PISA 2012.

2.6.3. Discussion

To address the two research aims that investigates to what extent the OTL aspects can be measured and related to the student achievement, which in this case is the student performance in PISA 2012 math, several PCM models have been developed for the OTL measures calibration, ranging from the unidimensional model, the consecutive-approach models, and the multidimensional (between-item) models as defined in PISA to several hypothesized multidimensional (between-item) models. Findings from the model calibration suggest that multidimensional models fit better for representing the OTL construct consistent with past studies (Wang, 1998; Floden, 2002). Borrowing the pragmatic-realist perspective to reach decision in defining which model should be used (Maul, Wilson, & Irribarra, 2013), I opted for a model that promotes the highest utility and practicality in measuring OTL using items provided in PISA 2012. Taking into account the issue of item sensitivity when operationalizing the OTL aspects (Muthén et al., 1998) and the idiosyncrasy of the learning context (Cogan & Schmidt, 2015), Alternative Model 3 was then chosen to be the best representation of the OTL measure that can provide more information to explain the relationship between OTL and student performance in PISA 2012 math. More detailed description on the model characteristics of the between-item five-dimensional Alternative Model 3, referred to as *the best-fit OTL measure* throughout the rest of this chapter, is presented below.

2.6.3.1. Item fit

After the DDA approach was applied on the best-fit OTL measure, no item misfit was found. All items are located within the recommended bounds of the infit MNSQ, i.e. 0.75 – 1.33 (Wilson, 2005). It should be noted that many items have a reverse category, and thus these

items' ratings were reversed during the calibration process. As a result, higher difficulty of all items corresponds to higher frequency of positive occurrence. The rating categories have also behaved as expected since the mean location of each response category increases as the response increases for each item.

2.6.3.2. Reliability

The EAP/PV reliability estimates for each dimension of the Alternative Model 3 are as follow: 0.65 for CE, 0.59 for DI, 0.54 for SO, 0.52 for HA, and 0.6 for SF, respectively. As anticipated, this 5-D model has considerably smaller EAP/PV reliability values than that of the unidimensional model (i.e. 0.85 as listed in Table 2.3) since apart from the CE dimension that has 30 items, the other dimensions only contain 5 – 8 items each. These EAP EAP/PV reliability values are higher than those[6] of the consecutive-approach models because they also accounted for the interrelation among the constructs. Hence, the 5-D Alternative Model 3 provides better precision for the estimates of OTL.

2.6.3.3. Evidence of validity

Drawn from the standards of validity evidence developed by the AERA, APA, and NCME (2014), this study has investigated whether the best-fit OTL measure validly assesses what it has claimed to measure in terms of some of the strands of validity evidence. First, in assessing the first piece of validity evidence, which refers to "content validity", the best-fit OTL measure has been developed based on prior work, specifically from the classroom teaching and learning frameworks of Kurz (2011), Pianta and Hamre (2009), and Hattie (2009). The levels of OTL for each of the aspects of the best-fit OTL measure in ascending order can actually be anticipated from the frequency rating of each of the questionnaire items.

Regarding the internal structure of the assessment, the Wright Maps shown in Figure 2.8 – 2.14 suggest that the best-fit OTL measure did adequately span the distribution of the OTL estimates. Figure 2.8 presents the Wright-map of the best-fit OTL measure showing the distributions of the students' OTL estimates (left) for each dimension on the same logit scale as the average item difficulties (right). Having applied the DDA method, comparison of the OTL level across dimension is now possible. The means of OTL level for each dimension are 0.5 logits for CE, 0.46 for DI, 0.4 for SI, 0.39 for HA, and 0.39 for SF. The Thurstonian thresholds for the item-step difficulties for each dimension are depicted on the right columns of Figure 2.6 – 2.15. The item threshold is represented the notation "*i.k*" for item *i* at response category *k*, indicating the location on the latent OTL scale at which the students have a 50 percent chance of having such frequency at or above level *k* for item *i* (Wu et al., 2007).

[6] EAP/PV reliability values of the consecutive approach models of Alternative Model 3 are 0.58 for CE, 0.49 for DI, 0.38 for SO, 0.47 for HA, and 0.42 for SF.

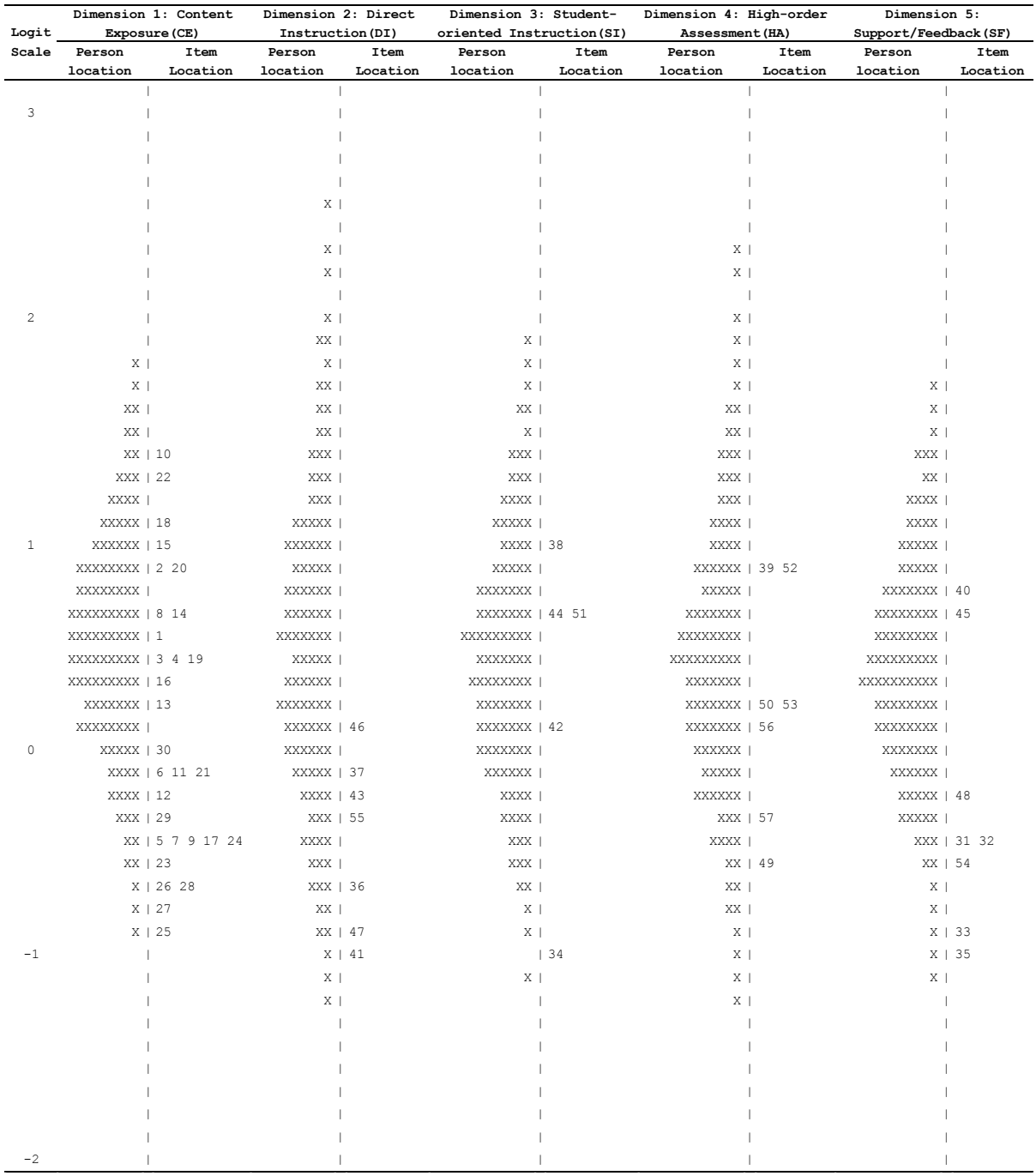


Figure 2.8. A Wright-map of the between-item 5-dimensional Alternative Model 3. Each "x" in the person location distribution represents approximately 48 cases. Numbers 1 to 62 on the item location column denote the item number whose location represents the average item difficulty.

The internal structure of the best-fit OTL measure with respect to each of its aspect/dimension is described as follows:

1. Content Exposure (CE)

Figure 2.8 shows that the CE items adequately spanned the distribution of the students' OTL levels. The distribution of the person estimates was matched by the item threshold difficulties as illustrated in Figure 2.9. Item #10, #22, and #18 appeared to be the most difficult items, which mean that the Indonesian students seemed to have less familiarity with mathematics terms such as exponential function, probability, and polygon, respectively. More than half of the participating students also reported less familiarity to the terms *complex number* (Item #15), *cosine* (Item #20), and *vectors* (Item #14). Since the mean of OTL level in the CE aspect is located at 0.5 logits, any item appeared above the mean would indicate that the students experienced less exposure on such items. Hence, less exposure to specific mathematics tasks such as "Working out from a <train timetable>[7] how long it would take to get from one place to another" (Item #1), "Calculating how much more expensive a computer would be after adding tax" (Item #2), or "Calculating the power consumption of an electronic appliance per week" (Item #8), was also reported. These findings may suggest that there was a lack of math word problems being exposed in math lessons in Indonesia. From a more detailed Wright-map as presented in Figure 2.9, almost half of the participating students only heard once or twice such math terms as probability, exponential function, or polygon. In the sample, about forty percent of the participants was in 9^{th} grade, while forty-six percent was the 10^{th} graders (modal group). These students would have been exposed to early topics in geometry and statistics and probability in the first (Fall) semester of the 9^{th} grade (BNSP, 2006a). Meanwhile, the 10^{th} graders would have been taught about trigonometry and geometry in the second (Spring) semester. Since PISA 2012 was held around late April in Indonesia, there are some potential explanations about these findings: (1) the participating students were unfamiliar with the terminology used in the questionnaire, (2) the students did not grasp the topics well when taught, and thus forgot about them, or (3) the topics had not actually been covered in their math lesson. However, the possibility of the third reason was slim because Indonesian math teachers typically implemented strict curriculum following the national standards.

[7] The term inside < > is to be replaced by a country-specific term that would be familiar to the participating students.

Logit Scale	Dimension 1: Content Exposure (CE)	
	Person location	Item Location
		|
3		|
		|
		|
		|
		| 10.4
	X	| 15.4
	X	| 18.4 22.4
		|
		|
2	X	| 14.4
	XX	| 2.3
	XXXX	|
	XXXX	| 8.3 19.4
	XXXXXXX	| 13.4
	XXXXXXXX	| 1.3 4.3 10.3 12.4
	XXXXXXXXX	| 3.3 22.3
	XXXXXXXXXXXXXX	| 17.4 21.4
	XXXXXXXXXXXXXXXXX	| 18.3
	XXXXXXXXXXXXXXXXXXX	| 10.2 30.3
1	XXXXXXXXXXXXXXXXXXXXXX	| 6.3 15.3 22.2 24.3
	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	| 20.3
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX	| 2.2 18.2
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	| 14.3 23.3 29.3
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	| 8.2 20.2 28.3 ← Mean OTL of 0.5 logits
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	| 5.3 7.3 9.3 15.2 16.3 19.3 22.1
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	| 1.2 3.2 4.2 10.1 26.3
	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	| 13.3 14.2 18.1 20.1
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX	| 11.3 27.3
0	XXXXXXXXXXXXXXXXXXX	| 2.1 12.3 19.2 21.3 25.3
	XXXXXXXXXXXXXXX	| 6.2 16.2 30.2
	XXXXXXXXXXXXXXX	| 13.2 15.1 17.3
	XXXXXXXXXXXX	| 1.1 8.1 14.1
	XXXXXXX	| 3.1 4.1 5.2 7.2 9.2 21.2 24.2 29.2
	XXXXXXX	| 11.2 19.1
	XXX	| 12.2 23.2
	XXX	| 17.2 26.2 28.2
	XX	| 13.1 16.1
-1	X	| 25.2 27.2
	XX	|
	X	| 5.1 6.1 21.1 30.1
		| 7.1 29.1
		| 9.1
		|
		| 23.1 27.1
		| 11.1 12.1 24.1 25.1 26.1 28.1
		|
-2		|
		|
		| 17.1
		|
		|
		|

Figure 2.9. The Wright Map of the 1st dimension – *Content Exposure* aspect – of the between-item 5-dimensional Alternative Model 3. Each 'x' represents 12 cases. The item label shows the threshold level of such item.

2. *Direct Instruction (DI)*

This aspect has been endorsed well with most of the Indonesian students who had expressed high frequency of direct teaching strategies (see Figure 2.8). All items appeared to be overly easy as probably such teacher-dominant practices are common in Indonesian's math lessons (MoECI, 2013b). On average, the Indonesian students reported such practices occurred in most math lessons as shown by having item step 2 located below the mean OTL level of 0.46 logits in Figure 2.10. This one-way or exposition way of teaching style also appeared predominant in the TIMSS' video study of math teaching in Indonesian classrooms (Worldbank, 2011).

Logit Scale	Dimension 2: Direct instruction (DI) Person location	Item Location
	X	
3	X	
	X	
	X	
	XX	
	X	
	XX	
	XXX	
	XX	
2	XXX	
	XXXXXXX	
	XXXXX	
	XXXXXXXXX	
	XXXXXXX	
	XXXXXXXXX	37.3 46.3 55.3
	XXXXXXXXXXX	
	XXXXXXXXXXXXX	43.3
	XXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXX	36.3
1	XXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXX	41.3 47.3
	XXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXX	Mean OTL of 0.46 logits
	XXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXX	
0	XXXXXXXXXXXXXXXXXXXX	46.2
	XXXXXXXXXXXXXXXXXX	37.2
	XXXXXXXXXXXXXXXXX	43.2
	XXXXXXXXXXXX	
	XXXXXXXXXXXXX	
	XXXXXXXXXXXX	55.2
	XXXXXXXXXX	
	XXXXXXXX	36.2
	XXXXXX	
-1	XXXX	
	XXX	41.2 47.2
	XXX	
	XX	46.1
	X	
	X	
	X	
		37.1 43.1
-2		55.1
		36.1
		47.1
		41.1

Figure 2.10. The Wright Map of the 2nd dimension – *Direct Instruction* aspect – of the between-item 5-dimensional Alternative Model 3. Each 'x' represents 12 cases. The item label shows the threshold level of such item.

3. Student-oriented Instruction (SI)

The five items allocated to operationalize this aspect did span the distribution of the person estimates. The student-oriented instructional approach tends to get the students involved in planning and executing the learning in the classrooms. The most difficult item in this aspect is task differentiation based on the students' learning capacity, i.e. the extent to which the teacher would give different work to classmates who have difficulties learning and/or to those who can advance faster (Item #38, see Table 2.11). In the spirit of standardization and equality, this innovative student-center approach has yet been fully implemented in math classrooms throughout Indonesia. Zooming in to the item-step Wright map in Figure 2.11, there was hardly any student who indicated high frequency in getting autonomy to decide their own procedures for complex problem-solving (Item #51). The most common practice of student-centered learning in this case was that the math teacher would continue teaching until the students understand in every classroom (Item #34) as illustrated by having an item location of 34.3 below the mean OTL of 0.4.

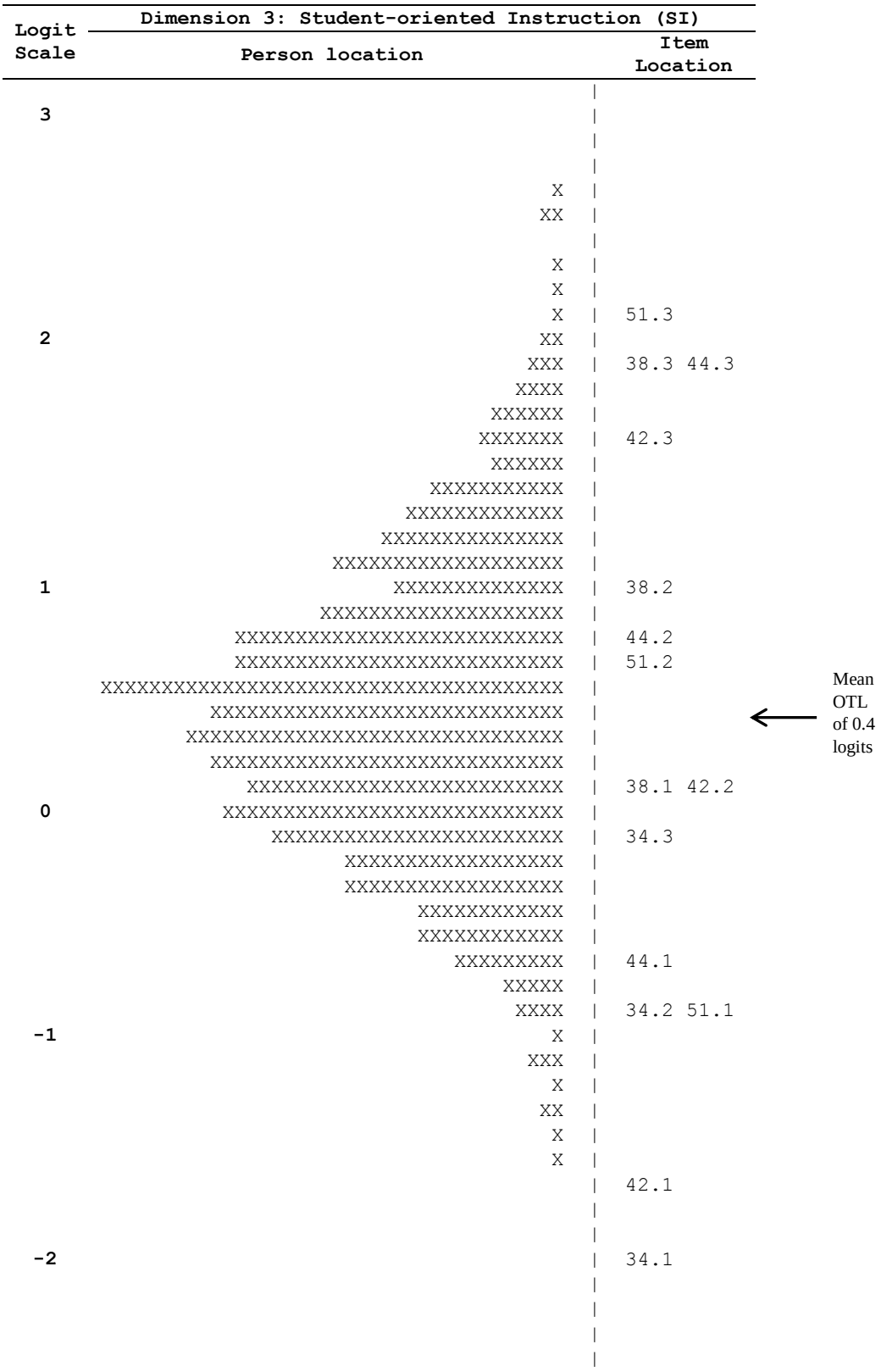

Dimension 3: Student-oriented Instruction (SI)
Logit Scale
Person location
Item Location
3
2
1
0
-1
-2
51.3
38.3 44.3
42.3
38.2
44.2
51.2
38.1 42.2
34.3
44.1
34.2 51.1
42.1
34.1
Mean OTL of 0.4 logits

Figure 2.11. The Wright Map of the 3rd dimension – *Student-oriented Instruction* aspect – of the between-item 5-dimensional Alternative Model 3. Each 'x' represents 12 cases. The item label shows the threshold level of such item.

4. Higher-order Assessment (HA)

Based on the classroom learning framework by Kunter & Baumert (2006), teachers, who challenged their students by giving them cognitively-activated or intellectually-challenged assessment or classroom activity, would create positive atmosphere for learning and hence endorse a higher degree of opportunity to learn. Two difficult items are noted in Figure 2.8, that is a long-term and possibly complex project assessment that would require at least one week to complete (Item #39) and assessment problem with no immediate obvious method of solution (Item #52). Students with an average level of this aspect indicated that these cognitively-activated learning opportunities did not happen often in their typical math lesson. Similarly, the distribution of students' OTL estimates was matched by the item threshold difficulties (see Figure 2.12).

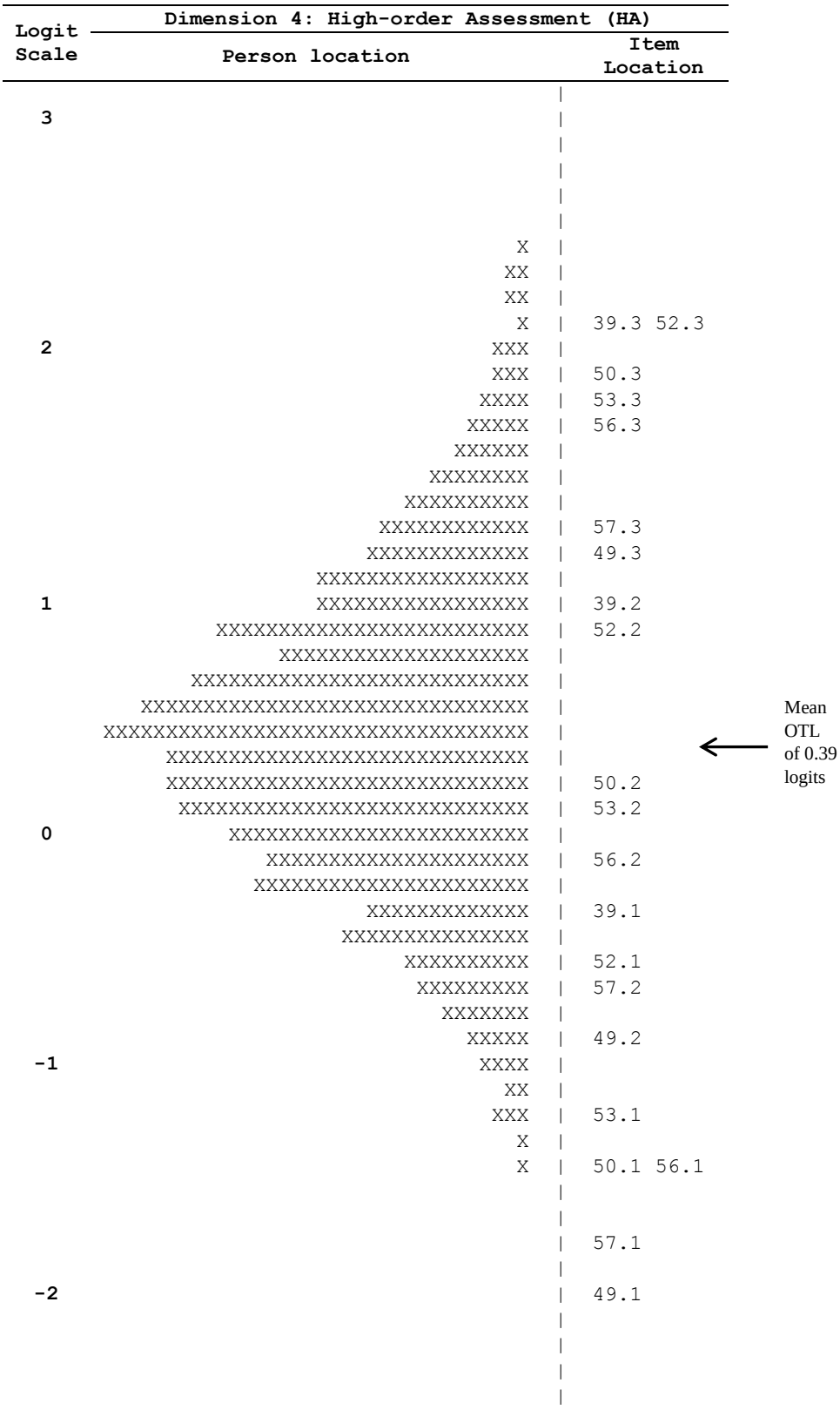

Dimension 4: High-order Assessment (HA)
Logit Scale
Person location
Item Location
3
2
1
0
-1
-2
39.3 52.3
50.3
53.3
56.3
57.3
49.3
39.2
52.2
50.2
53.2
56.2
39.1
52.1
57.2
49.2
53.1
50.1 56.1
57.1
49.1
Mean OTL of 0.39 logits

Figure 2.12. The Wright Map of the 4th dimension – *Higher-order Assessment* aspect – of the between-item 5-dimensional Alternative Model 3. Each 'x' represents 12 cases. The item label shows the threshold level of such item.

5. Teacher Support/Feedback (TF)

As previously discussed, this aspect represents teacher's emotional support and feedback with the goal of improving learning. There seems to be two groups of items that were located above the mean and below the mean as shown in Figure 2.8. This situation can also be seen in Figure 2.13 that illustrates the distribution of item thresholds. The two most difficult items located in the upper part stem from the notion of giving praise to students, that is the extent to which the math teacher informs how well a particular student is doing in the math class (Item #40) and provide feedback on the strengths and weaknesses in mathematics (Item #45). More than half of the participating students found praise and feedback only occurred in some lessons (having the second thresholds of Item #40 and #45 are located above the mean). But, almost half of the participating students with average OTL level reported that they are given an opportunity to express opinions (Item #35) in almost every lesson as shown by having the highest threshold of Item #35 located below the mean.

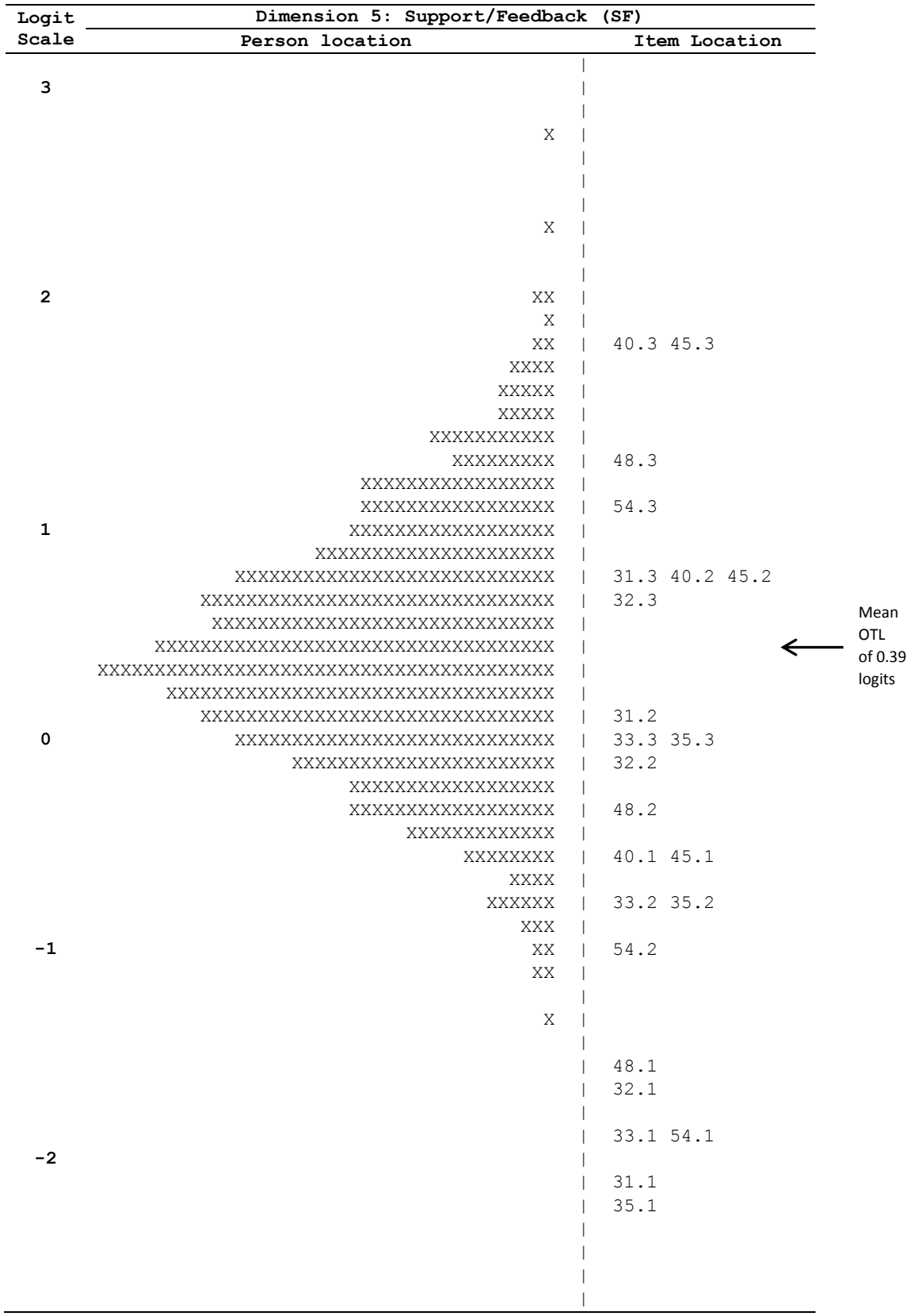

Logit Scale
Dimension 5: Support/Feedback (SF)
Person location
Item Location
3
2
1
0
-1
-2
40.3 45.3
48.3
54.3
31.3 40.2 45.2
32.3
31.2
33.3 35.3
32.2
48.2
40.1 45.1
33.2 35.2
54.2
48.1
32.1
33.1 54.1
31.1
35.1
Mean OTL of 0.39 logits

Figure 2.13. The Wright Map of the 5th dimension – *Teacher's Support/Feedback* aspect – of the between-item 5-dimensional Alternative Model 3. Each 'x' represents 12 cases. The item label shows the threshold level of such item.

The best-fit OTL measure is also useful in providing the disattenuated correlations among the five aspects. Table 2.19 provides the disattenuated and the observed correlations between each of the five domains. The disattenuated correlations are written below the diagonal, while the observed correlations/covariances are above the diagonal. Here, the disattenuated correlations are higher because the measurement error has been removed from the calculation of the coefficients. As shown in Table 2.19, the disattenuated correlations among the instruction-related aspects are moderately high, i.e. ranging from .73 to .81, whilst the CE aspect correlated less with all other aspects, as anticipated (range from .2 to .4). Having some covariance among these aspects support the notion that OTL is indeed a multidimensional concept, but yet comprised of inter-related aspect/dimension.

Table 2.19
Covariance matrix of the 5-dimensional DDA-adjusted Alternative Model 3*

Dimension	Dim-1	Dim-2	Dim-3	Dim-4	Dim-5
Dimension 1: Content Exposure (CE)		.20	.08	.11	.16
Dimension 2: Direct Instruction (DI)	.41		.45	.45	.41
Dimension 3: Student-oriented Instruction (SI)	.19	.80		.40	.33
Dimension 4: Higher-order Assessment (HA)	.26	.76	.86		.32
Dimension 5: Teacher Support/Feedback (SF)	.45	.83	.85	.70	
Variance	.34	.67	.47	.53	.36
(Std. Err.)	(.006)	(.013)	(.009)	(.010)	(.008)

Note. No. of items = 57 OTL-related items, N = 5584. The covariances are depicted above diagonal, whilst the disattenuated correlations are written below diagonal.
* Delta-Dimensional Alignment (DDA) was applied to allow direct comparison across dimension.

For validity evidence based on relations to other variable, I have correlated this best-fit OTL measure with the student performance in PISA 2012 math as an indication of math achievement (Table 2.17) and performed a simple structural model with the math performance as the effect variable as shown in Figure 2.14 with results in Table 2.18. As previously discussed, most aspects of the best-fit OTL model, except the SF aspect, were able to explain the variability in the math performance by having non-zero correlations (except for the SF aspect) and giving statistically significant associations with the math performance, and having the highest variance accounted for.

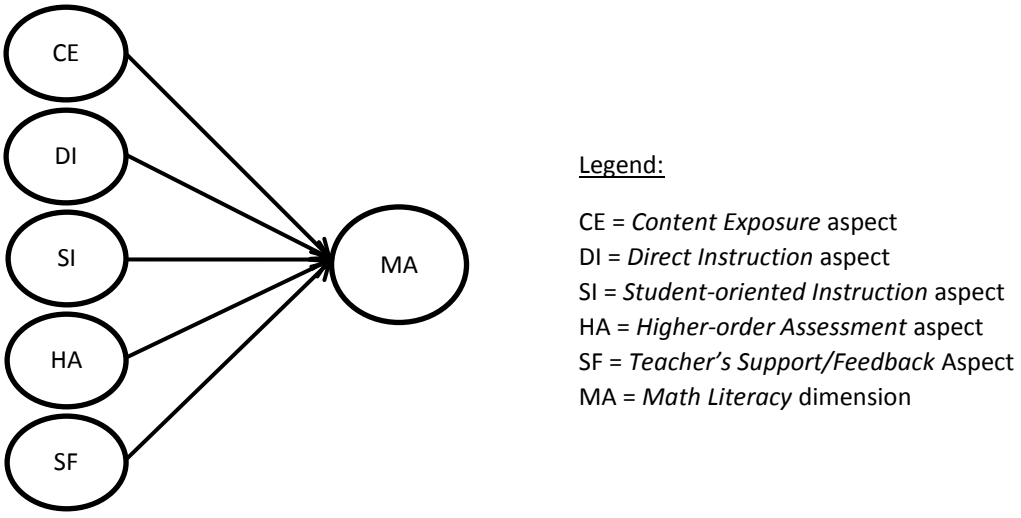


Figure 2.14. Illustration of the simple structural path Alternative Model 3 of the OTL aspects having the Indonesian students' math performance in PISA 2012 as the effect variable.

2.7. Limitations and Future Directions

There are some limitations of this study. First, by applying Rasch item response models, this study assumes that the item scores derived from the 57 items are multidimensional and locally independent of one another (Wilson, 2005). However, some of OTL aspects seem to violate these assumptions both conceptually and empirically. The inter-connected task prompts in the CE-related items suggest task interdependence (see Table B.3). For example, the two ST73 items asked the student to indicate the frequency of encountering the same prompt of a math problem in the math lessons or in the tests taken at school. Future work should address local dependence nature, such as testlet or item-bundle models, can be explored to accommodate the existing task structure.

Second, the fact that the OTL-related indicators in PISA 2012 were obtained from student self-report may potentially bring in a question about the reliability of the ratings. A potential social desirability bias may occur if the sampled students inflate or deflate their perception of their own teacher, which may also depend on the student performance in class. However, it can be argued that these 15-year olds would be mature and experienced enough to recall what they had been exposed to in the math class (Schmidt, Zoido, & Cogan, 2013). Although some items did prompt the students to only think about the mathematics teacher who just taught them in the last mathematics class, I should note that the student ratings on the OTL-related items might also represent their accumulative perceptions of OTL throughout their schooling. To strengthen the students' self assessment of their OTL, albeit

being time and cost intensive, a structured or semi-structured observational survey of classroom interaction may be proposed to better portray the effectiveness of the learning environment.

Furthermore, classroom management and organization is not yet included in the proposed OTL measures as the available *Disciplinary Climate* (DC) items could not represent the anticipated operationalization of the respective aspect, where one may expect more observable items depicting how the math teacher organizes the lesson. Given that a teacher's capability for managing and organizing the classroom can increase student learning (see Porter, 1995a, 1995b; Allen et al., 2013), new items need to be developed to operationalize this aspect. Similarly, more items may also be needed to span the person's OTL level (see Figure 2.8). Currently, most dimensions, which are mainly related to the teaching practices and supports, only contain 5 – 8 items. Increasing the number of items per dimension can be considered for a future avenue of research as it would provide more informative results and a better evidence of reliability.

Finally, the cross-sectional nature of PISA data provides a limitation as the timing of the one-time data collection is not guaranteed to be representative. Therefore, the PISA outcomes should always be interpreted with caution as they might not fully represent the whole picture of the students' OTL being investigated. Should it be required to assess the development or improvement of OTL over a period of time for a national policy evaluation, for instance, expanding it to a longitudinal data collection by tracking some sets of a particular sample of students or schools within a country on each round of PISA may be an option. By doing so, the effectiveness of a curriculum and/or school policy reform can be assessed better.

The data analyses performed in this chapter constitutes one step of the development of OTL measures and prompts further studies in the near future. The demographic variables in the current dataset from each institution would make it possible to perform context effect analysis. Hence, this suggests a need to continuously and rigorously investigate the extent to which the selected student and school background variables impact the OTL aspects and the OTL measures differ between- and within-school. These analyses are performed and discussed in Chapter 3.

2.8. Conclusion

A series to data analyses have been performed in order to investigate (1) to what extent the OTL aspects as defined in PISA 2012 can be measured using the related OTL items provided and related to the student math performance in PISA 2012, and (2) if there can be a better operationalization of OTL measures, to what extent the proposed alternative model(s) of OTL measures can explain the student's math performance in PISA 2012. Models of OTL measures were developed based on the existing definition and operationalization given by PISA 2012 (OECD, 2014), a thorough qualitative investigation

of the related items' wordings, and also on prior work found in the literature of effective classroom teaching and learning (Kunter & Baumert, 2006; Kurz, 2011; Hattie, 2009; Allen et al., 2013). These models were then compared and the best-fit model, Alternative Model 3, was selected according to the pragmatic use of the OTL measures, i.e. to provide insights on how these aspects of OTL can explain the variability in the student academic achievement a.k.a performance in PISA 2012 math assessment. The best-fit OTL measures are comprised of five aspects, i.e. content exposure, direct instruction, student-oriented instruction, higher-order assessment, and teacher's support/feedback.

From the discussion of results in Section 2.6, it can be concluded that the proposed OTL measures are multidimensional, whose effects on the student performance would be highly dependent upon the operationalization of such measures (Muthén et al., 1998; Floden, 2002; Schmidt & Maier, 2009). The disattenuated correlations among the aspects of Alternative Model 3 are low to moderately high and hence, distinct information can be obtained from different OTL aspect estimates. In addition, the best-fit OTL measures have been shown to be able to adequately measure the distribution of the Indonesian students' OTL level as expected. The EAP/PV reliability values for each dimension ranges from 0.5 to 0.65, which shows that the measures can moderately discriminate among the levels of OTL. The range of the OTL measures' item difficulties is appropriate with increasing values of the item thresholds. However, the Wright maps in Figure 2.8 indicate that the items could not adequately span the OTL levels across dimensions, except in the *Content Exposure* dimension. Providing more items in these aspects in the next round of PISA or in any other large-scale OTL-related assessment would be more likely to improve the internal structure validity of the OTL construct itself. Furthermore, the Wright maps suggested that in the typical math lessons, (1) there was a lack of math word problems being exposed, (2) teacher-dominant practices were common, (3) there was a lack of task differentiation based on the students' learning capacity, (4) the stated cognitively-activated learning opportunities happened infrequently, and (5) teachers did not praise and point out students' strength and weaknesses often enough.

Findings from this chapter have two major contributions to the discourse surrounding OTL in large-scale assessment, particularly regarding PISA. First, it provides substantial guidance on the development of the appropriate OTL measures that can explain the variability of OTL across the defined aspects and in turn, the student's academic performance. Second, the proposed OTL measures can also provide a basis for further investigations on the effect of student and school background information across the OTL aspects. The relative differences of the effects can then be utilized to hypothesize ways to improve the curriculum structure, the content pedagogical knowledge, and the school delivery standards as discussed in the next chapter.

Appendix B. Description of Opportunity-to-learn Related Items in PISA 2012

Table B.1
Itemset ST61 related to Content Exposure Aspect – Experience with Math Task sub-aspect
Prompt: "How often have you encountered the following types of mathematics tasks during your time at school?"
Response category: Frequently – Sometimes – Rarely – Never (Scale was reversed for the item parameter estimation)

Item No	Item Code	Item
1	ST61Q01	a) Working out from a <train timetable> how long it would take to get from one place to another
2	ST61Q02	b) Calculating how much more expensive a computer would be after adding tax
3	ST61Q03	c) Calculating how many square metres of tiles you need to cover a floor
4	ST61Q04	d) Understanding scientific tables presented in an article
5	ST61Q05	e) Solving an equation like $6x^2 + 5 = 29$
6	ST61Q06	f) Finding the actual distance between two places on a map with a 1:10 000 scale
7	ST61Q07	g) Solving an equation like $2(x+3) = (x + 3)(x - 3)$
8	ST61Q08	h) Calculating the power consumption of an electronic appliance per week
9	ST61Q09	i) Solving an equation like $3x+5=17$

Note. The word inside <> was to be replaced by a country-specific term.

Table B.2
Itemset ST62 related to Content Exposure Aspect – Familiarity with Math Concepts sub-aspect
Prompt: "Thinking about mathematical concepts: how familiar are you with the following terms?"
Response category: Never heard of it – Heard of it once or twice – Heard of it a few times – Heard it often – Know it well, understand the concept

Item No	Item Code	Item
10	ST62Q01	a) Exponential Function
11	ST62Q02	b) Divisor
12	ST62Q03	c) Quadratic Function
13	ST62Q06	e) Linear Equation
14	ST62Q07	f) Vectors

15	ST62Q08	g) Complex Number
16	ST62Q09	h) Rational Number
17	ST62Q10	i) Radicals
18	ST62Q12	k) Polygon
19	ST62Q15	m) Congruent Figure
20	ST62Q16	n) Cosine
21	ST62Q17	o) Arithmetic Mean
22	ST62Q19	p) Probability

Note. This table excluded the 3-foils items intended for overclaiming detection.

Table B.3

Itemset ST73, ST74, ST75, and ST76 related to Content Exposure Aspect – Exposure to Types of Math Tasks sub-aspect

General Prompt: "The next four questions are about your experience with different kinds of mathematics problems at school. You will see descriptions of problems and grey-coloured boxes, each containing a mathematics problem. Please read each problem. You do NOT need to solve it."

Response category: Frequently – Sometimes – Rarely – Never (Scale was reversed for the item parameter estimation)

Item No	Item Code	Item
	ST73 Prompt	**In the box is a series of problems. Each requires you to understand a problem written in text and perform the appropriate calculations. Usually the problem talks about practical situations, but the numbers and people and places mentioned are made up. All the information you need is given. Here are two examples:** **1) <Ann> is two years older than <Betty> and <Betty> is four times as old as <Sam>. When <Betty> is 30, how old is <Sam>?** **2) Mr <Smith> bought a television and a bed. The television cost <$625> but he got a 10% discount. The bed cost <$200>. He paid <$20> for delivery. How much money did Mr <Smith> spend?** **We want to know about your experience with these types of word problems at school. Do not solve them!**
23	ST73Q01	OTL - How often have you encountered these types of problems in your math lessons?
24	ST73Q02	OTL - How often have you encountered these types of problems in the tests you have taken at school?

	ST74 Prompt	**Below are examples of another set of mathematical skills.** **1) Solve 2x + 3 = 7.** **2) Find the volume of a box with sides 3m, 4m and 5m.** **We want to know about your experience with these types of problems at school. Do not solve them!**
25	ST74Q01	OTL - How often have you encountered these types of problems in your math lessons?
26	ST74Q02	OTL - How often have you encountered these types of problems in the tests you have taken at school?
	ST75 Prompt	**In the next type of problem, you have to use mathematical knowledge and draw conclusions. There is no practical application provided. Here are two examples.** **1) Here you need to use geometrical theorems:** **<graph omitted> Determine the height of the pyramid.** **2) Here you have to know what a prime number is.** **If n is any number: can (n+1)^2 be a prime number?** **We want to know about your experience with these types of problems at school. Do not solve them!**
27	ST75Q01	OTL - How often have you encountered these types of problems in your math lessons?
28	ST75Q02	OTL - How often have you encountered these types of problems in the tests you have taken at school?
	ST76 Prompt	**In this type of problem, you have to apply suitable mathematical knowledge to find a useful answer to a problem that arises in everyday life or work. The data and information are about real situations. Here are two examples.** **1) A TV reporter says "This graph shows that there is a huge increase in the number of robberies from 1998 to 1999." [Chart omitted] Do you consider the reporter's statement to be a reasonable interpretation of the graph? Give an explanation to support your answer.** **2) For years the relationship between a person's recommended maximum heart rate and the person's age was described by the following formula:** **Recommended maximum heart rate = 220 - age** **Recent research showed that this formula should be modified slightly. The** **new formula is as follows:** **Recommended maximum heart rate = 208 – (0.7 × age)** **From which age onwards does the recommended maximum heart rate increase as a result of the introduction of the new formula? Show your work.** **We want to know about your experience with these types of problems at school. Do not solve them!**

29	ST76Q01	OTL - How often have you encountered these types of problems in your math lessons?
30	ST76Q02	OTL - How often have you encountered these types of problems in the tests you have taken at school?

Note. The word inside <> was to be replaced by a country-specific term.

Table B.4
Itemset S79 related to Teaching Practices Aspect
Prompt: "How often do these things happen in your mathematics lessons?"
Response category: Every lesson – Most lessons – Some lessons – Never or hardly ever

Item No	Item Code	Item
36	ST79Q01	a) The teacher sets clear goals for our learning
37	ST79Q02	b) The teacher asks me or my classmates to present our thinking or reasoning at some length
38	ST79Q03	c) The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster
39	ST79Q04	d) The teacher assigns projects that require at least one week to complete
40	ST79Q05	e) The teacher tells me about how well I am doing in my mathematics class
41	ST79Q06	f) The teacher asks questions to check whether we have understood what was taught
42	ST79Q07	g) The teacher has us work in small groups to come up with joint solutions to a problem or task
43	ST79Q08	h) At the beginning of a lesson, the teacher presents a short summary of the previous lesson
44	ST79Q10	i) The teacher asks us to help plan classroom activities or topics
45	ST79Q11	j) The teacher gives me feedback on my strengths and weaknesses in mathematics
46	ST79Q12	k) The teacher tells us what is expected of us when we get a test, quiz or assignment
47	ST79Q15	l) The teacher tells us what we have to learn
48	ST79Q17	m) The teacher tells me what I need to do to become better in mathematics

Note. The highlighted items were included in the Teacher Support aspect in Model A-III.

Table B.5
Itemset ST77 related to Teaching Quality Aspect – Teacher's Emotional Support Sub-aspect
Prompt: "How often do these things happen in your mathematics lessons?"
Response category: Every lesson – Most lessons – Some lessons – Never or hardly ever

Item No	Item Code	Item
31	ST77Q01	a) The teacher shows an interest in every student's learning
32	ST77Q02	b) The teacher gives extra help when students need it
33	ST77Q04	c) The teacher helps students with their learning
34	ST77Q05	d) The teacher continues teaching until the students understand
35	ST77Q06	e) The teacher gives students an opportunity to express opinions

Table B.6
Itemset ST81 related to Teaching Quality Aspect – Disciplinary Climate Sub-aspect
Prompt: "How often do these things happen in your mathematics lessons?"
Response category: Every lesson – Most lessons – Some lessons – Never or hardly ever

Item No	Item Code	Item
58	ST81Q01	a) Students don't listen to what the teacher says
59	ST81Q02	b) There is noise and disorder
60	ST81Q03	c) The teacher has to wait a long time for students to <quiet down>
61	ST81Q04	d) Students cannot work well
62	ST81Q05	e) Students don't start working for a long time after the lesson begins

Note. The word inside <> was to be replaced by a country-specific term.

Table B.7
Itemset ST80 related to Teaching Quality Aspect – Cognitive Activation Sub-aspect
Prompt: "Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen?"
Response category: Always or almost always – Often – Sometimes – Never or rarely

Item No	Item Code	Item
49	ST80Q01	a) The teacher asks questions that make us reflect on the problem
50	ST80Q04	b) The teacher gives problems that require us to think for an extended time
51	ST80Q05	c) The teacher asks us to decide on our own procedures for solving complex problems
52	ST80Q06	d) The teacher presents problems for which there is no immediately obvious method of solution
53	ST80Q07	e) The teacher presents problems in different contexts so that students know whether they have understood the concepts

54	ST80Q08	f) The teacher helps us to learn from mistakes we have made
55	ST80Q09	g) The teacher asks us to explain how we have solved a problem
56	ST80Q10	h) The teacher presents problems that require students to apply what they have learned to new contexts
57	ST80Q11	i) The teacher gives problems that can be solved in several different ways

Chapter 3

Investigating Effects of Background Variables on the Indonesian Students' Opportunity to learn in PISA 2012 Mathematics

3.1. Introduction

Opportunity to learn (OTL) is frequently cited as a significant factor underlying the disparities in academic achievement between individual students, classrooms, schools, and even states/countries (Floden, 2002; Pullin & Haertel, 2008; Schmidt & Maier, 2009). Defined initially as the opportunity that a student has in learning the topics and the ways of solving problems presented in the test, the OTL definition has been broadened to include the quality of instruction experienced by the student as well. Apart from being used as technical validity evidence for achievement test results (Husén, 1967; Floden, 2002), examining students' OTL across all of its related aspects, i.e. content and instructional quality, can also help evaluate the effectiveness of schooling and ensuring equitable education for all students (McDonell, 1995; Porter, 1995b). Before developing an accountability measure or deriving a policy based on test results, understanding how OTL is situated in the specific context under study is essential to fathom its utility for explaining the variability of educational outcomes at all levels. Hence, an investigation on student- and school-level background information influencing the extent to which OTL induces achievement is warranted, specifically when such information are related to the key issues of policy reforms in education: teaching and learning.

This chapter investigates and discusses the utility of student and school background information collected by the Program for International Student Assessment (PISA) to provide insights on how they are associated with the degree of OTL. By understanding deeply how the background information is correlated with each aspect of OTL, specific student and/or school policy can be targeted to improve the curricular and resource allocation management, which in turn leverage the degree of students' OTL. In addition, the use of such information can justify that the anticipated benefit of participating in PISA as it would outweigh the paid cost and potential risk, specifically for developing countries with limited capacity in making evidence-based policy decisions. In the mean time, Indonesia – the fourth largest country in the world in terms of population – always falls short in all of the international tests, including PISA, that it has participated in. Indonesian students have performed very poorly and are always ranked close to the bottom in comparison to students from other countries. For my dissertation, I focus on the PISA data for Indonesia and utilize the student background survey outcomes to examine their association with the students' classroom learning environment as past studies have indicated that good student learning experiences lead to success in student achievement (Gamoran, 1987; Stevens & Grymes, 1993; Floden, 2002; Schmidt & Maier,

2009). This investigation will not only provide significant guidance for policy makers in ameliorating the national educational system and resource allocation, but will also improve the utility of PISA results to support evidence-based policy-making.

The triennial PISA tests, sponsored by the Organization for Economic Cooperation and Development (OECD), collect a vast amount of background information from the sampled students, schools, and parents along with the administration of cognitive tests in mathematics, reading, and science. There is a different test-subject focus in each round, and the 2012 round of PISA focused on mathematics. Mathematics is often perceived as being mainly a product of schooling, and hence an appropriate provision of OTL is consequential in developing mathematical literacy at school. In contrast with a typical mathematics test that assesses student proficiency on standards prescribed in the school mathematics curriculum, the PISA mathematics literacy framework places its emphasis on evaluating the students' "capacity to formulate, employ, and interpret mathematics in a variety of contexts" (OECD, 2013c, p. 25). In other words, PISA assesses the accumulated knowledge in mathematics that can guide students to put mathematics to functional use in their every day lives (Wu, 2003). In order to support the attainment of these necessary life skills, student- or school-level background variables related across all aspects to the student's opportunity to learn mathematics at school need to be investigated. In particular, due to the two-stage stratified sampling method used by PISA, students were randomly sampled within the selected schools, and thus, some school clustering effect may result when analysing the degree of OTL.

3.2. Concept of Opportunity-to-learn (OTL)

This section reintroduces the concept of opportunity to learn as described in greater detail in Chapter 2. The OTL concept was first presented by Carroll (1963) who postulated that learning is a function of student and school factors. When identifying factors related to the success of school learning, he described OTL as time allowed for learning. This definition was embraced and expanded by Benjamin Bloom to fit in with his own concept of mastery learning, and used in the First International Mathematics Study (FIMS, Cogan & Schmidt, 2015). In FIMS, OTL referred to the opportunity that a student has to learn the particular topics and the ways to solve a problem as presented in the test (Husén, 1967, p. 162). Since then, the notion of OTL has been incorporated in most subsequent ILSAs to help explain the variability of the student performance across nations, for example: the IEA[1]'s Second International Mathematics Study (SIMS) and the Third International Mathematics and Science Study (TIMSS). It was not until the 2012 round that the OECD's PISA formally introduced the concept of OTL by collecting indicators of classroom learning environment in the student questionnaire.

[1] IEA stands for the International Association for the Evaluation of Educational Achievement. It is an international, independent organization that collaborates with cross-national research institutions and government bodies to conduct large-scale comparative studies in education (IEA, 2011).

The concept of OTL gained popularity during the NCLB era in the 1990s to 2000s, in which the notion of OTL was emphasized, redefined, and measured to validate the needs and the implementation of standardized state testing whose results became crucial for school accountability purposes (Porter, 1995b; McDonell, 1995). In 1993, Stevens and Grymer (1993) identified four main OTL aspects, related to successful learning, emerging from past studies: content coverage, content exposure, content emphasis, and quality of instructional delivery. Many other studies have also found OTL to be a significant factor in explaining the variability of students' academic achievement (Guiton & Oakes, 1995; MacDonnell, 1995; Porter, 1995b; Herman & Klein, 1997; Wang, 1998; Floden, 2002; Flores, 2007; Schmidt & Maier, 2009).

The importance of this idea of OTL has been highlighted by having it explicitly prescribed in the new 2014 Testing Standards, which defines OTL as "the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test" (AERA, APA, & NCME, 2014, p. 56). The new Testing Standards consider OTL as a fairness issue wherein students should not be held accountable for their academic performance unless they have had the opportunity to learn the content/knowledge expected and have received appropriate instructions about them (Porter, 1995b; Wang, 2008; Pullin & Haertel, 2008). Hence, the OTL concept embodies the student's exposure to both good quality of content knowledge and instructional experience. As Cooper and Liou (2007) pointed out, OTL "refers to the conditions or circumstances within schools and classrooms that promote learning for all students" (p. 44). Therefore, the student's OTL, and the student/school background information associated with it, need to be taken into account when evaluating any student academic outcomes; without accounting for such background information, a full and fair understanding of the variability in the expected outcomes might be difficult to achieve.

Although the definition of the OTL concept may sound trivial, it is not very clear how to measure it for several reasons: (1) common standards of OTL have yet been defined, (2) the operationalization of measuring OTL depends on their definitions, and (3) data gathering on OTL aspects are costly and complex (McDonnell 1995; Porter, 1995b; Guiton & Oakes, 1995; Herman, Klein, & Abedi, 2000; Floden, 2002). O'Day and Smith (1993) have advocated a systemic reform for school improvement by developing school standards for accountability purposes. These school standards should contain three parts: (1) *Resource* – the provision of all of the necessary means (e.g. teachers, school materials, curriculum) to allow all students have the opportunity to learn the curriculum contents to a high performance level, (2) *Practice* – the implementation of school programs or activities that provide such opportunities, and (3) *Performance* – the achievement of high performance goals. The definitions of school standards promote aspects of OTL that need to be included and measured in the attempt of leveling out the playing field before any accountability assessment based on test scores should be imposed on students, teachers, or schools.

However, difficulties in achieving and maintaining a consensus on the acceptable standards have arisen as some of the OTL-related standards depend on the individual characteristics of the students and their interactions with the teacher and peers at school (McDonnell, 1995, Floden, 2002). OTL is not defined simply by the curriculum exposure

and time allocated for it; thus, it is also crucially important how the content was presented, who delivered it, and when such observation/measure was obtained would influence the interpretation of its measurement results (Porter, 1995b; Herman, Klein, & Abedi, 2000; Long, 2014). In addition, the ideological perspectives of the researchers or policymakers could also drive the analysis and decision-making related to how OTL has been and will be provided equally and equitably (Guiton & Oakes, 1995). To avoid such conflicting perspectives, the use of multiple measures of OTL obtained at different time frames is strongly recommended; however, in doing so, the cost for gathering OTL-related information will likely be much greater (Porter, 1995a). To sum up, the strength of the OTL measure(s) and their relationship with student performance depend on how it is defined and operationalized (Floden, 2002; Schmidt & Maier, 2009).

Typically, the main objective in conducting research on OTL is to provide explanatory information about the differences in student performance and identify possible causes and thus (hopefully) solutions for poor schooling outcomes (Porter, 1995a; Floden, 2002; Schmidt & Maier, 2009). However, the success of finding the effect of OTL on student achievement greatly depends on how OTL is operationalized and measured (Muthén et al., 1998; Floden, 2002; Schmidt & Maier, 2009). Chapter 2 has delineated some examples of OTL's operationalization and measures based on past studies and proposed the most appropriate model for measuring OTL with the limitation of the PISA 2012 data. The past models of OTL measures include the OTL-like indices defined by PISA 2012 (OECD, 2014), Kurz (2011), and the Classroom Learning Assessment System at secondary level (CLASS-S) framework (Allen et al., 2013), all of which pertained to the development of the proposed model of the OTL measures. In Chapter 3, student- and school-level background information associated with the proposed OTL measures were investigated in order to provide a deeper understanding on the extent to which this information can be associated with the degree of OTL. Findings will help policy-makers in developing the efforts for improving students' learning experiences.

3.2.1. Measures of opportunity-to-learn in PISA 2012[2]

PISA, which is held every three years, aims to evaluate the readiness of 15-year-old students to meet future challenges. In addition to the administration of cognitive tests to randomly sampled students within each country, PISA also collects information about the classroom learning environment experienced by the students as well as other student and school background information. All of this information is essential for understanding the context and factors pertaining to the students' OTL and their performance in PISA. As PISA results are often used to inform changes in the national policies of the participating countries (Breakspear, 2012; Lockheed, 2015), a detailed investigation on how the diverse student background may influence PISA outcomes in general, as well as the student learning experience, will help the policy makers develop education policy reforms.

[2] Most parts in this section have already been discussed in Chapter 2.

Initially, PISA 2012 defined three aspects of OTL: *content exposure*, *teaching practices*, and *teaching quality*, each of which is categorized further into three sub-aspects (OECD, 2014). These aspects were developed according to the general principles of effective teaching promulgated by Good, Wiley, and Flores (2009) and the factors of successful learning described by Stevens and Grymes (1993). As mathematics was the focus of the PISA 2012 round, most of the items are concentrated on mathematics learning.

The first aspect – ***Content Exposure*** **(CE)** – asks students to rate the degree of (1) their familiarity with certain formal mathematics contents, (2) their experiences with various types of math problems during their schooling, and (3) the frequency of being taught to solve specific math problems involving formal or applied mathematics. Table B.1.1 – B.1.3 in Appendix B lists items corresponding to each of these three CE sub-aspects. For the second aspect, i.e. ***Teaching Practices*** **(TP)**, students report on how often the teacher directs instruction (*teacher directed instruction* sub-aspect), orients student participation (*student orientation* sub-aspect), and provides feedback to students on their assessment and learning (*formative assessment* sub-aspect). This aspect denotes the specific practices/strategies of effective teaching, which were adapted from the OECD's Teaching and Learning International Survey (TALIS) (OECD, 2009b). The complete list of items related to this aspect is provided in Table B.1.4. Lastly, the aspect of ***Teaching Quality*** **(TQ)** embraces three loosely defined characteristics of a supportive learning environment: teacher's socio-emotional support, classroom organization and management, and instructional support/cognitive activation (Pianta & Hamre, 2009; Stuhlman, Hamre, Downer, & Pianta, 2015). Each of these characteristics is represented by the following sub-aspects:

(1) *Teaching Support* (TS), includes questions on how often the teacher shows interest in student learning, gives extra help when needed, helps students with their learning, continues teaching until the students understand, and allows students to express opinions in math lessons (see Table B.1.5).
(2) *Disciplinary Climate* (DC), illustrates several disciplinary-related classroom learning situations that may occur during math lessons such as students not listening to what the teacher says, noise and disorder, the need for the teacher to wait a long time for students to quiet down, the difficulty for students to work well, and a delayed work start-time after the lesson begins (see Table B.1.6).
(3) *Cognitive Activation* (CA), asks students to think about the mathematics teacher that taught their last class and indicate how often a series of teaching strategies, which can cognitively activate students to be able to solve problems in different contexts with multiple solutions, has happened (see Table B.1.7).

After a thorough evaluation of the internal structure of the defined OTL measures and a qualitative investigation on the OTL-related item wordings given by PISA 2012 student questionnaire, these three aspects were thoroughly analyzed to derive five-aspect measures of OTL as shown in Figure 3.1. Taking into account the issue of item sensitivity when operationalizing the OTL aspects (Muthén et al., 1998) and the idiosyncrasy of the learning context (Cogan & Schmidt, 2015), this proposed 5-aspect model was chosen to be the best representation of the OTL measures since it can provide more information to explain the relationship between OTL and the Indonesian student performance in PISA 2012 math. Details on the model building and rationale are given in Chapter 2. However, the following

paragraphs present brief descriptions on the aspects of the proposed OTL measures and how they were derived from the previously defined aspects of OTL.

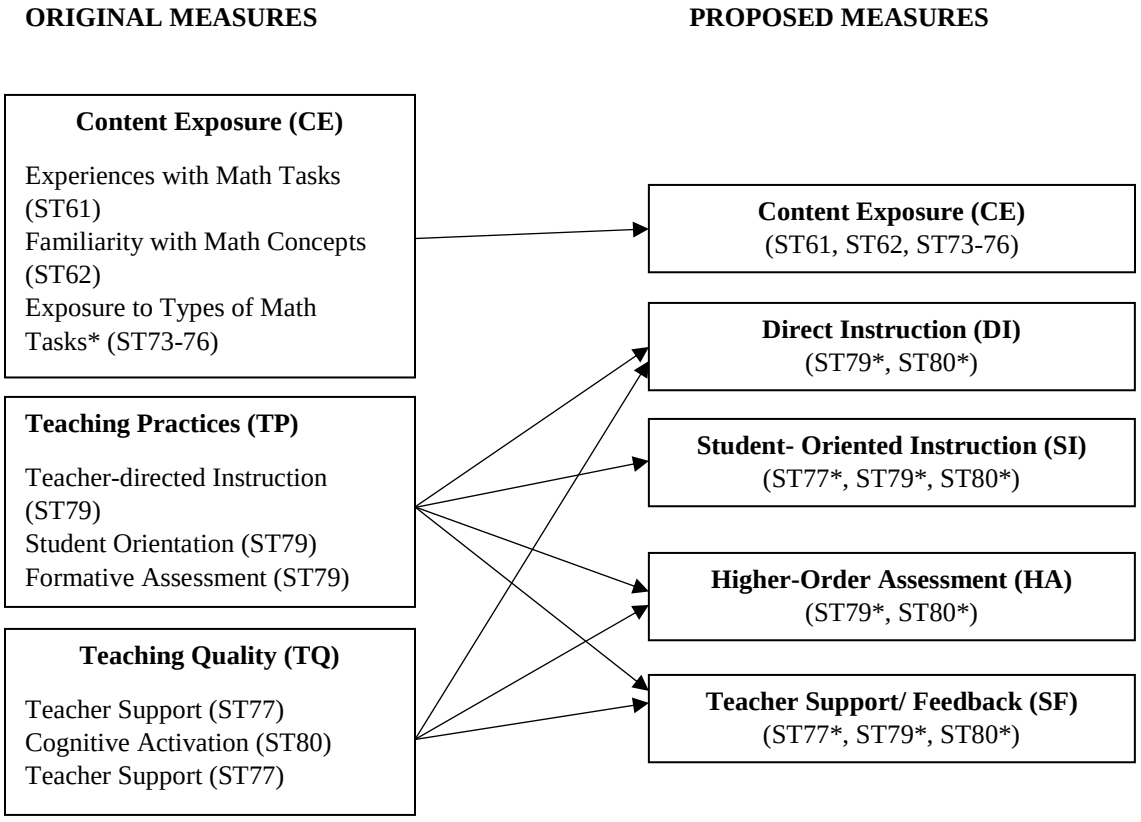


Figure 3.1. Illustration of the original definition of opportunity-to-learn (OTL) aspects in PISA 2012 and their relationships with the proposed measures. Arrows indicate how the related items get distributed, while '*' denotes the inclusion of some parts of the corresponding itemset.

3.2.1.1. Aspect of Content Exposure (CE)

The proposed model embraces the three sub-aspects of CE as a whole because keeping them intact was thought to be more representative in reflecting the students' exposure to math content-related matters and serving the purpose of the study better. The CE sub-aspects reflect factors pertaining to the content coverage and exposure variables of Stevens and Grymes (1993) and the content emphasis of Schmidt (2009). To capture time-on-task or how the students were engaged in learning the curriculum content, time-related scales are utilized: both the familiarity with math concepts sub-aspect (Table B.1.1) and the exposure to types of math tasks sub-aspect (Table B.1.3) used a 4-point rating scale with "frequently–sometimes–rarely– never" options. Meanwhile, the sub-aspect of experiences with math tasks (Table B.1.2) used a 5-point rating scale with "never heard of it – heard of it

once or twice – heard of it a few times – heard it often – know it well, understand the concept" response categories. The item set ST76 Example 1 and 2 listed in Table B.1.3 include two sample PISA math items, which were items from Unit 9 Robberies and Unit 46 Heartbeat, respectively (OECD 2009a; Cogan & Schmidt, 2015). For this CE aspect, hypothetically, the more frequent a student responded to being exposed to a particular topic/task in math, the higher the degree of content exposure and the more OTL the particular student is assumed to have on mathematics.

3.2.1.2. Aspect of Direct Instruction (DI)

The aspect of the Direct Instruction approach encompasses direct teaching practices that state learning goals, inform success criteria, build commitment and engagement, provide guided lessons and practices, provide summary/conclusions, and allow independent practice after direct instruction (Hattie, 2009; Kurz & Elliott, 2014; Stuhlman et al., 2015). Items reflecting this DI aspect were teased out from the TP aspect (Table B.1.4) and CA aspect (Table B.1.7), and recompiled in Table 3.1. This classification was similar to the grouping of the *teacher directed instruction*-related items as originally defined in the TP aspect, but included two extra items: how the teacher sets the expected outcomes from a test, quiz, or assignment (ST79Q12) and how the teacher asks for an explanation of a problem solving approach (ST80Q09). These two extra items seemed to represent the sub-aspect of *teacher directed instruction* more than the *formative assessment* sub-aspect of the TP's original classification. It is hypothesized that having more frequent instructional strategies in which the teacher holds the main authority and becomes dominant would represent a higher degree of the direct instruction aspect of teaching. How this aspect would influence a student's OTL may depend on the cultural context of a specific teaching and learning environment. In Indonesia, for example, the teacher-centered practices were more predominant than the student-centered approaches (Worldbank, 2010).

3.2.1.3. Aspect of Student-oriented Instruction (SI)

The Student-oriented Instruction aspect depicts how teachers tailor their instruction or assessment per the student needs such as grouping and allowing students' self-selected approach to problem-solving (Hattie, 2009; Allen, et al., 2013). Table 3.2 presents the related items, which were teased out from the previous TS, TP, and CA aspects. This aspect contains three items out of five items that operationalized the *student orientation* sub-aspect in the previous TP aspect. The two additional items refer to how the teacher would orient their teaching to suit the students' needs (see ST77Q05) and allow some flexibility and autonomy for students in problem-solving (see ST80Q05). Hypothetically, when teachers weigh in student factors for their instructional strategies more frequently, the degree of student-oriented instruction increases. The direction of this student-oriented instruction's effect (SI) on student achievement may contradict the effect of the direct instruction (DI) approach. Depending on the context, when the DI aspect positively influences the academic achievement, such environment may have a negative effect of the student-oriented approach.

Table 3.1
Items related to the Direct Instruction (DI) aspect

Item No	Item code	Item Wording
36	ST79Q01[1]	a) The teacher sets clear goals for our learning
37	ST79Q02[1]	b) The teacher asks me or my classmates to present our thinking or reasoning at some length
41	ST79Q06[1]	f) The teacher asks questions to check whether we have understood what was taught
43	ST79Q08[1]	h) At the beginning of a lesson, the teacher presents a short summary of the previous lesson
46	ST79Q12[1]	k) The teacher tells us what is expected of us when we get a test, quiz or assignment
47	ST79Q15[1]	l) The teacher tells us what we have to learn
55	ST80Q09[2]	g) The teacher asks us to explain how we have solved a problem

Note. This table is identical with Table 2.10.
[1] Item prompt: How often do these things happen in your mathematics lessons? Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen? Scale category: Always or almost always-Often-Sometimes-Never or rarely.

Table 3.2
Items related to the Student-oriented Instruction (SI) aspect

Item No	Item code	Item Wording
34	ST77Q05[1]	d) The teacher continues teaching until the students understand
38	ST79Q03[1]	c) The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster
42	ST79Q07[1]	g) The teacher has us work in small groups to come up with joint solutions to a problem or task
44	ST79Q10[1]	i) The teacher asks us to help plan classroom activities or topics
51	ST80Q05[3]	c) The teacher asks us to decide on our own procedures for solving complex problems

Note. This table is identical with Table 2.11.
[1] Item prompt: How often do these things happen in your mathematics lessons? Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen? Scale category: Always or almost always-Often-Sometimes-Never or rarely.

3.2.1.4. Aspect of High-order Assessment (HA)

Based on the cognitive activation concept described by Kunter and Baumert (2006) and Baumert et al. (2010), this aspect represents the provision of intellectually challenging tasks to students that requires higher-order thinking or cognitive processes. It is hypothesized that to have a higher degree of higher-order assessment/task, the students math teachers assign them projects that require more than a week to complete, problems with no immediately obvious solutions or many different approaches for solutions, tasks that require students to think for an extended time, and so on (see Table 3.3). Most of the items were derived from the *cognitive activation* (CA) sub-aspect, except one item that asked if the teacher assigned projects with a due date of at least one week (ST79Q04) – derived from the previous TP aspect (i.e. from the *student-orientation* sub-aspect). Having taken a week to complete, a project could be assumed to have a considerable degree of complexity. From past studies, being exposed to progressive instructional strategies that promote high-order cognitive skills/thinking can improve students' OTL, which in turn is positively associated with higher academic performance (Lee, 2006; Kunter & Baumert, 2006; Allen et al., 2013).

Table 3.3
Items related to the Higher-order Assessment (HA) aspect

Item No	Item code	Item Wording
39	ST79Q04[1]	d) The teacher assigns projects that require at least one week to complete
49	ST80Q01[2]	a) The teacher asks questions that make us reflect on the problem
50	ST80Q04[2]	b) The teacher gives problems that require us to think for an extended time
52	ST80Q06[2]	d) The teacher presents problems for which there is no immediately obvious method of solution
53	ST80Q07[2]	e) The teacher presents problems in different contexts so that students know whether they have understood the concepts
56	ST80Q10[2]	h) The teacher presents problems that require students to apply what they have learned to new contexts
57	ST80Q11[2]	i) The teacher gives problems that can be solved in several different ways

Note. This table is identical with Table 2.12.
[1] Item prompt: How often do these things happen in your mathematics lessons? Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen? Scale category: Always or almost always-Often-Sometimes-Never or rarely.

3.2.1.5. Aspect of Teacher Support/Feedback (SF)

Teachers' emotional support and sensitivity (Stuhlman et al., 2015), sense of immediacy (Hattie, 2009, p. 183), and feedback system (Hattie, 2009, p. 173) are captured in this Teacher Support/Feedback aspect. Table 3.4 delineates the corresponding items, which were teased out from the TS, TP and CA aspects. Similarly, this aspect contains four items related to the original definition of the TS aspect (ST77Q01-06). The additional three items were obtained from the *formative assessment* sub-aspect of the TP aspect (ST79Q05-17) because they reflected more on how the teacher provides feedback for improvement. Lastly, one item from the prior CA aspect was included in this aspect because it asked how the teacher helped students to learn from their own mistakes (ST80Q08). When the math teacher shows interest in every students' learning, gives help with student learning and even extra help when needed, while also giving feedback about students' strengths and weaknesses as well as guidance and praise, hypothetically such student would have a high opportunity to learn in this aspect.

Table 3.4
Items related to the quality of teacher's support/feedback (SF) aspect

Item No	Item code	Item Wording
31	ST77Q01[1]	a) The teacher shows an interest in every student's learning
32	ST77Q02[1]	b) The teacher gives extra help when students need it
33	ST77Q04[1]	c) The teacher helps students with their learning
35	ST77Q06[1]	e) The teacher gives students an opportunity to express opinions
40	ST79Q05[1]	e) The teacher tells me about how well I am doing in my mathematics class
45	ST79Q11[1]	j) The teacher gives me feedback on my strengths and weaknesses in mathematics
48	ST79Q17[1]	m) The teacher tells me what I need to do to become better in mathematics
54	ST80Q08[2]	f) The teacher helps us to learn from mistakes we have made

Note. This table is identical with Table 2.13.
[1] Item prompt: How often do these things happen in your mathematics lessons? Scale category: Every lesson-Most lessons-Some lessons-Never or hardly ever.
[2] Item prompt: Thinking about the mathematics teacher that taught your last mathematics class: How often does each of the following happen? Scale category: Always or almost always-Often-Sometimes-Never or rarely.

The five-aspect model as described above proposes an alternative definition for the OTL measures in PISA 2012 as illustrated in Figure 3.2. This proposed model serves the idiosyncratic feature of any OTL measure: context-based (Cogan & Schmidt, 2015), particularly when assessing the Indonesian students' learning experience and environment. Furthermore, the subsequent data analysis incorporated some selected student- and school-

level background information as collected by PISA 2012 to investigate their effects, or can be referred to as associations in this study, across the OTL aspects.

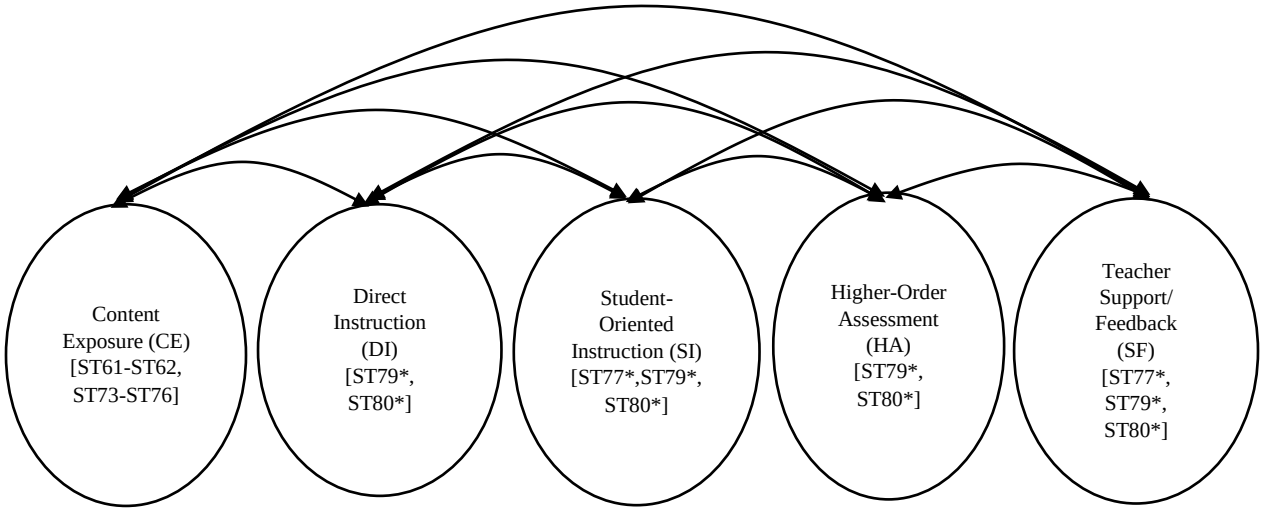


Figure 3.2. Illustration of the proposed measures of the inter-related aspects of opportunity-to-learn (OTL) using PISA 2012 survey items.

3.2.2. Background information related to the opportunity-to-learn measures

The degree to which an opportunity-to-learn (OTL) is experienced by a student may vary depending on many background variables including the student's economic, socio, and cultural condition (home and community), curriculum standards, teaching practices and quality, school programs, and so on (Gamoran, 1987; Porter, 1995b; Schwatz, 1995; Floden, 2000; Schmidt & Maier, 2009, Schmidt et al., 2015). In an effective school, the prescribed/intended curriculum standards are translated into appropriate levels of detail that are taught to students while utilizing adequate infrastructure, human and financial resources. Apart from being contingent on the school resources, the implementation of the curriculum also depends on the teacher's capacity to interpret, teach, and assess the prescribed standards. Despite learning in the same classroom with the same teacher, each individual student may also benefit from the classroom activities to a varying degree. Students' unequal learning experiences may also be associated with school characteristics such as intra- or extra-curricular programs, curricular differentiation, class grouping/tracking, types of assessments, and school type/location (O'Day & Smith, 1993; Guiton & Oakes, 1995; Herman & Klein, 1996; Schwatz, 1995; Cooper & Liou, 2007). Therefore, it is important to investigate the extent to which student and school background information can be associated with a student's OTL; recognizing these background information may inform effective interventions in leveraging the student learning experience. Some of the background variables included in the analysis of this chapter are discussed as follows.

Even though the measurement of socio-economic status (SES) is still imperfect, student SES is often linked to many aspects of educational outcome, either directly or indirectly (Gamoran, 1987; Herman & Klein, 1996; Considine & Zappala, 2002; Perry & McConney, 2010). Students with low SES tend to live in more disadvantaged areas, attend schools with limited educational resources, and get tracked to lower quality school programs, all of which in turn lead to an adverse association on achievement. Gamoran (1987) suggested that high school students with high SES had higher achievement because they had a relatively advantaged schooling experience. The more economically-advantaged students were more likely to attend schools with more academic tracks and high quality coursework programs. When these positive experiences are held constant, SES has little effect on student achievement (Gamoran, 1987). In addition, Perry and McConney (2010) argued that the school mean SES had a strong effect on the student achievement, regardless of the individual student SES. For example, in Australia, the composition of the SES within school mattered greatly for student performance. Furthermore, Schmidt and his colleagues (2013) also suggested that in most countries, the SES and OTL variation was greater within schools than between them when analyzing PISA 2012 data. However, in their 2015 paper, Schmidt et al. identified different patterns of the source of OTL inequalities when exploring the joint relationship of students' exposure to formal mathematics as a measure of OTL and their SES to the students' mathematics performance in PISA 2012 test. Some countries had greater between-school OTL inequalities (e.g. Germany and the Netherlands), while some other countries had a larger gap of OTL within schools (e.g. the U.S, Australia, and the U.K). Nevertheless, the findings of their study confirmed a strong direct relationship between the OTL measure and student performance, and revealed that students with high SES were more likely to receive more rigorous OTL in mathematics (Schmidt et al., 2015). The link between SES and the mathematics performance was also found to be largely due to the association of SES with OTL. Therefore, the inclusion of SES as a control variable in the investigation of students' OTL is essential to drive the attention of public policy makers to improve the schooling process in economically-disadvantaged areas.

Furthermore, the effect of school location on students' academic achievement and their learning experience is highly contextual, depending on the socio-economic and cultural status of the community in which such school serves. In a study that evaluated the effect of school types (i.e. urban, rural, and suburban) on the 8th grade students' opportunity to learn with respect to the new performance-based assessment[3] in mathematics, Herman and Klein (1996) found no consistent differences on the students' OTL across school types. However, the students attending suburban schools had more preparation for the assessment than their cohorts living in other areas. These students mostly came from affluent families, who were more likely to attend schools that provided more educational resources. They also found that students attending urban schools, which were mainly located in inner cities with economically-disadvantaged communities, were less likely to have recent mathematics textbooks, indicating that they might not be sufficiently prepared for the assessment. Meanwhile, the rural school students, who lived in more geographically remote areas with mixed SES, were observed to be more engaged with constructivist learning experience related to the recent reforms in curriculum and assessment system.

[3] The 1993 California Learning Assessment System (CLAS).

In addition, Reeves (2012) contended that attending rural high schools in the U.S. did not seem to give less advantage on the students' OTL. The students, typically with low family SES, still experienced high quality instructional practices in mathematics, which compared favorably with their cohorts in non-rural high schools. Instead, the achievement gap between rural and non-rural students in the U.S appeared to strongly stem from low influence and motivational support from family and friends.

On the contrary, rural schools in many developing countries were associated with low income and economically-disadvantage communities; hence they tended to perform less well compared to their cohorts in bigger cities (MacJessie-Mbewe, 2004; Epstein & Yuthas, 2012; Luschei & Zubaidah, 2012). These schools, which are normally located in remote and/or geographically challenged areas, were less likely to have sufficient educational resources and attract high quality teachers. In Indonesia's rural areas, teachers' absenteeism was high, whereas multi-grade teaching often occurred due to lack of teachers (Luschei & Zubaidah, 2012; MoECI, 2013b.) Hence, the students attending such impoverished schools were more likely to have less adequate OTL than students at schools in bigger cities with more educational resources, which eventually costed them lower academic achievement and attainment, and thus lost of rewarding job opportunities. Since the dropout rates were also high, some studies recommended tailoring the curriculum and instructional strategies to the needs and characteristics of the respective rural community (MacJessie-Mbewe, 2004; Epstein & Yuthas, 2012).

Apart from the content exposure aspect, about half of the OTL related items correspond to the aspects of instructional strategies, assessment, and support for students. Past studies have indicated that teachers matter, as gaps in teaching capacity are often associated with gaps in student academic performance (Darling-Hammond, 2000; Betts, Zau, & Rice, 2003; Goe & Stickler, 2008). Among many variables that assess teacher quality, teacher certification is one way that can provide a benchmark for pre-service teachers and quality assurance for in-service teachers. In a typical teacher certification program, the pre-certified teachers are required to demonstrate their ability in subject matter knowledge and their knowledge in teaching, learning, and assessment, as well as their capacity in providing support for students and their different needs. In a performance-based assessment for teacher certification, teachers are not only assessed on their content knowledge, but are also required to demonstrate such teaching skills as planning, instruction, assessment, reflection, and development of academic language (Darling-Hammond, 2011). Although further studies are warranted to define what type of certification program would be the most productive (Goe & Stickler, 2008), having qualified teachers leverages students' OTL since these teachers can develop an effective learning environment (Betts, Zau, & Rice, 2003).

The final covariate that may be associated with the students' OTL in school settings as discussed in this current study is the notion of grade retention as to whether retaining students would actually provide these students more OTL to achieve the expected knowledge or skills. However, unless it is accompanied by an increase in instructional support, past studies indicate that student retention was associated with adverse effects on student learning (NASP, n.d.; McCoy & Reynolds, 1999; Silberglitt, Appelton, Burns, & Jimerson, 2006; Allen et al., 2009). Comparing data from PISA 2003 and PISA 2012 (OECD, 2013e)

revealed that in some countries, the percentage of students, who had ever repeated a grade in primary, lower secondary or upper secondary school, increased in a decade, e.g. Finland (up 1%), the U.S. (up 1.7%), and Korea (up 3.2%). On the other hand, countries like France (down 11.1%) or Indonesia (down 0.4%) had decreased the percentage of students retained, whilst Japan and Norway had zero retention. Although the majority of retained students were reported to come from the minority groups and/or from families with low SES background (NASP, n.d.; Jimerson, 2001), it is beyond the scope of the current study to seek an association between student retention and increase in poverty. Nevertheless, an understanding on how the retained students perceived their OTL can provide evidence of these students' learning experience, which could eventually benefit the school policy makers as to ensure that all students from diverse background have access to an equitable OTL.

3.3. Research Objectives

Due to the nature of sampling, it is also interesting to see how schools vary in the background information (covariates) associated with the students' OTL. I would argue that the variability of the school clusters would relate to the variability of OTL as perceived by the sampled students. Specifically in developing countries, the process of schooling reflected in the concept of OTL varies among schools following the structure and distribution of the schools. Therefore, using the PISA 2012 data on Indonesian students, I investigate how background information shows associations across the different aspects of the OTL measures, after accounting for the school differences. To do so, I have the following research questions:

1. To what extent are the student- and school-level covariates correlated with the aspects of the OTL measures? How can these covariates contribute to inform OTL improvement?
2. To what extent do the covariates of gender, SES, and grade levels differ across the aspects of the proposed OTL measures after accounting for school clusters?

3.4. Brief Description of Education System in Indonesia[3]

With a population of around 245 million people (Trading Economics, 2012), Indonesia is the fourth largest country in terms of population in the world, covering about 17,000 islands with more than 200 ethnic groups and at least 300 different languages/dialects. With a diverse pattern of population demographics, it is not surprising that there is a large socio-economic gap and income inequality throughout the country, which in turn is closely related to student academic performance differences (CIA, 2013; OECD, 2013a). The stagnantly low performance (i.e. in the ten bottom ranks) of the Indonesian students in PISA and other international student assessment programs on all tested subjects is

of significant concern and calls for a major national education reform (see Table 1.2 for Indonesian students' performance on PISA mathematics). Please note that the mean scores of the student math performance in Table 1.2 cannot be directly compared without linking the common items from the 2000 round to the 2012 round of the PISA test[4]. However, the students' low performance in the five consecutive PISA rounds has strongly informed for the national education rhetoric (Kompas, 2013d).

The school system in Indonesia is comprised of 3 levels: primary/elementary school (1^{st} – 6^{th} grades), junior secondary school/middle school (7^{th} – 9^{th} grades), and senior secondary school/high school (10^{th} – 12^{th} grades) (MoECI, 2013b). Previously, compulsory and public education was only offered for the first nine years of schooling (i.e., primary to junior secondary levels). But recently, the MoECI introduced the formal plan for a twelve–year compulsory education program that will include free public education for the three years of senior secondary level (MoECI, 2015).

Progression between grade levels is very restricted in Indonesia since each student must pass a final year assessment developed by local schools in order to move up the ladder. Upon completion of each of the three levels of schooling, the students must take a high-stakes NE over several nationally-defined major subjects such as mathematics, language (Indonesian and English), and science at the end of 6^{th}, 9^{th}, and 12^{th} grade, respectively. Results from these NEs become the major factor for not only passing the current school level, but also for school admission at the next level. Although currently the stakes of the NE have been greatly reduced, the results are still used for admission to the best schools[5] at any level (Sindo, 2014; MoECI, 2016). Meanwhile, the passing rates on these NEs, are almost always 100% (Kompas, 2012; Metro, 2012; Kompas, 2013c; MoECI, 2014). There was only 0.04% of junior secondary students who failed in the 2012 NE in mathematics (Metro, 2012). The striking difference between the NE results and the PISA outcomes has prompted the Indonesian government to review its national curriculum and introduce new standards (MoECI, 2012). The development and administration of the new, yet controversial, curriculum faced significant challenges, disputes, and oppositions from the wider community (Kompas, 2013a; Kompas 2013b). Quoting the poor performance on international assessments, the Indonesian Minister of Education and Culture re-evaluated the new curriculum standards and practices, and targeted a 20% performance increase in the next round of the PISA tests (Baswedan, 2014). Having only been implemented for six months in almost all public schools and some private schools throughout Indonesia, the MoECI halted and revisited the implementation of the new curriculum in December 2014.

Based on the national school management policy, there are three categories of school programs: (1) general school – under the jurisdiction of the Ministry of Education and Culture (MoEC), (2) Islamic school – under the jurisdiction of the Ministry of Religion Affairs (MoRA), and (3) vocational school – under the jurisdiction of the MoEC. The first two categories provide education for K-12 programs, while the vocational school only caters for the equivalent of 7^{th} to 12^{th} grade programs with an emphasis on specific vocational

[4] Although the latest round of PISA was held in 2015, the data has yet been made available publicly.
[5] In this case, the best schools refer to schools to which students with high national examination results competed to get admitted.

education programs for training students with specific skills in a variety of fields such as mechanics, business, home economics, tourism, handicraft, art, agriculture and aquaculture.

In adherence to the Indonesian President's top nine development agenda (*Nawacita*), this year the MoEC highlighted some significant efforts in improving the quality of the vocational education programs (VEPs) and their graduates (KSP, 2016; Media Indonesia, 2016). MoECI (2013b) reported a 158% increase in vocational school enrollment during 20001 – 2010, specifically for VEP at senior secondary level, due to the government-driven strategies to utilize the program outputs for fostering rapid economic development. As the growth was seen to be more supply-driven rather than demand-driven (MoECI, 2013b), MoEC needs to ensure that these programs can produce high quality and job-ready graduates as needed by the job markets. Some efforts for the VEP's improvement include revitalizing the VEP curriculum, providing support for quality academic resources (teaching media, professional development training for teachers, laboratories), and opening new VEPs with a specific priority for tourism, agriculture, and maritime/aquaculture related programs.

In addition, for each of the school categories – depending on the financial sources, there are two major distinctive types of school: public and private. Public schools are managed and controlled by "a public education authority, government agency, or governing board appointed by government or elected by public franchise"; whereas private schools are managed by private institutions such as religious and non-religious foundations (OECD, 2013c). Per the MoEC's 2010/2011 data, about 78% of schools at junior secondary level are public, whilst the public schools' share was only 32% at the senior secondary level (MoECI, 2013b).The rest were private schools. To help serve the poor community, MoRA builds and manages many Islamic schools at all levels in both rural and urban areas, whose funds can be provided either fully by the government or partly by particular community organizations.

Globally, private school students typically outperformed those from public schools (OECD, 2013d), but for Indonesian students, this might not be true. As MoECI (2013b) indicated, students who failed in getting admission to the public schools are more likely to enter the private schools. A similar case also happens with the vocational programs. Only those who need to seek employment and earn a living after graduation would opt for vocational schools[6]. Hence, public schools tend to have the advantage of getting good students and thus, are more likely to pay closer attention to their teaching, learning, and assessment process due to strict and rigorous monitoring from the local district offices. Using national examination (NE) results for primary and secondary school levels and three rounds of the Family Life Survey during 1997 – 2000, Newhouse and Beegle (2006) found that the Indonesian students attending public schools and non-Islamic private schools had generally performed better than their cohorts at the private schools, after controlling for several background variables. They asserted that the high performance of public schoolers at junior high school level was most likely due to the upward selection bias of higher quality admits. Thus this finding is associated with the grade repetition/promotion practices discussed in the

[6] However, the enrollment in the vocational schools has grown fast since 2004 due to a special endorsement from the MoECI, specifically in three major provinces in Indonesia (DKI Jakarta, central Java, and DI Yogyakarta) in which more students enrolled in the vocational schools than those did in general senior secondary education (MoECI, 2013b).

previous section. Their study also found that students at city schools performed worse than those located in non-urban areas.

Teacher quality is an on-going issue in Indonesia. MoCEI (2013b) reported that the there was a surplus in teacher workforce in comparison to student enrollment. In 2009, the student-teacher-ratio (STR) at the primary education level fell down to 16:1 from 22:1 nine years before. A similar scenario was also observed at the junior secondary level. This low STR did not correspond with smaller class sizes, but instead, an oversupply of teachers occurred in popular districts in economically-advantaged areas. Given the vast geographical challenge, the archipelagic Indonesia consists of many rural and remote regions that are far removed from cities. As anticipated, a shortage of teachers was found in remote regions with limited school and educational resources (Luschei & Zubaidah, 2012; MoECI, 2013b). Often, these teachers had to practice multigrade teaching. This phenomenon can potentially reduce students' OTL, because students might not receive sufficient quality and quantity of learning and instruction (Luschei & Zubaidah, 2012).

According to the latest Teachers Law No. 14 in 2005, the Indonesian teachers, at either public or private schools, are required to have attained at least a Bachelor or Diploma 4[7] degree and then successfully complete the teacher certification process (TCP). This teacher certification program aims at ensuring a high quality benchmark for pre-service teachers as well as improving in-service teachers' capacity to "educate, teach, train, guide, and assess students' learning" (Jalal et al., 2009, p. 29), which can also give public assurance that teachers do have the necessary knowledge and skills in teaching. Performed by several certifying universities, the certification program includes a written test on basic skills in writing, reading and mathematics and a portfolio assessment. Teachers who pass the test are labeled as *certified* and entitled to have their base salary doubled. In 2007, 52% of teachers who took the inaugural portfolio test failed. These teachers should then attend a remedial course provided by the certifying university; 96% of participating teachers passed. Based on the active learning curriculum, the comprehensive TCP program runs for 90 hours in which teachers spend 30 hours of theory class and 60 hours of actual teaching in their classrooms using the taught instructional strategies (MoECI, 2013b). Hence, the certified teachers are expected to have gained the necessary content knowledge and the pedagogical content knowledge for teaching.

3.5. Data Sample

3.5.1. Description of opportunity-to-learn related items in PISA 2012[3]

This chapter uses item responses to the 57 OTL-related items as obtained from the student background questionnaire (see Table B.1.1 – B1.3 and Table 3.1 – 3.4). The dataset

[7] Diploma 4 is a professional degree conferred by a teacher training program in which about 60% of its curriculum requires the student to practical work outside the school.

.

consists of all 5622 participating students from 209 schools, each of which had 2 – 35 sampled students. The item responses are polytomous (scored as 0 to 3 or 4 depending on the rating scale used by each item), in which a higher rating indicates a higher frequency of occurrence for each event/statement. Hence, most of the items' scales were reversed from their original rating in the calibration process. Due to the nature of two-stage stratified sampling in PISA, student survey weights were incorporated in the calibration of the item parameters of the proposed OTL measures, as discussed in Chapter 2. The student survey weight (provided by PISA) is calculated from the school base weight, the within-school base weight, some adjustment factors to compensate for non-participation by school/students, the school base weight trimming factor, and the final student weight trimming factor. Further details about each of these weight components are described in the PISA 2012 Technical Report (OECD, 2014). However, the later regression analyses presented in this chapter did not use weights[8].

In PISA 2012, the student questionnaire also followed a rotational design in which each participating student did not receive the same amount of items as others (see Table 3.5). Only about one-third of the sampled students took the complete set of items. Similar to the matrix sampling of the cognitive tests, this approach was taken to increase the content coverage of surveyed topics without increasing the response time for the respondent, since the administration time was only 30 minutes. The three forms of the questionnaire were distributed randomly to the students. Each of these forms has a common part and a rotated part. The common part included items about student demographics, whereas the rotated parts dealt with the attitudinal, perception, and other non-cognitive constructs. The OTL related items were distributed as an intact set across the three different forms in such a way that any missing values are due to non-response only.

Table 3.5
The distribution of opportunity-to-learn related items (coded as ST61 – ST80) based on the rotational design of the PISA 2012 Student Questionnaire

Form A	**Form B**	**Form C**
ST61	ST77	ST61
ST62	ST79	ST62
ST73	ST80	ST73
ST74		ST74
ST75		ST75
ST76		ST76
		ST77
		ST79
		ST80

[8] I found that the estimated item parameter values, calibrated with or without weights, were almost similar. Hence, the use of weights in the subsequent analyses is negligible.

3.5.2. Description of the selected background variables[3]

The students' item level scores on the OTL related items were obtained along with information about their gender, grade level, socio-economic status (SES), school type, program, and location, as well as information about the availability of additional math lessons offered by the school and the proportion of certified teachers in the school from publicly available datasets. The list of background variables then can serve as covariates related to a student's OTL is given in Table 3.6.

Table 3.6
List of student and school background variables used in the data analysis

Name	Definition	Description
Content Exposure	Student's perceptive on their exposure to specific mathematics contents, tasks, and problems	30 items: 4-response and 5-response categories
Direct Instruction	Student's perceptive on the teacher's directed instruction	7 items: 4-response categories
Student-oriented Instruction	Student's perceptive on the teaching instruction based on student needs	5 items: 4-response categories
Higher-order Assessment	Student's perceptive on cognitively challenged assessment/task	7 items: 4-response categories
Teacher Support/Feedback	Student's perceptive on the emotional support and feedback from teacher	8 items: 4-response categories
Male	Gender	Dummy variable: 1 = yes, 0 = no
SES	Index of socio-economics status	A composite index (from both parents' educational levels and occupational status, number of certain home possessions)
Grade 7-8	15-year-old student who was in 7th and 8th grade	Effect-coding variable: 1 = yes, 0 otherwise, -1 = modal group
Grade 9	15-year-old student who was in 9th grade	Effect-coding variable: 1 = yes, 0 otherwise, -1 = modal group
Grade 11	15-year-old student who was in 11th grade	Effect-coding variable: 1 = yes, 0 otherwise, -1 = modal group
Grade 12	15-year-old student who was in 12th grade	Effect-coding variable: 1 = yes, 0 otherwise, -1 = modal group
Vocational Secondary School	Attendance in a vocational school	Dummy variable: 1 = yes, 0 = no
Private school	Attendance in a privately subsided school	Dummy variable: 1 = yes, 0 = No

Rural schools	Attendance in a school located in a community less than 3000 inhabitants	Dummy variable: 1 = yes, 0 = No
Extra math lesson in school	Attendance in a school that offers additional math lessons	Dummy variable: 1 = yes, 0 = No
Proportion of certified teachers	Proportion of fully certified teacher in the school	The ratio of number of fully certified teacher to the total numbers of teachers

Note. The effect coding assigns 10th grade students as the reference group.

The distributions of the sampled students across gender and grade levels are presented in Table 3.7. The majority of students were in 9th grade (40%) and 10th grade (46%). Furthermore, about ten percent of the students were in 7th and 8th grade, whereas a little more than four percent were above 10th grade. Interestingly, nineteen students were already in 12th grade. Such students might have either started school at an early age or skipped class due to their special talent. As the 10th grade is the grade level at which the number of participating students is the largest, it is considered the modal grade. At the international level, 10th grade is also the modal grade. Among these students, 51% of them are female.

Table 3.7
The distribution of Indonesian students across gender and grade levels in PISA 2012. Grade 10 is the modal grade (bolded) for both Indonesia and International data

Gender	**Grade Level**						**Total**
	7	**8**	**9**	**10**	**11**	**12**	
Female	39	169	1,104	1,411	128	9	2860 (50.9%)
Male	61	268	1,150	1,186	87	10	2762 (49.1%)
Total (count)	100	437	2254	**2597**	215	19	5622
(%)	1.8	7.8	40.1	**46.2**	3.8	.3	100

As presented in Table 3.8, the school location variable indicates the community context in which the school is located. It has five categories including: village (serving less than 3000 inhabitants), small town (3000-15,000 inhabitants), town (15,000 – 100,000 inhabitants), city (100,000 – 1 million inhabitants), and large city (> 1 million inhabitants). The distribution of sampled students across the different school categories is also delineated in Table 3.8. In PISA 2012, a quarter of the sampled students (25%) attended schools located in a rural community of less than 3000 inhabitants. On average, these students came from low-income families, as compared to those who attended schools in bigger communities (see Figure 3.2). There was also a higher percentage of sampled students attending public schools than private ones with more students coming from the general junior secondary programs (about 40%). Almost half of the sampled private schools were Islamic religious schools and vocational schools located in non-urban areas. As previously discussed, these schools were more likely to be serving students from poor communities (OECD-ADB, 2015). In addition,

Table 3.8 shows that almost 20% of the sampled students attended vocational secondary schools, in which only 37% of the schools is public. Many of these public vocational schools are located in small towns, signifying that these schools serve more economically-disadvantaged students than their cohorts in more populated areas.

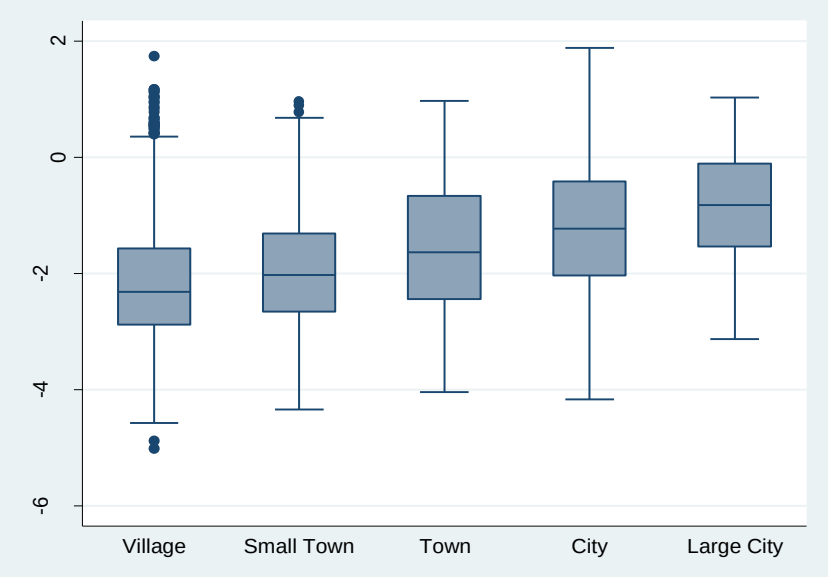
Table 3.8
The percentage of sampled student distribution across secondary school types, programs, and locations in PISA 2012, calculated with respect to the whole country sample

School location	General Junior (SMP)[1]		Islamic Junior (MTs)[2]		General Senior (SMA)[3]		Islamic Senior (MA)[4]		Vocational Secondary (SMK)[5]		Total Public	Total Private	Total
	Pu[6]	Pv[7]	Pu[6]	Pv[7]	Pu[6]	Pv[7]	Pu[6]	Pv[7]	Pu[6]	Pv[7]			
Village	8.7	3.6	-	6.6	2.2	0.6	.1	1.5	1.2	0.8	12.1	13.2	25.3
Small Town	15.0	3.5	.9	1.3	9.9	1.2	-	1.7	4.0	2.8	29.8	10.5	40.4
Town	3.3	0.1	-	1.2	3.6	0.9	1.1	.4	.5	3.3	8.5	6.0	14.5
City	2.7	1.8	-	-	2.9	2.3	.6	-	1.2	5.0	7.4	9.1	16.5
Large City	.5	.3	-	-	1.1	1.2	-	-	-	.2	1.6	1.7	3.3
Total	30.2	9.4	.9	9.2	19.7	6.2	1.8	3.7	6.9	12.0	59.5	40.5	100.0

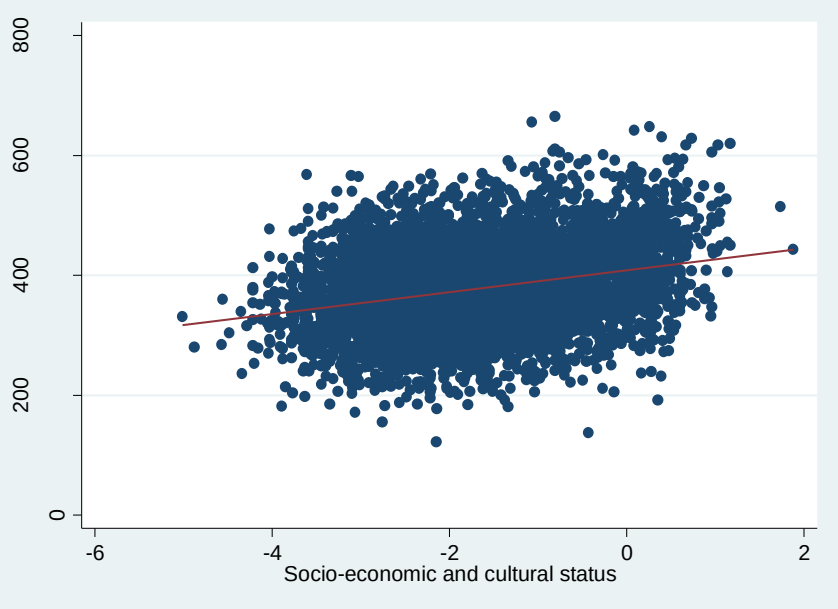
Notes. [1] SMP (Sekolah Menengah Pertama) = general junior secondary school, [2] MTs (Madrasah Tsanawiyah) = Islamic junior secondary school, [3] SMA (Sekolah Menengah Atas) = general senior secondary school, [4] MA (Madrasah Aliyah) = Islamic senior secondary school, [5] SMK (Sekolah Menengah Ketrampilan) = vocational secondary school, [6] Pu stands for public schools, and [7] 'Pv' means private schools.

The SES indicator is a composite scale for the economic and socio-cultural status of an individual student. It is derived from the indices of highest occupational status of parents, highest educational level of parents (in years of education), and home possessions using a principal component analysis (OECD, 2014). Higher values indicate relatively higher levels of SES and are standardized to a mean of zero for the population of students in OECD countries. A one-unit difference on the scale represents a difference of 1 standard deviation on the scale distribution. Most of the Indonesian sampled students have an SES index below the OECD average with a mean of -1.77 and a standard deviation of 1.09. The minimum and maximum values of SES are -5.01 and 1.88, respectively. With large economic gaps, students with moderately high SES tend to attend schools located in the cities. In Figure 3.3(a), the mean students' SES attending schools in the less populated areas are lower than those who lived in more densely-populated cities. However, there are some anomalies as shown in Figure 3.3(a), i.e. village/rural students with high SES. This phenomenon might correspond to students who live in small villages near natural resource plantations or mining fields in remote areas throughout Indonesia and attend a special private school. As anticipated, there was an indication that the student performance in PISA 2012 math was positively associated with their SES levels (Figure 3.3(b)), which also varies according to their school type and locations (Figure 3.3(c)).

(a)


2
0
-2
-4
-6
Village
Small Town
Town
City
Large City

(b)


800
600
400
200
0
-6
-4
-2
0
2
Socio-economic and cultural status

(c)

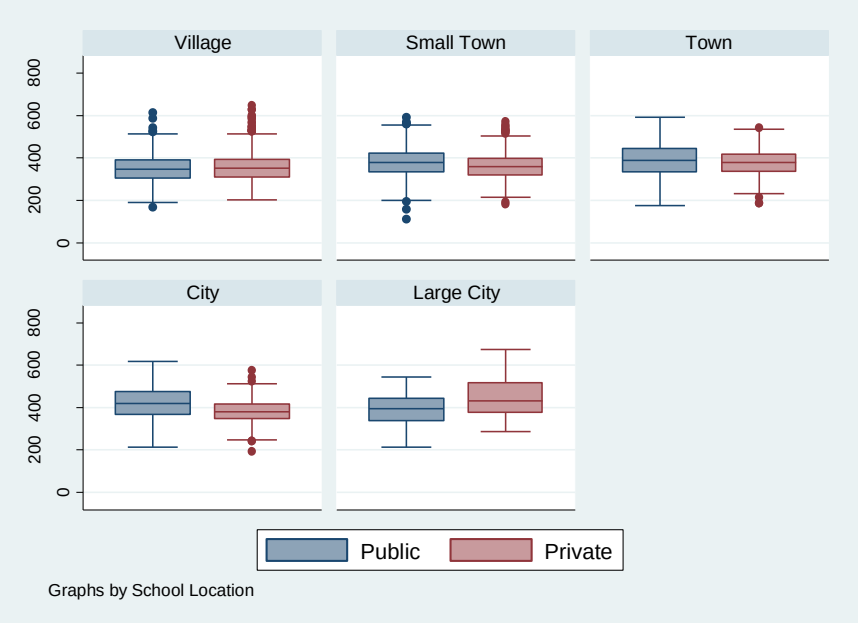


Figure 3.3. Descriptive figures of the Indonesian students participating in PISA 2012: (a) the distribution of the students' SES across different school locations[9], (b) the scatter plot of the students' math scores (from 1 plausible values) against their socio-economic and cultural status, and (c) the distribution of the students' math scores across different school type and locations.

The school survey, completed by the principal of the sampled school, also asked whether the school offered mathematics lessons "in addition to the mathematics lessons offered during the usual school hours" (OECD, 2014). If it did, then the school was instructed to specify the purpose of these additional mathematics lessons: for either enrichment, remedial, both, or even without differentiation depending on the students' prior achievement level. Although only schools that provided such additional lessons needed to specify its purpose, other schools, which did not, still responded to such questions. It was noted that this survey item could be misleading because it was unclear whether such additional lessons referred to one that could be offered during or outside school hours. As Table 3.9 shows, the majority of the sampled schools (75%), which provided additional mathematics lessons, did not distinguish between lessons that were solely aimed for enrichment and those intended for remedial purposes. This tendency was also echoed by the other schools, which did not indicate such offer of the additional mathematics lessons in the first place. However, the number of schools that did not specify the purpose of their additional math lessons is quite significant, i.e. 64 schools out of 209. A dummy variable

[9] School location represents community context in which the school is located: village (serving less than 3000 inhabitants), small town (3000-15,000 inhabitants), town (15,000 – 100,000 inhabitants), city (100,000 – 1 million inhabitants), and large city (> 1 million inhabitants).

(Yes/No item) indicating whether or not a school provides additional math lessons is included in the data analysis, while assuming that the additional lessons serve the purposes of both enrichment and intervention.

Table 3.9
The percentage of sampled schools that offered additional math lessons in PISA 2012

Provision	Purposes				Total
	Enrichment	Remedial	Both Enrichment & Remedial	Without Differentiation	
Yes	18%	7%	59%	16%	100%
No	0%	33%	50%	17%	100%

Note. Missing values excluded (i.e. cases from 64 schools).

The teacher certification related variable is a school-level variable representing the proportion of fully certified teachers to the total number of teachers in a sampled school. The mean proportion of certified teachers is approximately 59%, which means that there are approximately 6 certified teachers in every 10 teachers per school. Teacher participation in the certification process is often by invitation as it also depends on the availability of such programs being offered in the nearby areas and the number of teachers such program can take in. Including the proportion of certified teachers as a covariate related to OTL makes the assumption that certified teachers have sufficient knowledge on content and appropriate pedagogies and that mathematics teachers have the priority of getting certified first as mathematics has always been one of the major subjects tested in the local/school-based and national examinations.

Therefore, investigation of the effects/associations of the above-mentioned covariates is necessary to understand how the students' OTL differs across the four intended content areas after controlling for these background variables. The terms "dimension" or "aspect" will be used interchangeably throughout this paper to label the aspects of the proposed OTL measures.

3.6. Methods[10]

Item response theory has been commonly used to describe patterns of survey responses and make inferences about them in order to extract information about a latent construct for which the items manifested. There are four item response modelling approaches; such approaches are used for a wide variety of evaluation (De Boeck and Wilson, 2004). The first approach is called 'doubly descriptive'. This is when there is

[10] Most part of this section has been described in Chapter 1 and Chapter 2.

absence of properties of the item and person predictors. For example, measurement models like Rasch models normally use this approach for describing the extent to which item response patterns can depict the latent construct they are trying to describe. When there is at least one property of item or person incorporated in the model, such approach is called 'person or item explanatory' with the property coming from the person or item side, respectively. Lastly, the inclusion of both person and item properties constitutes the 'doubly explanatory' approach since the properties can be used to explain the effects of having such person property on the latent construct being represented by such item property. Similar to the approach taken in Chapter 1, this chapter describes the item response model that is in itself explanatory because of the use of the person background information for the evaluation of covariates associated with OTL.

For addressing Research Aim 1, I applied a between-item multidimensional partial credit model with latent regression to examine effects of student and school covariates on the student scores in the OTL-related items (see Figure 3.2). This represents a between-item multidimensionality because each of the OTL related items taps onto a single dimension. With this modelling approach, the given scores can explain how students perceive their OTL across different aspects and under diverse contexts. This study performs sub-dimension level analyses in order to examine the students' use of score categories within and across dimensions. A multidimensional random coefficient multinomial logit model (MRCMLM) is employed to calibrate the item parameters and ability estimates (Adams, Wilson, & Wang, 1997). The parameter estimation software ConQuest was used for estimating the parameters (Wu, Adams, Wilson, & Haldane, 2007; Adams, Wu, & Wilson, 2015), the same as is used by the PISA 2012 analysis itself (OECD, 2012). The MRCMLM is a generalized Rasch item response model that uses a scoring function and a design matrix to accommodate the applications of many existing IRT models used in this study such as the simple logistic model (Rasch, 1960), the partial credit model (Masters, 1982), and the multidimensional versions of these models (Adams, Wilson, & Wang, 1997).

Master's partial credit model (PCM) is used to deal with the polytomous items. In multidimensional PCM, each student/person p's latent OTL level estimate in dimension d is estimated using a probability model where the probability (P_{ik}) of answering an item i in response category k is a function of the difference between the location of person p and the location of item i. Incorporating R number of person background variables, the multidimensional latent regression with PCM can be formulated as

$$\eta_{pi} = \theta_{pd} - (\delta_i + \tau_{ik}) \qquad (3.1)$$

Here, $\eta_{pi} = \log\left(\frac{P_{ik}}{P_{ik-1}}\right)$ is the logit link to represent the probability model as a linear function of person latent OTL level θ on each dimension d, the relative item difficulty δ for a particular item i along with its *k*-th threshold parameter (τ_{ik}) when using PCM. The threshold parameter (τ_{ik}) is the deviation from the mean item difficulty δ_i for item i at step category *k* (i.e. $k = 0, \dots, K$) and constrained such that $\sum_{k=0}^{K} \tau_{ik} = 0$.

For a latent multiple regression model (see Wright & Masters, 1982; De Boeck & Wilson, 2004), the θ effect can be decomposed into Equation 3.2:

$$\theta_{pd} = (\sum_{r=1}^{R} \beta_{rd} Z_{pr}) + \varepsilon_{pd} \quad (3.2)$$

Similarly in this case, d indicates a specific latent dimension (i.e. $d = 1, \ldots, D$); θ_{pd} represents person p's latent ability parameter on dimension/construct d; τ_{ik} is the item step difficulty parameter for item i at category k (i.e. $k = 0, \ldots, K$); β_{rd} is the fixed latent regression coefficient of person covariate r (i.e. r = 1, …, R) on each dimension d that can be student-level or school-level covariates; Z_{pr} denotes the value of person p on covariate r; and ε_{pd} is the remaining person effect after the effect of the person covariates is accounted for ($\varepsilon_{pd} \sim N(0, \sigma_{\varepsilon}^2)$).

As previously described, the OTL items are polytomous indicating multiple time-related response categories, ranging from never/rarely/hardly ever to always/almost always/every lesson (see Table B.1.1 – B.1.3 and Table 3.1 – 3.4). In the actual survey, most of these response categories were in the reverse order. Hence, the item responses of those particular items were reversed during model calibration so that higher item difficulty corresponds to higher frequency. Specifically, the PCM models the probability of going from category level *k* to *k* + 1, such as 2 to 3, given the student has completed one of these steps. For the OTL measure(s), there are four or five levels, and thus three or four step parameters, respectively, to be estimated for each item—step 1 represents the difficulty of going from a category of 1 to 2, step 2 is the difficulty of going from a category of 2 to 3 given the student has completed the step from score 1 to 2, and step 3 is the difficulty of going from a score of 3 to 4 given the student has completed the previous steps, and so on.

The item parameter estimates needed for fitting a latent regression on the OTL's five-dimensional partial-credit model was adjusted using the Delta Dimensional Alignment[11] (DDA) technique (Schwartz, 2012), as described in Chapter 2. These items became anchors in the latent regression model when estimating the effect/association of each covariate across the OTL aspects. When fitting a multidimensional model in MRCMLM, it should be noted that the unadjusted estimates of person and item location cannot be directly compared across dimensions because they are on different scales as each dimension is separately centered at zero. The calibration of these item parameters incorporated the student sampling weights.

The latent regression model was then developed to examine the effects of such covariates as gender, student's SES, grade levels, school type, program, and locations, availability of additional math lessons in school, and the proportion of certified teachers on the proposed five-dimensional OTL measures (see Table 3.5). By estimating the latent regression coefficient directly from the item response data, one can avoid potentially misleading inferences about the differences in the means contributed by the covariates due to measurement errors. In this study, the effect size of each covariate on a particular OTL dimension/aspect is computed by dividing the estimate of the covariate's regression coefficient by the standard deviation of the respective latent variable from the unconditional five-dimensional model with no covariates (Wu, Adams, Wilson, & Haldane, 2007, p. 112). I

[11] Delta-dimensional alignment (DDA) is one approach to transform the parameter estimates of a multidimensional model onto the same metric to allow direct interpretation and comparison across dimensions. Further explanation about the DDA technique can be found in Schwartz (2012).

also examined the R^2 between the unconditional and conditional models to show changes in the variance explained by the model after including the covariates.

Given that ConQuest's regression function implements a listwise deletion approach when dealing with missing values of a covariate, the analysis for addressing the research Aim 1 used a subset of the data with N=4800 (about 14.6% considered as missing data from the total sample size). The sampled data was greatly reduced because there are quite a number of cases with missing values for the intended school-level covariates and also from cases who had a zero response on all of the OTL-related items. The percentage of the number of excluded cases falls within the suggested limit of 5 – 20% of missingness from the whole cases in the dataset before the amount of missing data is detrimental for data analysis purposes (Wu, et al., 2016, p.102). As the number of cases was reduced, the PISA student sampling weights[12] would not be applicable as is. However, the item parameter values used as anchors had been calibrated using weights, and thus the weights were not incorporated in the latent regression models (see Chapter 2).

Unfortunately, I could not use this dataset to address my second research aim due to the complexity of the data structure. Only one-third of the sample had a complete set of items, in addition to missing values for the covariates from both the students' and school principals' surveys. The school sample size ranged from 2 to 35 (i.e. students who responded to at least one item related to the OTL aspects). Because of the large amount of missingness in the item responses (due to the rotational survey design) and the complex case membership of the schools, I derived another subset of the data with N = 5547 to address Research Aim 2, after excluding cases with the following properties: (1) who responded zero to all of the OTL-related item, (2) who had missing student-level covariates, and (3) who attended schools with less than six sampled cases. About six clusters/schools with a low number of group members were excluded to avoid complication in the estimation of the covariance matrix.

Furthermore, effect coding was used for a better interpretation of the effect of the grade level variables, in which the 10^{th} grade level (the modal grade) becomes the reference group. In this case, the mean of the dependent variable for a given level is compared to the mean of the dependent variable for all levels of the variable (Sundstrom, 2010; UCLA, n.d.).

Using the generalized linear mixed model framework, the partial credit model as delineated in Equation 3.1 is a 2-level mixed-effect model in which the items representing repeated observations are nested within persons (De Boeck & Wilson, 2004; Rabe-Hesketh & Skrondal, 2012). Persons are regarded as a random sample from a population where the person abilities, which are the students' perspectives of their OTL level in this context, are identically and independently normally distributed. In this approach, items are considered as fixed effect. Hence, accounting for the effect of school clustering to address the Research Aim 2 can be done in two ways: schools as fixed effects or schools as random effects.

[12] The student survey weight is provided by PISA to account for the two-stage stratified sampling technique. It is calculated from the school base weight, the within-school base weight, some adjustment factors to compensate for non-participation by school/students, the school base weight trimming factor, and the final student weight trimming factor (OECD, 2014).

Ideally, as a generalization is often preferable when making inference about the school effect, schools are considered as a random effect.

Therefore, I modeled the school clustering effect as another fixed effect in ConQuest[13] by including school indicators as dummy variables in the latent regression model. To do so, Equation 3.2 was extended, as shown in Equation 3.3:

$$\theta_{gpd} = \left(\sum_{r=1}^{R} \beta_{rd}\, Z_{pr}\right) + \left(\sum_{s=1}^{S} \gamma_{sd}\, W_{ps}\right) + \varepsilon_{pd}\,. \qquad (3.3)$$

In Equation 3, γ_{sd} is the fixed latent regression coefficient of a school s (i.e. $s = 1 \ldots S$) on each dimension d and W_{ps} denotes the indicator of a person p attending a particular school covariate s. The definition of Z_{pr} in this case was then adjusted to only represent the student-level covariate r since the fixed effect approach cannot accommodate school-level covariate in the model (Clarke, Crawford, Steele, & Vignoles, 2010).

3.7. Findings and Discussion

This section delineates the extent to which the selected covariates show associations across the OTL measures after applying the latent regression approach using the 5-dimensional partial credit model onto the Indonesian dataset. It is important to evaluate these effects in order to highlight the utility of the collected background information in supporting the notion of OTL and how resource allocation can be better organized to enhance student's learning experience, specifically since OTL is associated with student academic performance (Gamoran, 1987; Wang, 1998; Lee, 2004; Allen et al., 2013; Schmidt, Zoido, & Cogan, 2013; Kurz & Elliott, 2014). In this study, a latent construct θ represents the level of OTL perceived by the sampled students in their schooling experience. Therefore, there are five θ-s for each sampled student, each of which represents the level of the latent OTL construct on one dimension/aspect.

3.7.1. Research Aim 1 – Effects of background variables

For Research Aim 1, I fitted a latent regression with eleven covariates, i.e. gender, SES, four grade levels, three school types (vocational, private, and rural schools), availability of additional math lessons offered by the school, and the proportion of certified teachers at the school, on the between-item 5-dimensional model of the OTL aspects. For estimating the regression coefficients of this "conditional" 5-dimensional model, I used the previously calibrated item parameters from the DDA-adjusted 5-dimensional model – Alternative Model 3 as the anchor parameters, as discussed in Chapter 2. By using this approach, a more "correct" inference about the differences in the group means can be made; specifically since

[13] Currently ConQuest version 4.5 accommodates Rasch/1-PL and 2-PL item response models corresponding to the 2-level generalized linear mixed model approach.

the group means are estimated directly from the item response data. In this study, gender, school type, program, and location, and availability of additional mathematics lessons offered by the school for each student were used as covariates using dummy variables for each category of the corresponding covariates. Table 3.10 lists the standardized effect size[14] of each of the background variables across the OTL aspects. Cohen's classification of the magnitude of effect size is used in the discussion of the findings, in which an effect size of 0.2 is considered small, 0.5 is medium, and 0.8 is a large effect size (Cohen, 1992). For the reference groups, I chose female for gender, public schools for school type, non-rural schools (serving a community of greater than 3000 inhabitants) for school location, non-vocational schools (regular and Islamic secondary schools), 10^{th} grade level for the grade level covariates.

The Gender Effect. After controlling for other covariates, statistically significant gender effects appeared across the CE, SI, and HA aspects as shown in Table 3.10. Here, boys seemed to perceive their content exposure of mathematics topics and problems/tasks rather negatively compared to girls. This finding might also indicate that boys could be less able to recall their experience in math related topics or problems at school. The boys, however, responded positively on items related to SI and HA, with an effect size of at least 10%. This means that, on average, the mean perspectives of the boys on the SI and HA aspects were 10% of a standard deviation higher than the girls. A further investigation is warranted to explore potential reasons on why girls were more likely to perceive less on the occurrence of instructional strategies that promote collaboration, independence, and higher-order learning in their mathematics classroom. Hence, a similar suggestion as presented in Chapter 1 can be recommended by delivering better gender-friendly teaching and learning practices to improve the girls' understanding and interest in learning mathematics.

The Socio-economic Status Effect. As anticipated, SES showed a significant and positive effects on the students' perspective to the CE, DI, and HA aspects, after accounting for other covariates. A unit increase in one standard deviation of the SES indicator predicts an increase of about 18% of the students' opportunity to learn in the content exposure aspect, 6% of having experienced direct instructional strategies, and 9% of presented with higher-order assessment. These effects confirmed past studies that students from affluent families were exposed to more rigorous curriculum content so that these students could recall heir experience with the math topics and problems better than their less economically-advantaged schoolmates (Gamoran, 1987; Grisay, Gebhardt, Berezner, & Halleux-Monseor, 2007). The schools attended by the students with high SES were perceived to have a more direct-teaching approach and give more challenging tasks, although the respective effect sizes were significantly low. When relating the effect size of SES with that of the private school along the CE aspect, the effect size of attending private schools was a low negative, confirming that students with higher SES went to good schools, specifically public schools, which normally provide strong curriculum, teacher-centered learning, and a more robust assessment system (MoCEI, 2013b). In Chapter 1, it has also been reported that students with higher SES tended to perform better in all mathematics content areas.

[14] The standardized effect size is calculated from dividing the regression coefficient of a covariate by the standard deviation of the unconditional model.

Table 3.10
Effect sizes of the multidimensional latent regression with partial credit model (DDA-adjusted) of the Opportunity-to-learn measures in PISA 2012 addressing Research Aim 1 (N = 4800)

Covariate	Content Exposure (CE)	Direct Instruction (DI)	Student-oriented Instruction (SI)	Higher-order Assessment (HA)	Teacher Support/ Feedback (SF)
Male	**-.22***	-.02	**.13***	**.10***	.02
SES	**.18***	**.06***	.03	**.09***	.02
7^{th} – 8^{th} Grade	**-.56***	.09	**.34***	.05	**.13***
9^{th} grade	.03	**.12***	**.37***	.06	**.32***
11^{th} grade	.31	.00	-.13	-.08	.06
12^{th} grade	.01	-.36	**-.63***	-.14	**-.69***
Vocational High School	**-.17***	.04	.02	-.10	.07
Private schools	**-.08***	-.04	.07	.05	-.03
Rural schools	.04	-.03	**-.13***	**-.13***	-.06
Additional math lessons in school	**.10***	.07	**-.17***	.04	-.08
Proportion of certified teachers	-.02	-.08	**-.13***	.01	**-.19***
Conditional Variance	*.28*	*.77*	*.46*	*.52*	*.35*
(Std. Error)	*(.006)*	*(.016)*	*(.009)*	*(.011)*	*(.007)*
Unconditional Variance	*.32*	*.81*	*.49*	*.53*	*.36*
(Std. Error)	*(.007)*	*(.016)*	*(.010)*	*(.011)*	*(.007)*
R^2 (%)	13.62	4.84	6.75	1.33	2.79

Notes. *** (bolded)** statistically significant at $p < 0.05$. The effect size is calculated by dividing the estimate of the regression coefficients by the unconditional standard deviation of the respective latent variables, whereas the R^2 denotes changes in the variance explained by the model after including the covariates.

The Grade Effect. In this case, effect coding is used in which the comparison should be made between the mean of a given level and the mean of the means of the respective OTL aspect. Compared to the mean level of OTL on the CE aspect from all grades, the 7^{th} graders had approximately 50% of significantly negative effect size. This moderately negative perspective on the items related to the prescribed math content and types of tasks showed that these students might have missed out on the opportunity to learn such topics and problems in mathematics according to their growth rate since this group would likely be the 15 year-olds who repeated grade or started school late due to various reasons. Grade retention was significantly associated with lower reading and mathematics achievement at age 14, above

and beyond an extensive set of explanatory variables (McCoy & Reynolds, 1999; Silberglitt, Appelton, Burns, & Jimerson, 2006; Allen et al., 2009).

However, the 7th – 8th graders perceived the SI and SF aspects in a statistically significant positive way, and so did the 9th graders. In addition, the 9th grade students also perceived positively on the DI and SF aspects as indicated by the low to moderate positive effect sizes. One should note that the PISA 2012 test was administered in March – April 2012 (MoECI staff, personal communication, January 2015), right before the mandatory 9th grade and 12th grade high-stakes NEs were normally held (early-mid May). These positive effect sizes during the preparation of NEs may warrant further evaluation. However, the timing of the PISA survey can explain the large significantly negative effects from the 12th graders regarding the SI and SF aspects as they might be under a lot of pressure in preparing for the high-stakes NE[15] and the university entrance test[16]. The first possible explanation about the large negative effect sizes is that the classroom practices, which put emphasis on students' needs and provide adequate emotional support and feedback, did not occur frequently in the 12th grade classroom as teachers would tend to do more teacher-centered learning when both the teacher and the students focused on the preparation for the NE (MoCEI, 2013b). Second, the high pressure of learning for knowledge and for passing the high-stakes NE might have made the students become apathetic in learning and thus responded negatively on the related items. Furthermore, one should also note that the 15-year-olds who were already in 12th grade at the time of the test can also be a special group of students. They could have either skipped class or sat in a special advanced class due to their academic talent.

As in Chapter 1, these findings also suggest that intervention approaches other than grade retention are needed to better promote school achievement and adjustment (NASP, n.d.). More attentions should also be given to the 9th and 12th grade classrooms as the students preparing for the NEs should be provided with the most effective opportunity in learning and experience of all the materials needed to succeed in the tests.

The School Program Effect. As anticipated, students attending vocational schools gave negative responses on the CE aspect (effect size = -.16). Since vocational schools implemented more focused curriculum to skills mastery rather than the theoretical foundation of subject matter (MoCEI, 2013b), these students might not have the experience and exposure to the prescribed mathematics contents and types of problems/tasks. Having less OTL in the curriculum content would eventually bring a disadvantage to the vocational school graduates as they cannot be on par when competing in a job market with graduates from regular schools (MoECI, 2013b; OECD-ADB, 2015). Hence, this finding supports the MoECI's plans to revitalize the curriculum of the vocational education program and improve the quality of its educational resources (KSP, 2016; Media Indonesia, 2016).

The School Type Effect. After controlling other covariates, a statistically significant, but small, negative effect for the CE aspect signaled a less rigorous curriculum exposed to the sampled students attending private schools. This confirmed the superior performance of the public schooling system in Indonesia in terms of the quality of the implemented

[15] The NE normally covered all materials from the previous three grade levels.
[16] The university entrance test is held nationally for admission to the public universities located around Indonesia, to which any student can compete to be admitted.

curriculum (MoEC, 2013b; OECD-ADB, 2015). As Newhouse and Beegle (2006) suggested, public schools were the most preferable school choice in Indonesia. They argued that this superior benefit was mainly due to the high quality of students admitted to public schools. This finding suggests that Indonesia needs to put more attentions and efforts to provide equal and equitable OTL at private schools, or attracting more students to the public schools while maintaining the quality. In addition, monitoring of national curriculum implementation at these private schools needs to be done to ensure better quality of learning environment.

The Rural School Effect. A dummy variable to indicate rural schools is included in the proposed OTL measures to provide some insights on how the OTL was perceived at rural schools as it is also typically associated with the students' SES and the degree of the school's facility and capacity. Serving a community of less than 3000 inhabitants, the rural schools were most likely be located in remote regions with low socio-economic environments (see Figure 3.3). Students attending rural schools, on average, appeared to come from families with lower SES than the students from non-rural schools as Figure 3.4 shows. In the context of Indonesia, these areas are normally located in low socio-economic environments with minimal educational resources (Suryadarma & Jones, 2013). Controlling for the other covariates, compared to students attending non-rural schools, the rural school students reported to receive approximately 10% less of a student standard deviation in the mean quality of the classroom learning environment that supported student-oriented instructional strategies and higher-order assessment practices. Although the effect sizes were statistically significant and low for the SI and HA aspects, the findings likely reflect the real conditions in which the rural school teachers often found it difficult to implement innovative teaching strategies since they often were required to handle large and multi-grade classes with minimal classroom resources. In such conditions, these teachers tend to find the teacher-centered learning approach easier (Luschei & Zubaidah, 2012). Interestingly, there is no significant effect detected for the CE aspect, making one speculate that there was no substantial difference in the content exposure between rural and non-rural students due to the adherence to the national curriculum standards. Finally, the findings may reinforce the need to support the rural teachers by ensuring an equitable distribution of financial and educational resources so that the rural students can have the opportunity to learn in a student-oriented learning environment and be exposed to curriculum and instructions designed to encourage higher-order thinking and its corresponding assessment.

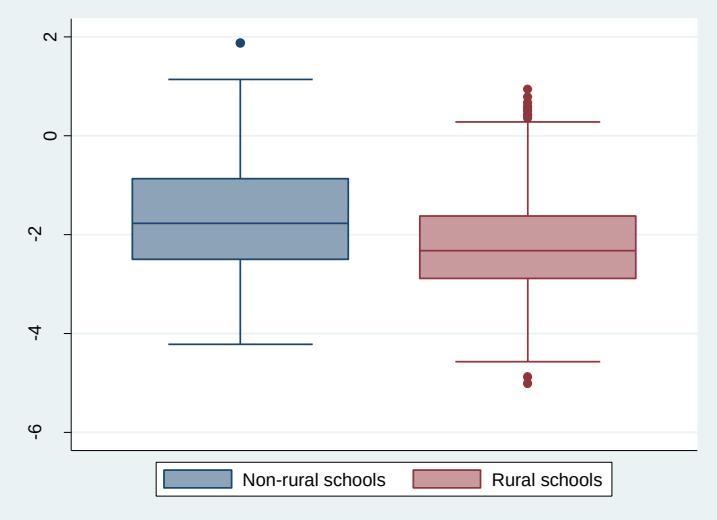


Figure 3.4. Distribution of the sampled students' socio-economic status (SES) who attended rural and non-rural schools in Indonesia.

The Additional Math Lessons Effect. This covariate represents schools that offered additional math lessons outside the school curriculum hours. Students attending such schools reported positively on the exposure of the prescribed math contents and types of problems/tasks by having a statistically significant positive effect size, as anticipated. These students, on average, have also seemed to experience less student-oriented learning (effect size = -.17), confirming the Indonesian teachers' persistence in using rote-learning and the teacher-centered approach (World Bank, 2010; MoCEI, 2013b; OECD-ADB, 2015).

The Teacher Certification Effect. The proportion of certified teachers indicates the number of certified teachers per the total number of teachers in each school. In the sampled schools, there was, on average, about 6 certified teachers per 10 teachers. In rural and/or private schools, the average ratio was less, i.e. only about 4 to 5 certified teachers per 10 teachers per school. Teachers who teach main subject matters, such as mathematics, science, and language, tend to be the ones who got to undertake the certification program (MoECI, 2013b). Apart from being trained on content knowledge, the teachers were also trained on using innovative and active instructional strategies to promote student-centered learning and create effective and pleasant learning environment. Hence, it is interesting to see the statistically significant negative effects on the SI and SF aspects, which suggest that having more certified teachers at school is associated with less students' opportunity to experience student-oriented learning environment in which the teacher showed positive support and feedback. Further studies are warranted to investigate these findings.

As the last rows of Table 3.10 show, the variance of each of the latent OTL aspects was only slightly reduced by the inclusion of the selected background variables. The ΔR^2 values indicate the changes in the variance explained by the regression model. In this case, the regression model can only explain marginally more variance for the CE related items (ΔR^2 = 13.62%) than it does for the other OTL aspects. This finding may suggest that the

content-related aspect of OTL was moderately influenced by the school settings, as confirmed by past studies (Schmidt & Maier, 2009; Schmidt, Zoido, & Cogan, 2013).

3.7.2. Research Aim 2 – School clustering effect

To address Research Aim 2, I included school clusters as fixed effects in the regression model (as school dummy variables, see Equation 3.3). Table 3.11 lists the effect sizes of the student-level background covariates, which were also included in the latent regression model in the preceding section. The school-level background covariates included in the preceding model are not incorporated in Model 2 and Model 4 in Table 3.11. These models have already included the school fixed effects and thus, the effects of all school-level covariates would have already been accounted for. The change in the proportion of explained variance for the models listed in Table 3.11 is presented in Table 3.12. The association between each of the student-level background covariates and the OTL aspects after accounting for the school clustering are described as follows.

The Gender Effect. After controlling for other covariates and school clustering, similar statistically significant gender effects appeared across the CE, SI, and HA aspects as they did in the previous latent regression model without the school fixed effect (see Model 4 in Table 3.11).

The Socio-economic Status (SES) Effect. In this case, SES also showed similarly statistically significant and positive associations on the students' perspectives to the aspects of CE, DI, and HA, after accounting for other covariates and school clustering. However, the magnitude of the effect size of SES on the CE aspect was reduced by about 8 percent-point after comparing the effect size of SES in Model 2 and that of SES in Model 4 (see Table 3.11). When the school clustering is held constant, every unit increase in the SES[17] measure is, on average, associated with a 12% of a student standard deviation, on the exposure and experience of the prescribed mathematics content topics and types of problem/tasks. This suggested that the students' OTL in the CE aspect varied depending on the specific school characteristics, confirming the previous finding that high-SES students were more likely to attend schools with good provision of OTL.

[17] In PISA 2012, the SES index is standardized to a mean of zero for the population of students in OECD countries. A one-unit difference on the scale represents a difference of 1 standard deviation on the scale distribution (OECD, 2014).

Table 3.11

Effect sizes of the multidimensional latent regression with partial credit model (DDA-adjusted) of the Opportunity-to-learn measures in PISA 2012 addressing Research Aim 2 (N = 5547)

Parameter/ Covariate	Content Exposure (CE)	Direct Instruction (DI)	Student-oriented Instruction (SI)	Higher-order Assessment (HA)	Teacher Support/ Feedback (SF)
Model 1 – Unconditional latent regression, with no covariates					
Variance	*.35*	*.70*	*.47*	*.56*	*.39*
(Std. Error)	*(.007)*	*(.013)*	*(.009)*	*(.010)*	*(.007)*
Model 2 – Conditional latent regression with student-level covariates					
male	**-.23***	-.02	**.15***	**.12***	.04
SES	**.20***	**.06***	**.04***	**.11***	.02
7^{th} – 8^{th} Grade	**-.47***	.05	**.31***	.05	**.14***
9^{th} grade	.04	.10	**.37***	.10	**.29***
11^{th} grade	**.29***	.01	-.15	-.07	.08
12^{th} grade	.03	-.36	**-.62***	-.24	**-.72***
Variance	*.30*	*.69*	*.47*	*.55*	*.38*
(Std. Error)	*(.006)*	*(.013)*	*(.009)*	*(.010)*	*(.007)*
Model 3 – Conditional latent regression with schools as fixed effects					
Variance	*.27*	*.66*	*.37*	*.48*	*.31*
(Std. Error)	*(.005)*	*(.012)*	*(.008)*	*(.008)*	*(.006)*
Model 4 – Conditional latent regression with student-level covariates + schools as fixed effects					
male	**-.21***	-.02	**.17***	**.12***	.03
SES	**.12***	**.05***	**.06***	**.07***	.02
7^{th} – 8^{th} Grade	.00	**.74***	.46	.51	**.95***
9^{th} grade	.46	**.73***	.48	**.58***	**1.14***
11^{th} grade	-.28	**-.54***	-.31	-.38	**-.44***
12^{th} grade	.19	-.55	-.68	-.54	**-1.31***
Variance	*.26*	*.57*	*.32*	*.48*	*.31*
(Std. Error)	*(.005)*	*(.011)*	*(.006)*	*(.009)*	*(.006)*

Notes. *** (bolded)** statistically significant at $p < 0.05$. The effect size is calculated by dividing the estimate of the regression coefficients by the unconditional standard deviation of the respective latent variables.

The Grade Effect. Among the different grade levels, the inclusion of school fixed effects have changed the magnitudes and the statistical significance of the effect sizes across the OTL aspects. Figure 3.5 – 3.9 depict the distribution of the mean raw scores for the students' perspectives on each of the OTL aspects with respect to grade levels. The higher mean values denote more opportunities a student has to experience or be exposed to a particular learning environment as stated in the specific item. After controlling for other

covariates and the school clustering, compared to the mean perspectives of all sampled students, the statistically significant negative perspectives of the 7^{th} – 8^{th} graders on their exposure to the prescribed mathematics contents and types of problems/tasks disappeared. Albeit most of these students rated the CE-related items comparably lower than their cohorts at different grade level as shown in Figure 3.5, this finding may have been due to the complex combination of the distribution of items and the distribution of this group among the sampled schools. As a matter of fact, the sampled 7^{th}-8^{th} graders were spread out across 74 schools.

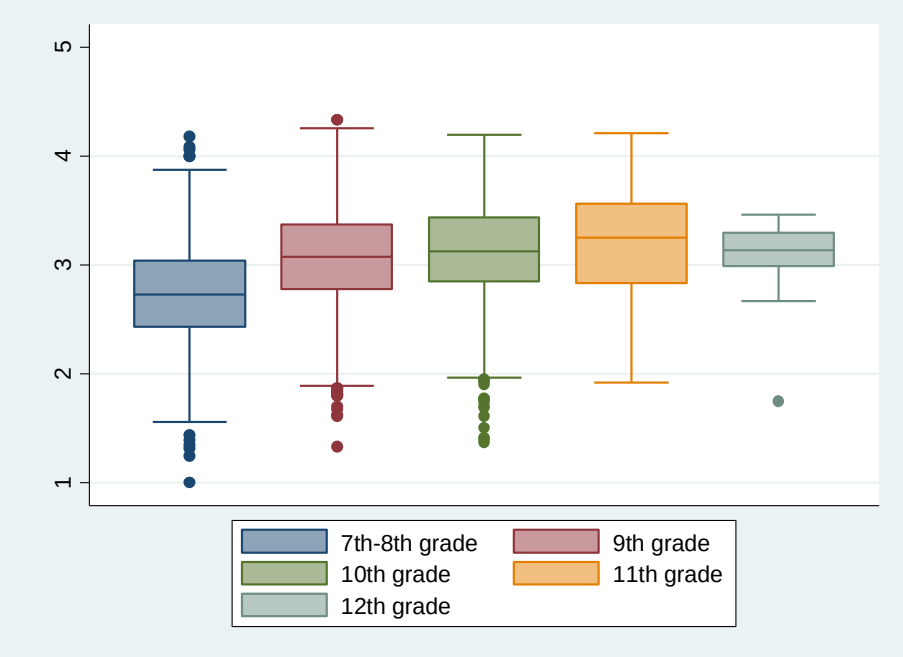


Figure 3.5. Mean raw score distribution of the students' perspectives on the Content Exposure (CE) aspect with respect to grade levels.

For the DI aspect, the school clustering effect has made the 7^{th}-8^{th} grade and the 9^{th} grade group show large statistically significant positive effects, after controlling for other covariates. As Figure 3.6 shows, the junior secondary school students rated higher on this DI aspect confirming that fact that the most Indonesian teachers predominantly used the direct-instruction approach in their teaching (MoECI, 2013; OECD-ADB, 2015).

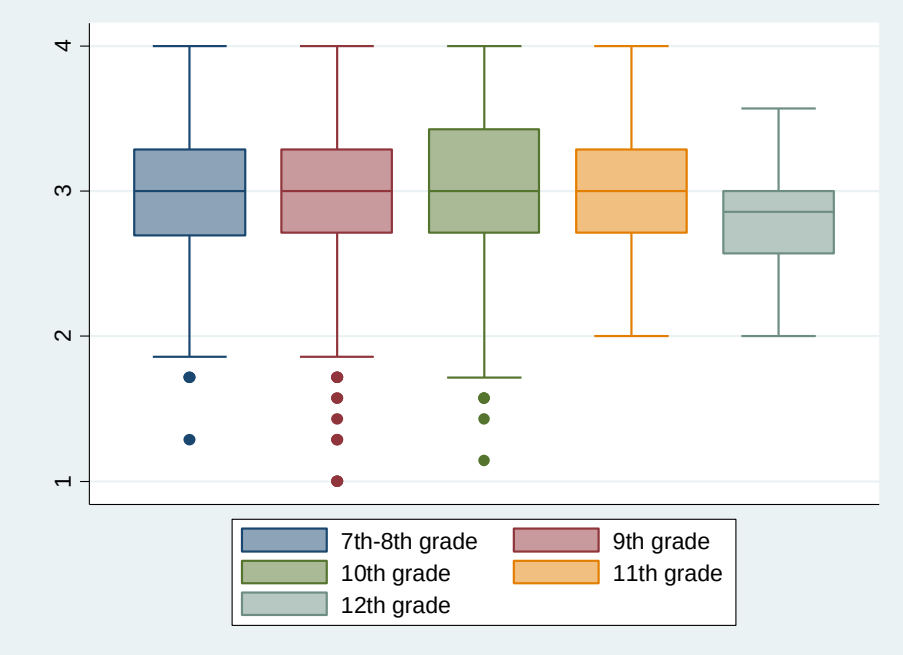


Figure 3.6. Mean raw score distribution of the students' perspectives on the Direct Instruction (DI) aspect with respect to grade levels.

The association of being in the junior secondary school groups (i.e. 7th -8th and 9th grade) with the SI aspect were diminished when the school cluster effect was incorporated (see Figure 3.7). This finding suggests that holding the school clustering constant, being in these particular group did not have a significant association with the students' OTL level on the SI aspect. Furthermore, in the case of the HA aspect as illustrated in Figure 3.8, the 9th grade group showed a statistically significant positive effect of 0.58, meaning that this student group felt itself to have more opportunity in learning practices that supported higher-order assessment than the average sampled students, after controlling for other covariates. Next, Figure 3.9 seems to indicate that the 12th grade students perceived lower teacher support/feedback compared to their cohort in the lower grades, which was also noted with the statistically significant and negative effect on the SF aspect in Table 3.11. However, there are only nineteen 12th grade students sampled from 3 schools in the dataset, and hence it is considered too small of a sample size to comment.

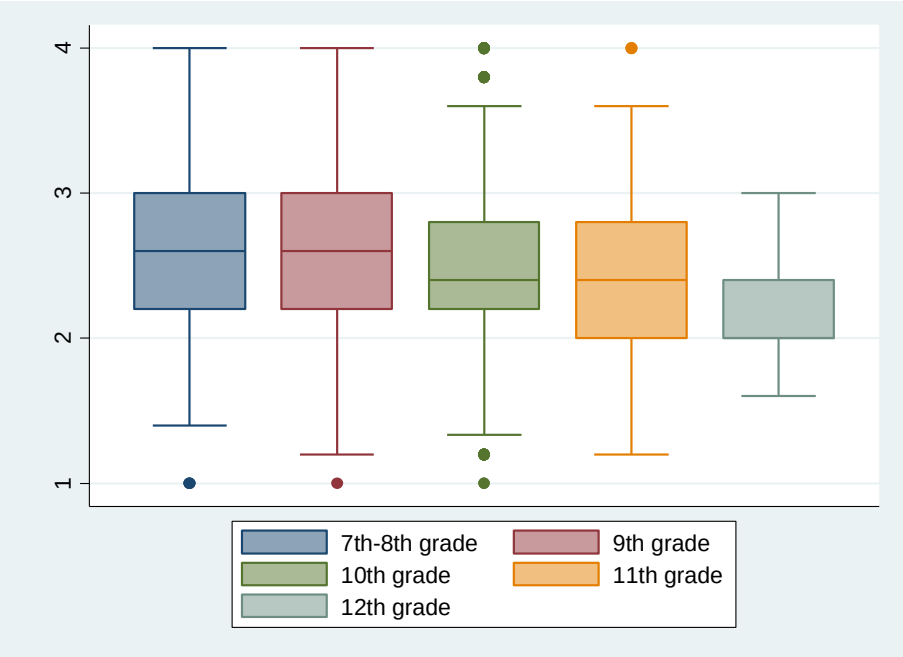


Figure 3.7. Mean raw score distribution of the students' perspectives on the Student-oriented Instruction (SI) aspect with respect to grade levels.

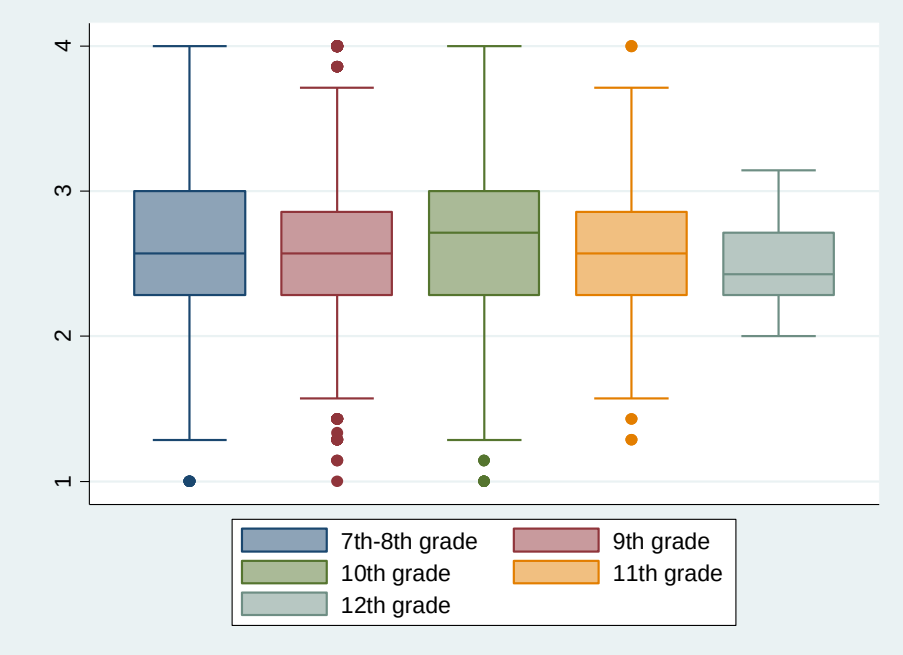


Figure 3.8. Mean raw score distribution of the students' perspectives on the Higher-order Assessment (HA) aspect with respect to grade levels.

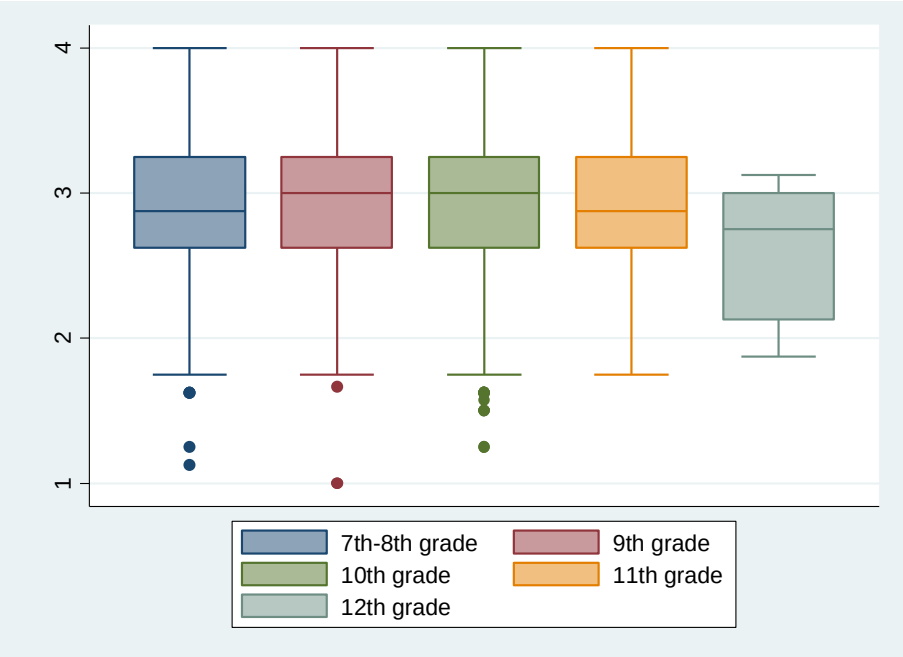


Figure 3.9. Mean raw score distribution of the students' perspectives on the Teacher support/Feedback (SF) aspect with respect to grade levels.

Table 3.12
The change in proportion of explained variance for the multidimensional latent regression with partial credit models across the opportunity-to-learn aspect/dimension

Change in Explained Variance* (ΔR^2)	**Aspect/Dimension of Opportunity-to-learn Measures**				
	Content Exposure (CE)	**Direct Instruction (DI)**	**Student-oriented Instruction (SI)**	**Higher-order Assessment (HA)**	**Teacher Support/ Feedback (SF)**
Model 4 vs. Model 2	13.6	17.6	31.6	12.8	17.6
Model 2 vs. Model 1	13.7	1.3	0.2	1.4	4.6
Model 3 vs. Model 1	23.7	5.7	19.7	13.9	22.1
Model 4 vs. Model 1	25.4	18.7	31.8	14.1	21.4

Notes. * The ΔR^2 denotes changes in the variance explained by the model after including the covariates (Model 2 – 4) with respect to the unconditional model (Model 1, without covariates).

As one can see from the Model 2 variance across the OTL aspects in Table 3.11, including only the student-level background covariates in the latent regression model did not explain large variance of the degree of the students' OTL aspects, except for the CE aspect. For this CE aspect, the change in the proportion of variance explained, ΔR^2, is around 14% as indicated in the second row of Table 3.12.

Furthermore, having compared the variance of Model 3 (including the school fixed effects only) and the unconditional Model 1 (without covariates) gives larger values of ΔR^2 ranging from approximately 6% to 24% across OTL aspects (see the third row of Table 3.12). These results suggest that the school clustering could explain more variance and thus, point to the importance of the variability of OTL level among the sampled schools. Next, when the student-level covariates, i.e. gender, SES, and grade levels, are added and accounted for in Model 4, the values of ΔR^2 increased and ranged from 14% to 32% across OTL aspects (see the last row of Table 3.12). The highest school clustering effect for Model 4 was found on the DI and SI aspects, i.e. ΔR^2 = 18.7% from 5.7% for DI and ΔR^2 = 31.8% from 19.7% for SI, as presented in the 4th and 3rd rows of Table 3.12. This suggested that the practices of direct and student-oriented teaching varied a lot across different schools. The lowest change in the proportion of variance explained was observed for the HA aspect (ΔR^2 = 14.05%), indicating that the school clustering effect provided less contribution in explaining the students' experience with higher-order assessment than it did on the other OTL aspects.

These findings revealed that the large school variance occurred between-schools, which might be due to the selective school admissions in Indonesia. As discussed previously, high-achieving students with high prior academic scores, normally with high SES background, were more likely to attend public schools that offered more rigorous curriculum and provided high quality teachers. Thus, the degree of OTL varies relatively more at the between school level than at the within school level, suggesting that schools did provide OTL via their enacted curriculum and instructional quality. Meanwhile, there was relatively less variance of OTL within schools as the students' OTL only varied in how much the students benefit in the rigor of curriculum and teaching quality in their own schools. These results were partly comparable with those of Schmidt et al. (2015) that in some countries, the variability of OTL were larger between-schools and high-SES students appeared to be provided with more OTL.

From the above discussions in Section 3.7.1 and 3.7.2, I assert that incorporating school clustering effect is essential in the investigation of the associations of the selected background variables on the OTL aspects. Including school fixed effects into the latent regression model did vary the associations and was able to explain more variability in the degree of OTL across its aspects.

3.8. Limitations and Future Directions

Although the current study has presented some useful utilizations of PISA's background information in terms of how such information shows associations across the aspects of OTL, several limitations should be noted. First, due to rotational matrix design of the student background survey, only one-third of the sampled students took the complete set of items. Thus, each student in the rest of the sample had missing values for at least one third of the OTL items by design, in addition to possibly more missingness from the items that the

student chose not to respond to. Next, there were 209 schools sampled in Indonesia, each of which further sampled 2 to 35 students to participate in the PISA 2012 test. The combination of the intricate missing-values pattern and the school membership has made the data structure so complex that fitting in a multidimensional partial credit model with school clustering as fixed effects was problematic. As incorporating the school clustering mediated the role of the selected background variables on explaining the variability of the students' OTL level, a better representation of the school clustering effect needs to be sought. For future research, a hierarchical linear item response model with school clustering as random effects may be developed, while using a better parameter estimation approach that can handle a complex pattern of missingness.

As also described in the Limitation section of Chapter 2, the OTL-related indicators were obtained from student self-report that can potentially be subject to social desirability bias, specifically if the sampled students systematically inflate or deflate their perspectives on their own classroom learning environment. However, Schmidt, Zoido, & Cogan (2013) have argued the 15-year olds would be mature and experienced enough to recall what they had been exposed to in the math class. Although some items did prompt the students to only think about the mathematics teacher who just taught them in the last mathematics class, I should note that the student ratings on the OTL-related items might also represent their accumulative perceptions of OTL throughout their schooling. To strengthen the students' self assessment of their OTL, albeit being time and cost intensive, a structured or semi-structured observational survey of classroom interaction could be proposed to better portray the effectiveness of the learning environment.

Another future study could explore the following topics to better investigate the effects of background information on the students' OTL level: (1) incorporating approaches like signal detection and anchoring vignette[18] to overcome potential bias due to social-desirability nature of the contextual items (OECD, 2014), and (2) investigating the extent to which the students' intrinsic characteristics (e.g. self-efficacy, self-concept, math anxiety, and work ethic in mathematics) relate to their perspectives of OTL as these characteristics might be the factors that drive the students to fully utilize the OTL provided to them at school.

3.9. Conclusion

I began this study with the primary aim of illustrating the utility of background information obtained along with PISA for helping decision makers in designing effective policy tools that can provide equitable and high quality of educational opportunities to improve student academic performance. Expanding on the study presented in Chapter 2, I investigated (1) the associations of selected background variables, such as gender, SES, grade level, school type, program, and location, the availability of additional math lesson,

[18] There were a few items dedicated to test these approaches in PISA 2012 round.

and the proportion of certified teacher per school, on the aspects of OTL as perceived by the students, and (2) whether these effects would differ when accounting for the school clusters.

For the 1st research aim, I fitted a five-dimensional (5-D) partial credit model with latent regression and the previously calibrated DDA-adjusted item parameter estimates as anchors. The 5-D model used in this analysis is the proposed OTL measures as developed in Chapter 2 that was chosen to better explain the variability of the students' performance in PISA 2012 math. By doing so, a more "correct" regression coefficient can be estimated as measurement error has been taken into account in the estimation process, and also direct comparison across dimension (or referred as to OTL aspect) would be possible with the application of the DDA method. Important statistically-significant positive effects or associations were found from SES and additional math lessons in school. These findings suggest that students with high SES were more likely to attend public schools that provided good opportunities to experience and be exposed to the prescribed mathematics contents and types of problems/tasks, and a learning environment that supported direct instructional strategies and higher-order assessment/school work. The statistically significant negative effect sizes on some OTL aspects were shown to come from students with the following characteristics: being retained (or not) in the $7^{th} – 8^{th}$ grade, accelerated to the 12^{th} grade, attending vocational schools, or attending private schools. When the school clustering effect was ignored, the inclusion of the selected background variables can help explaining the variability of the CE aspect more than the other aspects ($\Delta R^2 = 13.6\%$).

In addressing the 2nd research aim, I incorporated the school clustering as fixed effects added to the previously-developed latent regression model. The inclusion of school fixed effects by themselves could marginally increase the values of ΔR^2 from around 14% to 24% across OTL aspects. When the student-level covariates were also included, ΔR^2 of the latent regression Model 4 in comparison to the unconditional model increased slightly for the aspect of content exposure and moderately for the aspect of instructional strategies (either direct or student-oriented teaching). These results suggest that the school clustering could explain more variance and thus, gives more associations to the variability of the OTL level among the sampled schools. The highest school clustering effect was found for the SI aspect ($\Delta R^2 = 31.8\%$), suggesting that the practices of student-oriented teaching and learning were quite different across the schools. The lowest variance explained was observed for the HA aspect ($\Delta R^2 = 14.1\%$), which indicates that the school clustering effect provided less contribution in explaining the students' experience with higher-order assessment than it did on the other OTL aspects. From these results, one can see that the Indonesian students' OTL did vary between schools and that there was a considerable variability in the implemented curriculum at each school from what was intended, albeit the existence of mandated national curriculum standards for decades.

Findings from this chapter can make a major contribution to the discourse surrounding the OTL measures in large-scale assessment, particularly the use of student-level and school-level background information to help improve the educational opportunities for students from diverse backgrounds. The use of explanatory item response modelling approach, specifically the multidimensional partial credit model with latent regression, in the data analysis helps disaggregate the patterns of students' responses to the OTL-related survey

items so that the overarching latent OTL construct can be validly and comparably measured across its aspects, while estimating the effects/associations of such background information on each of the OTL aspects. The relative differences of the effects the OTL aspects can then be utilized to hypothesize ways to improve the curriculum structure, the content pedagogical knowledge, and the school delivery standards.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162-172.

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

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised Item Response Modelling Software* [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

American Educational Research Association, American Psychology Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (3rd Ed.). Washington, DC: Author.

Akiba, M., LeTendre, G. K., & Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational researcher, 36(*7), 369-387. doi: 10.3102/0013189X07308739

Albano, A. D., & Rodriguez, M. C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement, 75*(3), 836-856. doi: 10.1177/0013164413487375

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multi-level analysis. *Educational Evaluation and Policy Analysis*, 31(4), 480-499. doi:10.3102/0162373709352239

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the Classroom Assessment Scoring System – Secondary. *School Psychology Review, 42*(1), 76-98.

Anderson, S., & Mundy, K. (2014). *School improvement in developing countries: Experiences and Lessons learned*. Ontario Institute for Studies in Education. University of Toronto. Retrieved from https://www.oise.utoronto.ca/cidec/UserFiles/File/Research/School_Improvement/Anderson-SIP_Discussion_Paper-08042015.pdf

ASEAN. (n.d.). *About ASEAN*. Accessed on 23 Sep 2016. Retrieved from http://asean.org/asean/about-asean/

Badan Penelitian dan Pengembangan. (2016). *Tentang PISA*. Jakarta: Ministry of Education and Culture of Indonesia. Retrieved from http://litbang.kemdikbud.go.id/index.php/survei-internasional-pisa

Badan Perencanaan dan Pembangunan Nasional. (2016, Feb 23). *Prioritas nasional: Pembangunan Pendidikan*. Retrieved from http://www.bappenas.go.id/files/penyusunan_rkp_2017/seri_multilateral_meeting/Prioritas_Nasional_Pembangunan_Pendidikan.pdf

Badan Standar Nasional Pendidikan. (2006a). *Standar kompetensi dan kompetensi dasar SMP/MTs*. Jakarta, Indonesia: Author.

Badan Standar Nasional Pendidikan. (2006b). *Standar kompetensi dan kompetensi dasar SMA/MA*. Jakarta, Indonesia: Author.

Baswedan, A. (2014*). Gawat darurat pendidikan di Indonesia*. Jakarta, Indonesia: The Republic of Indonesia Ministry of Education and Culture. Retrieved from http://www.kemdiknas.go.id/kemdikbud/sites/default/files/Paparan%20Menteri%20-%20Kadisdik%20141201%20-%20Low%20v.0.pdf

Baumert, J., Kunter, M., Blum, W., Brunner. M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student Progress. *American Educational Research Journal, 47*(1), 133-180.

Betts, J. R., Zau, A. C., & Rice, L. A. (2003). *Determinants of student achievement: New evidence from San Diego*. San Francisco, CA: Public Policy Institute of California. Retrieved from http://www.ppic.org/content/pubs/report/R_803JBR.pdf

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system Performance. *OECD Education Working Papers*, No. 71. OECD Publishing. http://dx.doi.org/10.1787/5k9fdfqffr28-en

Briggs, D. C. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., and Easton, J. Q. (2010). *Organizing schools for improvement. Lessons from Chicago*. Chicago, IL: The University of Chicago Press.

Burge, B., Lenkeit, J. & Sizmur, J. (2015). *PISA in Practice: Cognitive Activation in Maths.* Slough: NFER. Retrieved from https://www.nfer.ac.uk/publications/PQUK04/PQUK04_home.cfm

Burstein, L. (1993). Studying learning, growth, and instruction cross-nationally: Lessons learned. About why and why not to engage in cross-national studies [prologue]. In L. Burstein (Ed.), *The IEA study of mathematics III: Student growth and classroom processes* (pp. xxvii-lii). New York: Pergamon.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Methropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75(1), 33-57. Doi: 10.1007/S11336-009-9136-X

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 6*4(8), 723-733.

Central Intelligence Agency. (2013). *The world fact book. Indonesia*. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/id.html

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement, 52*(2), 200-222.

Clarke, P., Crawford, C., Steele, F., & Vignoles, A. (2010, June). *The choice between fixed and random effects models: some considerations for educational research*. DoQSS Working Paper No. 1010. London: Institute of Education, University of London. Retrieved from http://eprints.ncrm.ac.uk/1285/1/qsswp1010.pdf

Cogan, L. S., & Schmidt, W. (2015). The concept of opportunity to learn (OTL) in international comparisons of education. In K. Stacey & R. Turner (Eds), *Assessing Mathematical Literacy*. Switzerland: Springer International Publishing.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. Retrieved from http://www2.psych.ubc.ca/~schaller/528Readings/Cohen1992.pdf

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., et al. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher, 38*(2), 85-99.

Considine, G., & Zappala, G. (2002). Factors influencing the educational performance of students from disadvantaged backgrounds. In T. Eardley and B. Bradbury (Eds.), *Competing Visions: Refereed Proceedings of the National Social Policy Conference 2001* (pp. 91-107). SPRC Report 1/02. Sydney, NSW: Social Policy Research Centre, University of New South Wales. Retrieved from https://www.sprc.unsw.edu.au/media/SPRCFile/NSPC01_7_Considine_Zappala.pdf

Cooper, R. & Liou, D. D. (2007). The structure and cultural of information pathways: Rethinking opportunity to learn in urban high schools during the ninth grade transition. *The High School Journal*, *91*(1), 43-56. Retrieved from http://www.jstor.org/stable/40367922

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved from http://epaa.asu.edu/ojs/article/view/392/515

Darling-Hammond, L. (2011). *Quality teaching: What is it and how can it be measured?* Stanford Center for Opportunity policy in Education. Stanford Graduate School of Education. Retrieved from https://edpolicy.stanford.edu/sites/default/files/events/materials/ldhscopeteacher-effectiveness.pdf

De Boeck, P. & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer-Verlag.

Detik. (2013, Dec 12). *Mendikbud: Survei PISA makin memperkuat pentingnya Kurikulum 2013*. Retrieved from http://news.detik.com/wawancara/2439467/mendikbud-survei-pisa-makin-memperkuat-pentingnya-kurikulum-2013/1

Devlin, K. (1994). *Mathematics: The Science of Patterns: The Search for Order in Life, Mind and the Universe*. New York: W. H. Freeman Scientific American Library.

Else-quest, N., Hyde, J. S., & Linn, M. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103-127. doi: 10.1037/a0018053

Epstein, M. J., & Yuthas, K. (2012, Winter). Redefining education in the developing world. *Stanford Social Innovation Review*, 19-20. Retrieved from https://ssir.org/pdf/Winter_2012_Redefining_Education_in_the_Developing_World.pdf

Feuer, M. J. (2012). *No country left behind: Rhetoric and reality of international large-scale assessment*. Princeton, NJ: ETS. Retrieved from https://www.ets.org/research/policy_research_reports/publications/publication/2012/jgbx

Floden, R. E. (2002). The measurement of opportunity to learn. In A. Porter & A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washington D.C.: National Academy Press. Retrieved from http://www.nap.edu/catalog/10322.html

Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap? *High School Journal, 91*(1), 29-42. doi: 10.1353/hsj.2007.0022

Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education*, 60, 135-155.

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational evaluation and Policy analysis, 19*(4), 325-338.

Goe, L., & Stickler, L. M. (2008). *Teacher quality and student achievement*. Teaching Quality Research and Policy Brief. National Comprehensive Center for Teacher Quality. Retrieved from http://files.eric.ed.gov/fulltext/ED520769.pdf

Good, T.L., Wiley, C. R. H., & Flores, I. R. (2009). Effective teaching: an emerging synthesis. In L. J. Saha and A. G. Dworkins (Eds), *International Handbook of Research on Teachers and Teaching,* 803-816.

Grisay, A., deJong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseor, B. (2007). *Translation equivalence across PISA countries*. Journal of Applied Measurement, *8*(3), 249-266.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis, 17*(3), 323-336. doi: 10.3102/01623737017003323

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analysis relating to achievement.* New York, NY: Routledge.

Herman, J. L., & Klein, D. C. D. (1996). Evaluating equity in alternative assessment: An illustration of opportunity-to-learn issues. *The Journal of Educational Research, 89*(4), 246-256. doi: 10.1080/00220671.1996.9941209

Herman, J. L., & Klein, D. C. D. (1997). *Assessing opportunity to learn: A California example* (CSE Technical Report 453). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., Klein, D. C. D., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement Issues and Practice, 19*(4), 16-24.

Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries*, Vol. 2, New York, NY: Wiley.

IEA. (2011). About IEA. Retrieved from http://www.iea.nl/about_us.html

Jalal,F., Samani, M., Chang, M. C., Stevenson, R., Ragatz, A., Negara, S. D. (2009) . *Teacher certification in Indonesian: A strategy for teacher quality improvement*. Jakarta: Ministry of Education and Culture of Indonesia and World Bank. Retrieved from http://documents.worldbank.org/curated/en/705901468283513711/pdf/485780WP0Box331ication0in0Indonesia.pdf

Jimerson, S. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review*, 30(3), 420-437.

Kantor Staf Presiden. (2016, June 2). *Pemerintah perbanyak SMK dan tingkatkan kompetensi pelaku pendidikan kejuruan*. Retrieved from http://ksp.go.id/pemerintah-perbanyak-smk-dan-tingkatkan-kompetensi-pelaku-pendidikan-kejuruan/

Kompas. (2012, Jun 2). *Banyak siswa tak lulus ujian matematika*. Retrieved from http://edukasi.kompas.com/read/2012/06/02/10035432/Banyak.Siswa.Tak.Lulus.Ujian.Matematika

Kompas. (2013a, Mar 27). *Tolak kurikulum instan*. Retrieved from http://edukasi.kompas.com/read/2013/03/27/1538083/Tolak.Kurikulum.Instan?utm_source=WP&utm_medium=Ktpidx&utm_campaign=

Kompas. (2013b, Apr 9). *Apapun kurikulumnya, yang penting gurunya*. Retrieved from http://edukasi.kompas.com/read/2013/04/09/16065817/Apapun.Kurikulumnya.yang.Penting.Gurunya

Kompas. (2013c, May 24). *Angka kelulusan UN tahun ini 99,48 persen*. Retrieved from http://edukasi.kompas.com/read/2013/05/24/08265868/angka.kelulusan.un.tahun.ini.99.48.persen

Kompas. (2013d, Dec 5). *Skor PISA: Posisi Indonesia nyaris jadi juru kunci*. Retrieved from http://www.kopertis12.or.id/2013/12/05/skor-pisa-posisi-indonesia-nyaris-jadi-juru-kunci.html

Kompas. (2014, Dec 8). *Surat keputusan Mendikbud menghentikan Kurikulum 2013*. Retrieved from http://edukasi.kompas.com/read/2014/12/08/11583761/Surat.Keputusan.Mendikbud.Menghentikan.Kurikulum.2013.

Kompas. (2016a, Nov 30). *Pentingnya moratorium ujian nasional*. Retrieved from http://nasional.kompas.com/read/2016/11/30/19395491/pentingnya.moratorium.ujian.nasional

Kompas. (2016b, Dec 7). *Usulan moratorium ujian nasional ditolak*. Retrieved from http://nasional.kompas.com/read/2016/12/07/18333851/usulan.moratorium.ujian.nasional.ditolak

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environment Research, 9*, 231-251. DOI 10.1007/s10984-006-9015-7

Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. N. Elliott, A. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students. Bridging the gaps between research, practice, and policy* (pp. 99-130). New York, NY: Springer Science+Business Media.

Kurz, A., & Elliott, S. N. (2014). Assessing students' opportunity to learn the intended curriculum using an online teacher log: Initial validity evidence. *Educational Assessment, 19*, 159-184. doi: 10.1080/10627197.2014.934606

Lee, J. K. (2004). Evaluating the effectiveness of instructional resource allocatioin and use: IRT and HLM analysis of NAEP teacher survey and student assessment data. *Studies in Educational Evaluation, 30*, 175-199.

Lokheed, M. (2015). *Why do countries participate in international large-scale assessments? The case of PISA*. Policy Research Working Paper, 7447, World Bank. Retrieved from http://documents.worldbank.org/curated/en/692561468000587471/pdf/WPS7447.pdf

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Liu, O.L., Wilson, M., & Park, I. (2008). A multidimensional Rasch analysis of gender differences in PISA Mathematics. *Journal of Applied Measurement, 9*(1), 1-18.

Long, D. A. (2014). Cross-national educational inequalities and opportunities to learn: Conflicting views of instructional time. *Educational Policy, 28*(3), 351-392. doi: 10.1177/0895904812465108

Luschei, T. F., & Zubaidah, I. (2012). Teacher training and transitions in rural Indonesian schools: a case study of Bogor, West Java. *Asia Pacific Journal of Education, 32*(3), 333-350.

MacJessie-Mbewe, S. (2004). Rural communities-education relationship in developing countries: The case of Malawi. *International Education Journal, 5(*3),308-330. Retrieved from http://files.eric.ed.gov/fulltext/EJ903857.pdf

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 49-174.

Maul, A., Wilson, M., & Irribarra, D. T. (2013). On the conceptual foundations of psychological measurement. *Journal of Physics: Conference Series 459*. doi:10.1088/1742- 6596/459/012008

McCoy, A. R., & Reynolds, A. J. (1999). Grade retention and school performance: an extended investigation. *Journal of School Psychology*, 37(3), 273-298.

McDonnell, L. M. (1995). Opportunity to learn as research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305-322.

Media Indonesia. (2016, Mar 29). *Kemendikbud pacu peningkatan mutu SMK*. Retrieved from http://www.mediaindonesia.com/news/read/37058/kemendikbud-pacu-peningkatan-mutu-smk/2016-03-29

Ministry of Education and Culture of Indonesia (2012). *Uji publik Kurikulum 2013: Penyederhanaan, tematik-Integratif.* Retrieved from http://www.kemdiknas.go.id/kemdikbud/uji-publik-kurikulum-2013-1

Ministry of Education and Culture of Indonesia. (2013a, June 13). *Wawancara dengan Menteri Pendidikan Terkait Kurikulum 2013 (Bagian 3) (Interviews with the Minister of Education and Culture on the 2013 Curriculum (Part 3))*. Retrieved from http://litbang.kemdikbud.go.id/index.php/index-berita-kurikulum/230-wawancara-dengan-mendikbud-terkait-kurikulum-2013-bagian-3

Ministry of Education and Culture of Indonesia. (2013b). *Overview of the education sector in Indonesia 2012. Achievements and challenges*. Jakarta: Author.

Ministry of Education and Culture of Indonesia. (2014). *Implementation of 2006 Curriculum and 2013 Curriculum*. Minister Regulation, No. 160. Jakarta: Author.

Ministry of Education and Culture of Indonesia. (2015). *Kebijakan perubahan ujian nasional*. Jakarta: Author.

Ministry of Education and Culture of Indonesia. (2014, May 19). *Tingkat kelulusan ujian nasional SMA capai 99,52 persen.* Retrieved from http://litbang.kemdikbud.go.id/index.php/home2-9/859-tingkat-kelulusan-ujian-nasional-sma-capai-99-52-persen

Ministry of Education and Culture of Indonesia. (2015, Dec 15). *Pemerintah siapkan perangkat untuk wajib belajar 12 tahun*. Retrieved from http://www.kemdikbud.go.id/main/blog/2015/12/pemerintah-siapkan-perangkat-untuk-wajib-belajar-12-tahun-4930-4930-4930

Ministry of Education and Culture of Indonesia. (2016). *Tanya jawab tentang ujian nasional*. Retrieved from http://un.kemdikbud.go.id/

Mo, Y., Singh, K., & Chang, M. (2013). Opportunity to learn and student engagement: a HLM study on eight grade science achievement. *Educational Research Policy Practice, 12*, 3-19. doi: 10.1007/s10671-011-9126-5

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independent assumption. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 4*, Educational Testing Service and International Association for the Evaluation of Educational Achievement. Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_04_Chapter_6.pdf

Muthén, B., Huang, L., Jo, B., Khoo, S., Goff, G. N., Novak, J. R., & Shih, J. C. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational evaluation and Policy Analysis, 17*(3), 371-403.

National Research Council (NRC). (2001). Teaching for Mathematical proficiency. In J. Kilpatrick, J. Swafford, and B. Findell (Eds.), *Adding it up: Helping children learn* Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Association of School Psychologist (NASP). (n.d.). *Grade retention and social promotion*. White paper. Retrieved from https://www.nasponline.org/research-and-policy/professional-positions/white-papers

National Council of Teachers of Mathematics (NCTM). (2012). *Closing the opportunity gap in mathematics education.* Retrieved from http://www.nctm.org/Standards-and-Positions/Position-Statements/Closing-the-Opportunity-Gap-in-Mathematics-Education/

Newhouse, D. and Beegle, K. (2006). The effect of school type on academic achievement: Evidence from Indonesia. *The Journal of Human Resources, 41*(3), 529-557.

Niederle, M. & Vesterlund, L. (2010). Explaining the gender gap in Math test scores: The role of competition. *The Journal of Economic Perspectives, 24*(2), 129-144. doi=10.1257/jep.24.3.129

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. Santa Monica: The RAND Corporation.

O'Day, J., & Smith, M. S. (1993). Systemic reform and educational opportunity. In S. H. Fuhrman (Ed.), *Designing coherent educational policy* (pp. 250-312). San Francisco: Jossey-Bass.

Opportunity Gap (2016, August 1). In S. Abbott (Ed.), *The glossary of education reform*. Retrieved from http://edglossary.org/opportunity-gap/

Organization for Economic Cooperation and Development (OECD). (2002). *PISA 2000 Technical Report*. OECD Publishing. Retrieved from http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33688233.pdf

Organization for Economic Cooperation and Development (OECD). (2004). *PISA 2003. Learning for tomorrow's world*. OECD Publishing. Retrieved from https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/34002216.pdf

Organization for Economic Cooperation and Development (OECD). (2005). *PISA 2003 Technical Report*. OECD Publishing. Retrieved from https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/35188570.pdf

Organization for Economic Cooperation and Development (OECD). (2007). *PISA 2006. Science competencies for tomorrow's world*. OECD Publishing. Retrieved from https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/pisa2006results.htm

Organization for Economic Cooperation and Development (OECD). (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics, and science* (Vol. I). Retrieved from http://www.oecd.org/pisa/pisaproducts/48852548.pdf

Organization for Economic Cooperation and Development (OECD). (2012). *PISA 2009 Technical Report*. OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/50036771.pdf

Organization for Economic Cooperation and Development (OECD). (2009a). *Take the test. Sample questions from OECD's PISA assessments*. Retrieved from http://www.oecd.org/pisa/pisaproducts/Take%20the%20test%20e%20book.pdf

Organization for Economic Cooperation and Development (OECD). (2009b). *Creating Effective Teaching and Learning Environments: First Results from TALIS.* OECD Publishing. Retrieved from *http://dx.doi.org/10.1787/9789264068780-en*

Organization for Economic Cooperation and Development (OECD). (2010). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science* (Volume I). Retrieved from http://dx.doi.org/10.1787/9789264091450-en

Organization for Economic Cooperation and Development (OECD). (2013a). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I). PISA, OECD Publishing. Retrieved from *http://dx.doi.org/10.1787/9789264201118-en*

Organization for Economic Cooperation and Development (OECD). (2013b). *PISA 2012 Results: Ready to Learn: Students' Engagement, Drive and Self-Beliefs (Volume III).* OECD Publishing. Retrieved from *http://dx.doi.org/10.1787/9789264201170-en*

Organization for Economic Cooperation and Development (OECD). (2013c). *PISA 2012 Assessment and analytical framework: mathematics, reading, science, problem solving, and financial literacy*. OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264190511-en

Organization for Economic Cooperation and Development (OECD). (2013d). *PISA 2012 Results: Excellence through Equity: Giving every student the chance to succeed* (Volume II). PISA, OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264201132-en

Organization for Economic Cooperation and Development (OECD). (2013e). *PISA 2012 Results: What makes schools successful? Resources, policies, and practices* (Volume IV). PISA, OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264201156-en

Organization for Economic Cooperation and Development (OECD). (2014). *PISA 2012 Technical Report*. OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD – Asia Development Bank. (2015). *Education in Indonesia. Rising to the Challenge*. Paris, France: OECD Publishing. Retreived from http://dx.doi.org/10.1787/9789264230750-en

Organization for Economic Cooperation and Development (OECD). (2016a). *About the OECD*. Retrieved from http://www.oecd.org/about/

Organization for Economic Cooperation and Development (OECD). (2016b). *PISA 2015 Results. Excellence and equity in education* (Volume I). PISA, OECD Publishing. Retrieved from https://www.oecd.org/education/pisa-2015-results-volume-i-9789264266490-en.htm

Organization for Economic Cooperation and Development (OECD). (2016c). *Results from PISA 2015. Indonesia*. PISA, OECD Publishing. Retrieved from https://www.oecd.org/pisa/PISA-2015-Indonesia.pdf

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109-119.

Porter, A. (1995a). Defining and measuring opportunity to learn. In W.J. Fowler, Jr. (Ed.), *Developments in school finances* (Report No. NCES-95-706, pp. 51-78). Washington, DC: National Center for Education Statistics.

Porter, A. (1995b). The uses and misuses of opportunity-to-learn standards. *Educational Researcher, 24*(1), 21-27.

Pullin, D. C., & Haertel, E.H. (2008). Assessment through the lens of "Opportunity to Learn". In P.A. Moss, D.C. Pulin, J.P. Gee, E.H. Haertel (Eds.), *Assessment, equity, and opportunity to learn* (pp. 17-41). New York, NY: Cambridge University Press

Rowan, B. (2002). Large-scale, cross-national surveys of educational achievement: promises, pitfalls, and possibilities. In A. Porter & A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washington D.C.: National Academy Press. Retrieved from http://www.nap.edu/catalog/10322.html

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *Elementary School Journal, 105*(1), 103-127.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: lessons from the study of instructional improvement. *Educational Researcher, 38*, 120. doi: 10.3102/0013189X09332375

Rabe-Hesketh, S. & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata. Volume I: Continuous responses* (3rd. Ed.). College Station, TX: Stata Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago:University of Chicago Press.)

Republika. (2016, November 24). *Kemendikbud usulkan moratorium ujian nasional mulai 2017*. Retrieved from http://www.republika.co.id/berita/pendidikan/eduaction/16/11/24/oh4x7z382-kemendikbud-usulkan-moratorium-ujian-nasional-mulai-2017

Reeves, E. B. (2012). The effects of opportunity to learn, family socioeconomic status, and friends on the rural math achievement gap in high school. *American Behavioral Scientist, 56*(7), 887-907. doi: 0.1177/0002764212442357

Schwartz, R. A. (2012). *The development and psychometric modeling of an embedded assessment for a data modeling and statistical reasoning learning progression* (Doctoral Dissertation, UC Berkeley). Retrieved from: http://escholarship.org/uc/item/96j6w7xk

Schwartz, W. (1995). Opportunity to learn standards: Their impact on urban students. ERIC/CUE Digest Number 110, ERIC Clearinghouse on Urban Education. Retrieved from http://eric.ed.gov/?id=ED389816

Schmidt, W. & Maier, A. (2009). Opportunity to learn. In G. Sykes, B. L. Schneider, D. N. Plank, & T. G. Ford (Eds.), *Handbook of Education Policy Research* (pp. 541-559). New York: Routledge.

Schmidt, W., Zoido, P., & Cogan, L. (2013). Schooling matters: Opportunity to learn in PISA 2012. *OECD Education Working Papers, 95*, OECD Publishing. Retrieved from http://dx.doi.org/10.1787/5k3v0hldmchl-en

Schmidt, W. H., Burroughs, N., Zoido, P., & Houang, R. T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher, 44*(7), 371-386. doi: 10.3102/0013189X15603982

Silberglitt, B., Appleton, J., Burns, M. K., & Jimerson, S. R. (2006). Examining the effects of grade retention on student reading performance: A longitudinal study. *Journal of School Psychology, 44*, 255-270.

Sindo. (2014, Dec 26). *UN diganti evaluasi nasional*. Retrieved from http://nasional.sindonews.com/read/942347/149/un-diganti-evaluasi-nasional-1419567391

Steen, L. (1990). *On the Shoulders of Giants: New Approaches to Numeracy*. Washington, DC: National Academy Press.

Stevens, F. I., & Grymes, J. (1993). *Opportunity to learn: Issues of equity for poor and minority students*. Report No: NCES-93-232. Washington, DC: National Center for Education Statistics.

Stuhlman, M.W., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2015). *Part 2: What should classroom observation measure*. Charlotsville, VA: Center For Advanced Study of Teaching and Learning (CASTL). Retrieved from http://curry.virginia.edu/uploads/resourceLibrary/CASTL_practioner_Part2_single.pdf

Sundstrom, S. (2010). *Coding in multiple regression analysis: A review of popular coding techniques*. U.U.D.M. Project Report 2010:14, Uppsala Universitet. Retrieved from http://www.diva-portal.org/smash/get/diva2:325460/fulltext01.pdf

Suryadarma, D. & Jones, G. W. (Eds.). (2013). *Education in Indonesia*. Singapore: Institute of Southeast Asian Studies.

Tempo. (2012, Jun 3). 1330 siswa SMP tak lulus UN Matematika. Retrieved from https://nasional.tempo.co/read/news/2012/06/03/079408015/1-330-siswa-smp-tak-lulus-un-matematika

Tornroos, J. (2005). Mathematics textbooks, opportunity to learn, and student achievement. *Studies in Educational Evaluation, 31*, 315-327.

Trading Economics (2012). *Populasi Indonesia*. Retrieved from http://id.tradingeconomics.com/indonesia/population

UCLA. (n.d.). *Regression with Stata. Chapter 5 – Additional coding systems for categorical variables in regression analysis*. Statistical Consulting Group. Accessed on November 1, 2016. Retrieved from http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter5/statareg5.htm

UNESCO. (2000). Education for All Goals. Retrieved from http://www.unesco.org/new/en/education/themes/leading-the-international-agenda/education-for-all/efa-goals/

UNESCO. (2016a). Education for all movement. Retrieved from http://www.unesco.org/new/en/education/themes/leading-the-international-agenda/education-for-all/

UNESCO. (2016b). SDG 4 Education 2030. Retrieved from http://www.unesco.org/new/en/education/themes/leading-the-international-agenda/education-for-all/sdg4-education-2030/

United Nations. (2016). Sustainable Development Goals. 17 Goals To Transform Our World. Retrieved from http://www.un.org/sustainabledevelopment/education/

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20, 137-155*. doi:10.3102/015237370/01623637020003137

Wihardini, D. (2012). *Multidimensional Rasch analysis on gender, grade and booklet effects of PISA 2009 Mathematics: A case of Indonesian students* (Research Report). Berkeley, CA: UC Berkeley Evaluation and Assessment Research (BEAR) Center.

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

World Bank. (2010). Inside Indonesia's mathematics classrooms: a TIMSS video study of teaching practices and student achievement. Retrieved from http://documents.worldbank.org/curated/en/151301468283518230/Inside-Indonesias-mathematics-classrooms-a-TIMSS-video-study-of-teaching-practices-and-student-achievement

Wu, M. (2003). The impact of PISA on mathematics education: Linking mathematics and the real world. Education Journal, 31(2), 121-140.

Wu, M.L., Adams, R.J., Wilson, M.R., and Haldane, S.A. (2007). *ACER Conquest version 2.0. Generalised item response modeling software*. *Technical Manual*. Camberwell, Vic: ACER Press.

Wu, M.L., Adams, R.J., Wilson, M.R., and Haldane, S.A. (2016, January). *ACER Conquest version 4.0. Generalised item response modeling software*. *ConQuest Command Reference.* Camberwell, Vic: ACER Press. Retrieved from https://www.acer.edu.au/conquest/acer-conquest1