

Boston College
Lynch School of Education and Human Development

Department of
Measurement, Evaluation, Statistics, and Assessment

FROM OLS TO MULTILEVEL MULTIDIMENSIONAL MIXTURE IRT:
A MODEL REFINEMENT APPROACH TO INVESTIGATING
PATTERNS OF RELATIONSHIPS IN PISA 2012 DATA

Dissertation
by

GULSAH GURKAN

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

August 2021

FROM OLS TO MULTILEVEL MULTIDIMENSIONAL MIXTURE IRT:
A MODEL REFINEMENT APPROACH TO INVESTIGATING
PATTERNS OF RELATIONSHIPS IN PISA 2012 DATA

by

Gulsah Gurkan, Author

Dr. Henry I. Braun, Dissertation Chair

ABSTRACT

Secondary analyses of international large-scale assessments (ILSA) commonly characterize relationships between variables of interest using correlations. However, the accuracy of correlation estimates is impaired by artefacts such as measurement error and clustering. Despite advancements in methodology, conventional correlation estimates or statistical models not addressing this problem are still commonly used when analyzing ILSA data.

This dissertation examines the impact of both the clustered nature of the data and heterogeneous measurement error on the correlations reported between background data and proficiency scales across countries participating in ILSA. In this regard, the operating characteristics of competing modeling techniques are explored by means of applications to data from PISA 2012. Specifically, the estimates of correlations between math self-efficacy and math achievement across countries are the principal focus of this study. Sequentially employing four different statistical techniques, a step-wise model refinement approach is used. After each step, the changes in the within-country correlation estimates are examined in relation to (i) the heterogeneity of distributions, (ii) the amount of measurement error, (iii) the degree of clustering, and (iv) country-level math performance.

The results show that correlation estimates gathered from two-dimensional IRT models are more similar across countries in comparison to conventional and multilevel linear modeling estimates. The strength of the relationship between math proficiency and math self-efficacy is moderated by country mean math proficiency and this was found to be consistent across all four models even when measurement error and clustering were taken into account. Multilevel multidimensional mixture IRT modeling results support the hypothesis that low-performing groups within countries have a lower correlation between math self-efficacy and math proficiency. A weaker association between math self-efficacy and math proficiency in lower achieving groups is consistently seen across countries. A multilevel mixture IRT modeling approach sheds light on how this pattern emerges from greater randomness in the responses of lower performing groups. The findings from this study demonstrate that advanced modeling techniques not only are more appropriate given the characteristics of the data, but also provide greater insight about the patterns of relationships across countries.

ACKNOWLEDGEMENTS

I would like to thank my dissertation chair and advisor Dr. Henry Braun for his continued mentorship, patience, and confidence in me throughout this work. It has been an absolute privilege working by his side for many years and I am, and forever will be, indebted to him for making me a better statistician and researcher. I also would like to thank Dr. Matthias von Davier, who co-advised and helped shape this dissertation from the very beginning. He generously offered his expertise, whenever I needed it, with encouragement and kindness. His guidance was invaluable particularly with some of the techniques that were central to this study. Finally, I would like to thank my readers, Dr. Zhushan Mandy Li and Dr. Michael Martin, for their thoughtful feedback that greatly improved the quality of this dissertation.

My doctoral studies have been even more rewarding thanks to many wonderful colleagues and friends I have made along the way. Special thanks to my dear friends Romita Mitra and Michael Kelly for their continuous cheering and never letting me believe that I was alone in this journey.

Words fail me to express my deepest gratitude to my parents, Emine and Orhan Gurkan, who have made my education as well as my sisters' their top priority. Any accomplishment I have achieved is in great part thanks to their unconditional love and support. I am truly grateful to them and my sisters for being only a call away regardless of the physical distance between us.

I owe the most thanks to my dearest husband, and my lifelong friend, Murat Kilicoglu. Pursuing my PhD degree required so much energy and so many sacrifices. I could not have endured any of these without his unwavering support and love. Thank you for always supporting me to pursue my dreams and believing in me undoubtedly.

Table of Contents

Symbols and Acronyms	iii
List of Tables	vii
List of Figures	ix
Chapter 1: Introduction	1
1.1 Problem Overview	1
1.2 Purpose of the Study	6
Research Questions	9
1.3 Significance of the Study	11
Chapter 2: Theoretical Background	12
2.1 ILSA Study Design and Data Properties	12
Clustered Nature of the Data	12
Measurement Error	15
2.2 Existing Advances in Modeling Data with Measurement Error and Clustering	19
2.2.1 Ordinary Linear Regression Model with Adjustments	19
2.2.2 Latent Variable Models	24
Item Response Theory	24
Structural Equation Modeling	28
Mixture Item Response Theory Models	31
2.2.3 Multilevel Models	35
Multilevel Linear Models	36
Multilevel Latent Variable Models	38
2.3 Relationship in Focus: Math Self-efficacy and Math Achievement	47
The PISA 2012 Assessment	47
Math Assessment	49
Math Self-efficacy	50
Domain-specific Self-related Constructs	52
Chapter 3: Methodology	55
3.1 Data	55

Sample.....	56
Variables	57
Missing Data	59
Weighting.....	60
3.2 Analysis Overview.....	61
Phase 1: Conventional Analysis of Within-Country Relationships	64
Phase 2: Refinement with Multidimensional IRT Modeling.....	67
Phase 3: Refinement with Multilevel Linear Modeling.....	71
Phase 4: Refinement with Multilevel Multidimensional Mixture IRT Modeling	74
Phase 5: Refinement at the School-level Within-countries	80
Phase 6: Refinement with a Composite Background Variable.....	84
Chapter 4: Results.....	86
4.1 Phase 1: Conventional Analysis of Within-Country Relationships.....	86
4.2 Phase 2: Refinement with Multidimensional IRT Modeling.....	90
4.3 Phase 3: Refinement with Multilevel Linear Modeling.....	105
4.4 Phase 4: Refinement with Multilevel Multidimensional Mixture IRT Modeling	110
4.5 Phase 5: Refinement at the School-level Within-countries	129
4.6 Phase 6: Refinement with a Composite Background Variable.....	147
Chapter 5: Discussion	161
5.1 Summary of Findings.....	161
5.2 Limitations and Future Research	177
5.3 Final Remarks	180
References.....	183
Appendix.....	212

Symbols and Acronyms

Acronyms

AIC:	Akaike information criterion
ANOVA:	Analysis of variance
ANXMAT:	Math anxiety index
BIB:	Balanced incomplete block
BIC:	Bayesian information criterion
CFA:	Confirmatory factor analysis
CTT:	Classical test theory
DIF:	Differential item functioning
EAP:	Expected a posteriori
EIV:	Errors-in-variables
EM:	Expectation-maximization
GDM:	General diagnostic model
GLLAMM:	Generalized linear latent and mixed models
GPCM:	Generalized partial credit model
GRM:	Graded response model
HGLM:	Hierarchical generalized linear model
HMGDM:	Hierarchical mixture general diagnostic models
ICC:	Intra-class correlation
IEA:	International Association for the Evaluation of Educational Achievement
ILSA:	International large-scale assessment
IQR:	Interquartile range
IRT:	Item response theory
LCA:	Latent class analysis
LL:	Log Likelihood
MAP:	Maximum a posteriori
MAR:	Missing at random

MATHEFF: Math self-efficacy index

MCMC: Markov Chain Monte Carlo

MESE: Mixed effects structural equations

MGDM: Mixture general diagnostic model

MIRT: Multidimensional item response theory

MixIRT: Mixture item response theory

ML: Maximum likelihood

MLE: Maximum likelihood estimator

MLIRT: Multilevel item response theory

MLM: Multilevel linear modeling

MLM_composite: Multilevel linear model with the composite measure

MLM_matheff: Multilevel linear model with the math self-efficacy variable

MLMIRT: Multilevel multidimensional item response theory

MLMixMIRT: Multilevel multidimensional mixture item response theory

MLMixMIRT_High: Multilevel multidimensional mixture item response theory - high math class

MLMixMIRT_Low: Multilevel multidimensional mixture item response theory - low math class

MML: Marginal maximum likelihood

MSA: Measure of sampling adequacy

NMAR: Not missing at random

NRM: Nominal response model

OECD: Organisation for Economic Co-operation and Development

OLS: Ordinary least squares

OLS_composite: Ordinary least squares model with the composite measure

OLS_matheff: Ordinary least squares model with the math self-efficacy variable

PCA: Principal component analysis

PCM: Partial credit model

PIRLS: Progress in International Reading Literacy Study

PISA: Programme for International Student Assessment

1PL: One-parameter logistic

- 2PL: Two-parameter logistic
- PPS: Probability proportional to size
- PV: Plausible value
- RQ: Research question
- RSM: Rating scale model
- SCMAT: Math self-concept index
- SD: Standard deviation
- SEM: Structural equation modeling
- SRS: Simple random sampling
- TIMSS: Trends in International Mathematics and Science Study
- VG: Validity generalization
- WLE: Weighted likelihood estimate

Symbols used in Chapter 3

r : Correlation estimate

Rubin's pooling method:

U: sampling variance

B: imputation variance

V: total variance

OLS models:

b : standardized regression coefficient

e_i : residual term

MIRT models:

$\vec{\theta}$: multidimensional continuous latent trait

$\vec{\alpha}_i$: multidimensional item discrimination/slope parameter

β_{ik} : item intercept parameter

x_i : item response for item i

k : an item response category of a polytomous item

m_i : total number of categories of a polytomous item

\vec{x} : item response pattern

$\hat{\sigma}^2$: square of the estimated measurement error

MLM:

β_{0j} : individual-level intercept parameter

β_{1j} : individual-level slope parameter

r_{ij} : individual-level residual term

γ_{00} : group-level intercept parameter

γ_{10} : group-level slope parameter

u_{0j} : group-level residual term

τ_{00} : variance between groups

σ^2 : within-group variance

MLMixMIRT models:

q_{ik} : binary entries in a Q-matrix indicating whether the item i measures the k^{th} dimension of the construct

g : latent class grouping indicator

s : observed clustering variable

π_g : class size or mixing proportion

β_{ihg} : class-specific item intercept parameters

$\vec{\alpha}_{ig}$: class-specific item slope parameters

List of Tables

Table 3.1. Items measuring math self-efficacy	57
Table 3.2. Items measuring math anxiety	58
Table 3.3. Items measuring math self-concept	58
Table 3.4. Model refinement steps.....	61
Table 4.1. Descriptive statistics for the within-country conventional correlation estimates from OLS models	87
Table 4.2. Model fit indices for the MIRT models.....	91
Table 4.3. Descriptive statistics for within-country correlation estimates from MIRT models ...	92
Table 4.4. Descriptive statistics for within-country correlation estimates from MLM models..	107
Table 4.5. Descriptive statistics for the within-country mean EAP math ability estimates by latent classes	111
Table 4.6. Descriptive statistics for within-country correlation estimates from MLMixMIRT models.....	113
Table 4.7. Pearson correlation between within-country correlation estimates and country-level math performance	117
Table 4.8. Descriptive statistics for the within-super-school conventional correlation estimates from OLS models.....	130
Table 4.9. Descriptive statistics for the weighted averages of PV1MATH for super-schools by country	132
Table 4.10. Descriptive statistics for the within-super-school correlation estimates from MIRT models.....	135
Table 4.11. Descriptive statistics for the weighted averages of MATHEFF for super-schools by country	146
Table 4.12. Descriptive statistics for the within-country conventional correlation estimates from OLS models with the composite measure and with math self-efficacy measure alone.....	150
Table 4.13. Descriptive statistics for the within-country correlation estimates from MLM models with the composite measure and with math self-efficacy measure alone.....	156
Table A.1. Country names and abbreviations.....	212
Table A.2. Total and final sample sizes by country.....	213
Table A.3. Q-Matrix used in MIRT and MLMixMIRT models.....	216

Table A.4. Sample size and mean PV1MATH range by super-school and country (Phase 5)...	218
Table A.5. Sample size by country (Phase 6)	220
Table A.6. Correlation estimates from the OLS models employed in Phase 1	221
Table A.7. Descriptive statistics for the item response data employed in Phase 2 and Phase 4.	223
Table A.8. Estimated item parameters from the multi-group MIRT model employed in Phase 2	228
Table A.9. Correlation estimates and empirical reliabilities from the MIRT models employed in Phase 2	233
Table A.10. Parameter estimates from the MLM models employed in Phase 3	236
Table A.11. Correlation estimates, class sizes, and class means from the MLMixMIRT models employed in Phase 4	239
Table A.12. Empirical reliabilities for the High Math Class from the MLMixMIRT models employed in Phase 4	242
Table A.13. Empirical reliabilities for the Low Math Class from the MLMixMIRT models employed in Phase 4	245
Table A.14. Correlation estimates by super-school and country from the OLS models employed in Phase 5	248
Table A.15. Correlation estimates and empirical reliabilities by super-school from the MIRT models employed in Phase 5.....	250
Table A.16. Results from the PCA analysis conducted in Phase 6.....	252
Table A.17. Correlation estimates from the OLS models employed in Phase 6	254
Table A.18. Parameter estimates from the MLM models employing MATHEFF only in Phase 6	256
Table A.19. Parameter estimates from the MLM models employing the composite measure in Phase 6	259

List of Figures

Figure 2.1. A simple path analysis example	28
Figure 2.2. A full SEM example.....	30
Figure 2.3. A Multilevel SEM example.....	43
Figure 3.1. Illustration of MLMixMIRT modeling process for one country in a unidimensional setting.....	77
Figure 4.1. Distribution of the within-country conventional correlations	87
Figure 4.2. Relationship between country-mean math performance and math self-efficacy	88
Figure 4.3. Relationship between within-country conventional correlations and country-mean math proficiencies	89
Figure 4.4. Relationship between within-country conventional correlations and country-mean math proficiencies	90
Figure 4.5. Distribution of the within-country conventional correlations and MIRT correlations	93
Figure 4.6. Relationship between within-country MIRT correlations and country-mean math proficiencies	95
Figure 4.7. EAP reliabilities gathered from the MIRT models for the math scale	96
Figure 4.8. EAP reliabilities gathered from the MIRT models for the math self-efficacy scale ..	97
Figure 4.9. Relationship between EAP reliabilities for math proficiency from MIRT models and the changes in the correlation estimates from conventional to MIRT models	98
Figure 4.10. Distribution of the EAP reliabilities gathered from the MIRT models for the math scale by country math performance level	99
Figure 4.11. Relationship between empirical reliability indices for math proficiency from MIRT models and country-mean math proficiencies	100
Figure 4.12. Distribution of the EAP reliabilities gathered from the MIRT models for the math self-efficacy scale by country math performance level	101
Figure 4.13. Relationship between empirical reliability indices for math self-efficacy from MIRT models and country-mean math proficiencies	102
Figure 4.14. Relationship between empirical reliability indices for math self-efficacy from MIRT models and country-mean math self-efficacy indices.....	103
Figure 4.15. Relationship between empirical reliability indices for math from MIRT models and MIRT correlation estimates	104

Figure 4.16. Relationship between empirical reliability indices for math self-efficacy from MIRT models and MIRT correlation estimates	105
Figure 4.17. Distribution of the within-country ICCs for the math performance outcome	106
Figure 4.18. Distribution of the within-country conventional and MLM correlations	108
Figure 4.19. Relationship between within-country MLM correlations and country-mean math proficiencies	109
Figure 4.20. Relationship between changes in the correlation estimates from OLS to MLM and within-country intra-class correlations for math proficiency outcome	110
Figure 4.21. Distribution of within-country mean math performance by latent class	112
Figure 4.22. Distribution of the within-country conventional and MLMixMIRT correlations ..	114
Figure 4.23. Relationship between within-country MLMixMIRT correlations and country-mean math performance	116
Figure 4.24. Relationship between within-country MLMixMIRT correlations and class-mean math performance	118
Figure 4.25. Distribution of within-country EAP reliabilities for math scale from the MLMixMIRT models by latent class	119
Figure 4.26. Distribution of within-country EAP reliabilities for math self-efficacy scale from the MLMixMIRT models by latent class	120
Figure 4.27. Relationship between within-country math reliabilities from the MLMixMIRT models and country-mean math performance	121
Figure 4.28. Relationship between within-country math reliabilities from the MLMixMIRT models and class-mean math performance	122
Figure 4.29. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and country-mean math performance	123
Figure 4.30. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and class-mean math performance	124
Figure 4.31. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and country-mean math self-efficacy index	125
Figure 4.32. Distribution of within-country mean math self-efficacy by latent class	126
Figure 4.33. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and class-mean math self-efficacy	127

Figure 4.34. Relationship between within-country math reliabilities from the MLMixMIRT models and MLMixMIRT correlation estimates	128
Figure 4.35. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and MLMixMIRT correlation estimates.....	129
Figure 4.36. Distribution of within-super-school conventional correlations by country.....	131
Figure 4.37. Relationships between conventional correlations and mean math proficiencies ...	133
Figure 4.38. Distribution of within-super-school conventional correlations and MIRT correlations.....	136
Figure 4.39. Relationships between MIRT correlations and mean math proficiencies	138
Figure 4.40. Distributions of EAP reliabilities gathered from the MIRT models for the math proficiency scale	139
Figure 4.41. Relationship between empirical reliability indices for math proficiency and mean math performance	141
Figure 4.42. Relationship between empirical reliability indices for math proficiency and MIRT correlation estimates	142
Figure 4.43. Distributions of EAP reliabilities gathered from the MIRT models for the math self-efficacy scale.....	143
Figure 4.44. Relationship between empirical reliability indices for math self-efficacy and mean math performance	145
Figure 4.45. Relationship between empirical reliability indices for math self-efficacy and mean math self-efficacy.....	147
Figure 4.46. Distribution of proportion of variance explained by the first principal component	149
Figure 4.47. Distribution of the within-country conventional correlations employing the composite measure and the math self-efficacy measure alone	150
Figure 4.48. Comparison of the relationships between OLS correlation estimates and country-mean math performance.....	152
Figure 4.49. Distribution of the within-country OLS and MLM estimates employing the math self-efficacy measure alone.....	154
Figure 4.50. Comparison of the relationships between OLS_matheff and MLM_matheff estimates and country-mean math performance	155

Figure 4.51. Distribution of the within-country MLM estimates employing the composite measure and the math self-efficacy measure alone.....	156
Figure 4.52. Comparison of the relationships between MLM correlation estimates and country-mean math performance.....	157
Figure 4.53. Distribution of the within-country OLS and MLM estimates employing the composite measure.....	158
Figure 4.54. Comparison of the relationships between correlation estimates and country-mean math performance	159
Figure 4.55. Relationship between changes in the correlation estimates from OLS to MLM and within-country intra-class correlations for math proficiency outcome.....	160
Figure A.1. PISA 2012 Student Questionnaire Rotated Design	215

Chapter 1: Introduction

1.1 Problem Overview

International comparative studies of student achievement such as the IEA's Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), and the OECD's Programme for International Student Assessment (PISA) receive a great deal of attention by researchers and policy makers across the world (Rutkowski & Rutkowski, 2010). Beside their many other uses, the main argument in favor of international large-scale assessments (ILSAs) is that countries can better understand the strengths and weaknesses of their education systems when they are compared to the education systems in other countries (Porter & Gamoran, 2002). ILSAs provide rich databases to examine students' cognitive skills as well as the relationships of those skills with numerous background variables, such as students' socioeconomic backgrounds and attitudes related to their learning and experiences in schools. For many years, the results from ILSAs have informed education policies in many ways such as pedagogy, teacher training, and hours of instruction (Heyneman & Lee, 2014). On the other hand, the secondary users of ILSA data are cautioned regarding the limitations of ILSAs and potential misinterpretations of the results from ILSA data analysis (Lockheed & Wagemaker, 2013; Singer & Braun, 2018).

Although educational assessment focused mostly on measuring students' cognitive skills in certain subject matters for quite some time, it has been well-accepted in the field that numerous contextual factors play important roles in educational practices and outcomes (Rutkowski & Rutkowski, 2010). Therefore, researchers have been careful about taking objective characteristics such as gender, socio-economic background, school characteristics as well as non-

cognitive factors such as students' attitudes and beliefs into account in the course of interpreting students' cognitive performances and making inferences to inform educational policies and practices. Thus, there have been substantial efforts to improve background questionnaires in ILSAs and to enrich the databases to allow for the examination of the relationships between students' cognitive skills and many other non-cognitive background measures (Porter & Gamoran, 2002). For instance, PISA distinguishes between minor and major assessment domains and includes additional context questionnaires specific to the major domain for any given cycle. In 2015, the major domain was science and the context questionnaire contained not only domain-specific measures for students such as enjoyment of science and environmental awareness, but also for teachers and school administrators such measures as science teacher collaboration and school learning environment for science. Although there are domain-independent measures that are used in every PISA cycle to make system-level trend data available (OECD, 2017), additional domain-specific measures are intended to capture in-depth information that is related to the major domain of the cycle. In PISA technical reports, the purpose of the background instruments is stated as helping policy makers and educators understand the reasons behind different levels of student achievement (OECD, 2014). Nevertheless, secondary analysts and policy makers are advised not to directly link differences in educational practices to differences in student performance (Singer & Braun, 2018).

To support inferences regarding educational outcomes and their relationships to other factors, it is critical for ILSAs to ensure cross-cultural comparability of the measures and assure the validity and the reliability of the data that is collected. In this regard, the development of both cognitive and non-cognitive measures is aided by enlisting committees of subject matter experts, developing assessment frameworks, conducting field studies, and using statistical modeling

approaches that provide evidence for the psychometric quality and construct validity of the measures. That said, ensuring cross-cultural comparability for non-cognitive measures in cross-national studies still requires attention and improvement (He & van de Vijver, 2015; Rutkowski & Rutkowski, 2010).

For instance, a recurring concern related to the background questionnaires in ILSAs is about a peculiar but consistent finding regarding certain non-cognitive measures and their relationships to achievement, commonly referred to as the “attitude-achievement paradox”. It has been shown for certain non-cognitive measures that averages of students’ attitudes are negatively correlated with average cognitive performance at the between-country level, even though the correlations are positive at the within-country level (Buckley, 2009; Kyllonen et al., 2010; Van de gaer et al., 2012). There have been many studies and potential explanations proposed regarding these correlations being of opposite signs. Although many promising approaches and explanations were suggested by researchers, a widely accepted explanation for this paradox remains to be found (Kyllonen & Bertling, 2014; Stankov et al., 2018).

It is known that data aggregation is associated with the problem of *ecological fallacy*, which is the tendency to draw conclusions about individual-level correlations based on aggregate-level (or ecological) correlations (Robinson, 1950). In ILSAs, schools are selected within countries and the data is collected from random samples of the population of eligible students within selected schools. The expectation that country-level correlations are in the same direction as individual-level correlations in ILSA data is an example of an ecological fallacy. On the other hand, many researchers are not convinced that this is the only – or even the best – explanation for the attitude-achievement paradox. It is argued that the reversal of direction demands more explanation than just having to deal with spurious correlations even though

aggregating data tends to change the magnitude of relationships seen at the within-country level (Kyllonen & Bertling, 2014; Stankov et al., 2018). Furthermore, the reversal of direction may be better explained by *Simpson's paradox* (Simpson, 1951), which suggests that a strong association observed within groups can disappear or reverse in direction due to a confounding variable when the groups are combined. Stankov et al. (2018) extended their study on the construct validity of the anchored scales used in PISA 2012 and briefly explored Simpson's paradox in relation to the attitude-achievement paradox. However, they were not able to find a satisfactory interpretation of the results.

An alternative explanation that deserves consideration lies in the reliability and cross-cultural comparability of non-cognitive measures. The attitudinal surveys in ILSAs typically employ response formats based on self-reported rating scales. Self-reports employing Likert-type scales are the main source of data available in ILSAs on students' affective states and behavioral dispositions. In other words, our knowledge about students' beliefs and attitudes is limited to what students choose to disclose about themselves by responding to probes using these rating scales. Self-reports using Likert scales are known to be vulnerable to factors such as social desirability or a tendency to extreme response style that influence respondents' objectivity in representing themselves (Khorramdel & von Davier, 2014). In addition, due to the concerns with test length and fatigue effects, the number of items that are used to measure each non-cognitive construct is often limited to between four and eight items. This limits the amount of information that can be obtained. These are all significant limitations of Likert-type background measures in the ILSA context, and ensuring reliability, validity, and cross-cultural comparability becomes a challenge under such conditions (Kyllonen & Bertling, 2014; Kyllonen et al., 2010).

Furthermore, underlying differences among groups can result in different amounts and distributions of uncertainty in the non-cognitive measures and this can threaten the cross-cultural comparability of the measurement. That is, among countries or among regions within a country there may be modest but non-trivial differences in how students' response styles or their understanding of the questions introduce both construct-irrelevant variance and greater measurement error. For instance, low scale reliability for certain developing countries raises concerns regarding the comparability of non-cognitive measures (Rutkowski & Rutkowski, 2010). The heterogeneity in the amount of measurement error across countries may explain some of the large differences among within-country correlations of attitudinal measures and achievement.

Response style differences and reference group effects have been widely examined in the field regarding the attitude-achievement paradox and the construct validity of the non-cognitive background measures (Heine et al., 2002; Kyllonen & Bertling, 2014; Min et al., 2016). It is argued that self-report rating scales suffer from individuals' tendencies to endorse extreme categories in items regardless of what the items are designed to measure (Baumgartner & Steenkamp, 2001; Ju & Falk, 2019; Khorramdel & von Davier, 2014; Lu & Bolt, 2015; Min et al., 2016). That is, respondents can favor the lower or upper-level response categories in an item no matter the statement in the item. This is arguably an important limitation of Likert-type scales. With regard to the cross-national studies, it has been shown that cultural differences in response style can especially lead to misleading results and should not be ignored (Van de gaer et al., 2012).

Although response bias could be a part of the reason behind the heterogeneous uncertainty in background measures across countries, the results from studies that take the

response bias into account while analyzing data from ILSAs are mixed (Ju & Falk, 2019). For instance, the relationships between extreme response style and other constructs have been found to be relatively small at the within-country level (Bolt & Newton, 2011). However, von Davier (2018) found that (lower) proficiency measures are consistently associated with membership in those subpopulations that exhibit extreme response styles. Additionally, studies that utilize some adjustments based on participants' response style did not result in substantial changes in the findings (He & van de Vijver, 2015; Lu & Bolt, 2015). Approaches such as anchoring vignettes (King et al., 2004) that were used in PISA 2012 for certain non-cognitive measures have been criticized in terms of leading to a significant change in the factorial structure of the constructs and introducing statistical artefacts (Stankov et al., 2018; von Davier, Shin, et al., 2018).

In light of unresolved issues regarding the substantial heterogeneity observed in the correlations across countries, it can be argued that the impact of measurement error, as well as the nested structure of the ILSA data, on the statistical inferences should be explored further. Although a single explanation may not be found regarding the attitude-achievement paradox, statistical inferences should account for the fact that measurement errors tend to be heterogeneous across countries and appropriate modeling approaches should be employed when analyzing data from ILSAs.

1.2 Purpose of the Study

The purpose of the study is to investigate, in a particular instance, the variation among within-country correlations based on conventional estimates gathered from analyzing international large-scale assessment data and to determine the extent to which this variation is

due to the heterogeneity in the amount of measurement error, as well as the clustered nature of the data.

In educational research, it is very common to sample schools or classrooms of students to collect data and ILSAs are no different. Even though it is more desirable to design studies with students being the unit of observation, it is neither practical nor cost effective to randomly select students without selecting multiple students from classrooms and schools for a study. Moreover, studies may rely on teacher information as well, and in order to link teachers to students, the most straightforward sampling approach would be to select classrooms taught by certain teachers, and then to sample students from within these classes. However, cluster sampling brings some challenges for secondary analysts. Different from studies in which each observation is assumed to contribute unique information, participants drawn from a particular cluster share the same context, so that the total amount of information is less than what would be obtained from a random sample of the same size. The impact of clustering on the power of the study can be quantified in terms of a design effect (Snijders, 2005). Clustering also limits the use of certain statistical modeling techniques such as standard regression analysis with ordinary least square estimation due to violations of certain independence assumptions (Raudenbush, 1993). The importance of separating within-group and between-group effects has been well recognized (Cronbach & Webb, 1975; Raudenbush, 1993). Therefore, appropriate statistical approaches such as multilevel linear modeling should be utilized while analyzing ILSA data.

Furthermore, the reliability of self-reported rating scales measuring non-cognitive skills in ILSAs is arguably still below desired levels (Rutkowski & Rutkowski, 2010). Measurement of latent constructs is at the heart of social sciences and the consequences of fallible measurement have to be acknowledged and understood better in order to avoid potentially misleading

inferences. With regard to the background measures in ILSAs, research regarding cultural differences in response style is extensive and well-known. Further, reference group effects on the construct validity can certainly be a part of the explanation for aforementioned attitude-achievement anomaly (Van de gaer et al., 2012). However, there may be other reasons for varying levels of uncertainty in the self-report measures. For instance, certain groups of students may have difficulty understanding the questions asked in background questionnaires due to low reading ability and/or unfamiliarity with the types of questions. This may result in heterogeneity of measurement error among and within countries. Additionally, students with low cognitive skills may tend to skip questions or to give the same response to every question without reading the question stems (i.e., straightlining behavior). This introduces construct irrelevant variance and, if not accounted for, leads to increased errors of measurement and to problems with student-level background measures, especially in lower-achieving countries. Rutkowski and Rutkowski (2010) show that student-level background measures tend to display problems in certain less economically developed countries.

It is well documented that the strength of the reported relationships between many background scales and cognitive measures differs across countries and, in some cases, these differences can be quite large. For example, students' math self-concept can be a significant predictor of their math achievement in Portugal ($r=0.64$) while this relationship is rather weak in Colombia ($r=0.15$) (OECD, 2014). However, this heterogeneity across countries may result from the imperfect measurement and the limitations of self-report attitudinal rating scales, and not be due to differences in the correlations of the underlying constructs of math literacy and math self-concept. When measurement error is not accounted for properly, having an apparently weaker relationship in one country and an apparently stronger relationship in another country may not be

due to the real differences in the magnitude of the relationships but, rather, due to the measure itself and differences in its precision across population. Therefore, statistical modeling techniques that simultaneously take into account measurement error and clustering should provide different and more credible results than those based on statistical analyses in which heteroscedastic nature of the measurement error is ignored (Fox, 2007; Kamata, 2001; Pastor, 2003; Sulis & Toland, 2017). In fact, with techniques that properly account for the uncertainty in measures, the results may yield a much more homogeneous picture regarding the distribution of within-country attitude-achievement relationships. In such a case, the results should offer more accurate country-level correlations for background measures with proficiency and may help to address the attitude-achievement paradox.

To summarize, the main goal of the study is to examine the impact of both the clustered nature of the data and the heterogeneous measurement error on the correlations reported between background data and proficiency scales across countries participating in ILSAs. In this regard, the operating characteristics of competing modeling techniques are explored by means of applications to data from PISA 2012 and the estimates of correlations between math self-efficacy and math achievement across countries are the focus of this study.

Research Questions

Based on the need for a better understanding of the impact of measurement error and clustering on the correlation estimates gathered from ILSA data, five main research questions have been formulated. The first three examine different modeling techniques and their performance in properly accounting for measurement error and clustering compared to conventional correlation estimates of the relationship between math self-efficacy and math achievement. Although within-country relationships and their distribution across the countries

are the main focus of the study, the fourth research question explores the distribution of the correlation estimates at the level of schools within a country to investigate whether similar patterns are seen at the level of schools and at the level of countries. Finally, the fifth research question examines whether the accuracy of the correlation estimates can be improved by combining multiple background measures. To this end, a composite measure that combines math self-efficacy, math anxiety, and math self-concept is used as the background measure in the models.

Research questions are formulated as follow:

1. When modeling techniques are used to properly account for *measurement error* in the observed data, do the estimates of within-country relationships between math self-efficacy and math achievement display greater homogeneity across countries than the conventional, within-country correlation estimates?
2. When the *clustered nature* of the observed data is taken into account, do the estimates of within-country relationships between math self-efficacy and math achievement have greater homogeneity across countries than displayed by the conventional correlation estimates?
3. When modeling techniques are used to properly account for *both measurement error and clustering* in the observed data, do the estimates of within-country relationships between math self-efficacy and math achievement show greater homogeneity across countries than that displayed by the conventional correlation estimates or the ones gathered from the models that account for either measurement error or clustering only?
4. Are the relationship patterns in school-level correlations and school-level proficiencies within countries similar to the patterns seen at the country-level?

5. When participants' attitudinal background measure comprises *multiple indicators* (math-self-efficacy, math anxiety, and math self-concept), how different are the changes seen in the correlation estimates when clustering is taken into account compared to those seen when math self-efficacy is used as a single background measure?

1.3 Significance of the Study

Conventional correlations are commonly calculated to characterize relationships between variables of interest in secondary analyses of ILSA data. However, the accuracy of correlation estimates can be impaired by statistical artefacts such as measurement error and clustering (Carroll et al., 2006; Fuller, 1987; Jones, 1979; Woodhouse et al., 1996). As noted earlier, there have been numerous advancements over the years in statistical modeling techniques to properly take both measurement error and the clustered nature of the data into account. Nevertheless, conventional correlation estimates (e.g., Pearson correlations) or statistical models which can only treat part of the problem (e.g., multilevel linear models) are still commonly used when analyzing data gathered from complex study designs such as ILSAs.

This study carries out a comprehensive investigation of the impact of measurement error and clustering on statistical inferences by employing data from the PISA 2012 database. In particular, unresolved issues regarding the heterogeneity in relationships between certain background and cognitive measures seen across the countries are explored. With the proposed step-wise model refinement strategy, it is possible to provide comparisons among available statistical modeling techniques and insights about how large-scale assessment data can be better analyzed.

Chapter 2: Theoretical Background

2.1 ILSA Study Design and Data Properties

International large-scale assessments have increased in number and participation as they have become the main resources for researchers and policy makers who are interested in educational accountability and monitoring at the system level (Rutkowski et al., 2014). With the help of many advancements in their methodology and design, ILSA data facilitate certain comparisons across countries in terms of differences among education systems and their impact on educational outcomes. Although they provide useful evidence regarding the school systems of participating jurisdictions, one should keep in mind that ILSAs are cross-sectional studies that are limited in the kinds of inferences they can generally support. That is, patterns of relationships revealed in ILSA studies are cross-sectional and observational in nature and, generally, are not suitable for making causal inferences. In fact, crude analysis of ILSA data and summary information such as rankings can be often misleading and lead to unwarranted policy changes (Goldstein & Spiegelhalter, 1996; Singer & Braun, 2018). The complex ILSA study designs permit flexible and affordable data collection at a large scale but require attention to data quality and mandate employing appropriate procedures in the data analysis (Rust, 2014).

Clustered Nature of the Data

Collecting comparable information about students' performance at a large-scale demands complex study designs and sampling approaches. Because the main goal of ILSAs is to make inferences at the population level, data are collected from representative samples of students within participating countries without requiring every student to participate in the assessment. Stratified sampling techniques are used in ILSAs to obtain representative samples of students

from participating countries. In PISA, for example, a two-stage stratified sampling design involves first sampling schools that have eligible 15-year-old students and then selecting a random sample of students from these randomly sampled schools (OECD, 2014).

A multi-stage sampling technique has consequences for the amount of information gathered from a sample. Students who attend the same school share a context that makes them more similar (in contrast to students who attend different schools) in terms of their background and educational outcomes such as what they learn and how motivated they are to learn. One consequence of this within-context correlation is that the information gleaned from each observation is not as unique as that obtained from a single observation in a simple random sampling (SRS) design. In terms of relative precision, the advantage of a SRS over a complex design with the same sample size is termed the *design effect*. The design effect can be calculated by accounting for the intra class correlation (ICC or ρ) which is a measure of the degree of dependence among the observations in terms of the outcome of interest (Kish, 1965), as shown below.

$$ICC = \frac{\text{Variance between schools } (\sigma^2_{\text{between}})}{\text{Total variance } (\sigma^2_{\text{between}} + \sigma^2_{\text{within}})} \quad (2.1)$$

$$\begin{aligned} \text{Design Effect} &= \frac{\text{Variance of the statistic under the actual design}}{\text{Variance of the statistic with SRS design}} \quad (2.2) \\ &= 1 + (n - 1) * ICC , \end{aligned}$$

where n represents the average number of individuals across schools. Effective sample size can also be calculated based on the ICC and the design effect. In studies with correlated observations, effective sample size is smaller than the actual sample size.

$$\text{Effective sample size} = \frac{N}{\text{Design effect}} = \frac{N}{1 + (n - 1) * ICC}, \quad (2.3)$$

where N represents the actual total sample size.

Secondary analysts should keep in mind that ILSA data are collected at the individual level and not at the school or country level. That is, data are collected from the students even though the goal is making inferences at the group level. The use of individual level data at the aggregate level may trigger an ecological fallacy. As mentioned earlier, ecological fallacy (or aggregation bias) occurs when the statistical inferences made at the group level are generalized to the individual level without enough support (Diez-Roux, 1998; Robinson, 1950; Schwartz, 1994; Wakefield, 2008). Therefore, to avoid aggregation bias, a trend observed at the aggregate level should not be expected to follow the trend observed at the individual level. For instance, a surprisingly powerful correlation between chocolate intake per capita and the number of Nobel laureates can be seen when the data is based on country-level averages (Messerli, 2012). It is unjustifiable, however, to conclude that chocolate consumption enhances cognitive function and increases the number of Nobel laureates in a country. Robinson (1950) demonstrates that group level (or ecological) correlations cannot validly be used as substitutes for individual-level correlations and can lead to meaningless or faulty conclusions. Many similar examples of spurious and unwarrantable correlations are demonstrated by Vigen (2019).

Measurement Error

Cognitive Measures

It is a challenge for ILSAs to collect sufficient data to achieve adequate construct representation while keeping individual testing time short enough to avoid fatigue effects, especially in a low-stakes context (von Davier & Sinharay, 2014). A comprehensive representation of the focal skill domains requires a large number of items that tap different facets of the construct, as well as varying in level of difficulty. In order to avoid hours of testing time per student, a matrix sampling (item) design is adopted by which each item is administered to only a random sample of respondents so that each student is administered only a portion of all items measuring the construct (Braun & von Davier, 2017). In order to ensure that each pair of items appears together a predetermined number of times, a balanced incomplete block (BIB) design is used in the development of the test forms. In BIB designs, items are organized in blocks and blocks are systematically rotated in the construction of each test booklet. What makes this block design ‘incomplete’ is that each form contains only some of the item blocks, and not all blocks appear together with all other blocks.

The complex item sampling design used in ILSAs has certain implications for achievement scaling methodology. Due to the matrix sampling, individual-level measurement is not as accurate as it could be when all items were taken by every student in the sample. Compared to testing programs yielding scores for individual students, ILSAs aim at group-level score reporting only and, thus, produce proficiency estimates that contain a comparably larger amount of measurement error at the student level (von Davier, Gonzalez, & Mislevy, 2009). Nonetheless, a set of intermediate individual-level performance values, termed plausible values, are generated by latent regression modeling which incorporates both students’ responses to the

items administered and the student background information collected by the context questionnaires (Braun & von Davier, 2017). Plausible values are randomly drawn from an empirically derived family of (posterior) distributions of score values, specific to each individual. The use of background information enhances the precision of sub-population achievement values (Mislevy, 1991; von Davier, Gonzalez, & Mislevy, 2009). Plausible values can be used to generate approximately unbiased estimates of sub-population distributional characteristics, as well as estimates of the variance in those estimates due to measurement error (Martin et al., 2016). However, plausible values should not be treated as (usual) test scores for individuals and they are intended to be used as a set (at least 5 imputed values are provided in ILSA databases) by following appropriate procedures as described in the literature (Little & Rubin, 2002; Mislevy, 1991; von Davier, Gonzalez, & Mislevy, 2009).

With matrix sampling, each student is administered only a portion of the item pool: different students may answer different sets of questions, thus limiting the use of traditional number-correct scoring (von Davier, Gonzalez, & Mislevy, 2009). Consequently, ILSAs designers employ item response theory (IRT; Lord & Novick, 1968) in the development and scaling of the measures. IRT, also known as latent trait theory, estimates individuals' ability levels on an underlying trait based on both their responses and the psychometric properties of items (Emberson & Reise, 2000). In the IRT framework, item parameters are assumed to be (approximately) sample independent. Hence, ability scores can be reported on the same scale even though students are not administered the same set of items. Although item parameters are assumed to be invariant over populations, this assumption can be tested to assure whether reported scores are indeed comparable across different groups (de Ayala, 2009).

Background Questionnaire and Self-reports

Over the years, there have been many studies demonstrating strong correlations between students' achievement and their self-reported attitudes and behaviors. For example, measures such as self-concept and self-efficacy have been shown to be strongly associated with student achievement, even more strongly than many other measures of attitudes towards school and learning (Lee, 2009; Lee & Stankov, 2013). Recent studies even suggest that the development of character skills (or soft skills) should be incorporated in school practices because they are related to broader life outcomes and quality of life (Kautz et al., 2014; Levin, 2012). As these variables and their inter-relationships are important for education policy, information about students' demographic characteristics, attitudes and school context, and their relationships to cognitive skills have become an important focus in ILSAs. As a result, background questionnaires have been expanded over time (Jude & Kuger, 2018).

It is necessary to ensure the psychometric quality of the component non-cognitive measures not only because it is important to have a better understanding of students' attitudinal backgrounds and learning context and their relationships to the cognitive outcomes, but also because background measures are used in the generation of plausible values. Therefore, problems with the background measures can influence both the measurement of the learning outcomes and the observed relationships between those learning outcomes and background measures (Rutkowski & Rutkowski, 2010).

The use of plausible values appropriately enables the estimation of measurement error in cognitive outcomes in ILSAs (Braun & von Davier, 2017) but the extent and impact of measurement error in background measures is still unresolved. In addition, construct validity and cross-cultural comparability of non-cognitive measures have become a major concern in recent

years. One of the reasons for these concerns originates from the frequent use of Likert scales in the measurement of non-cognitive skills. Likert scales are judged to be problematic for making cross-country comparisons (Kyllonen & Bertling, 2014). It has been shown that simple rating scales can introduce construct-irrelevant variance because of differences in response styles and non-equivalence across different languages and cultures (Buckley, 2009; He & van de Vijver, 2015). Even though there are studies which provide evidence of response bias issues within countries and languages (Yap et al., 2014), the concerns regarding response-style differences are mostly discussed in the context of cross-country comparisons (Jude & Kuger, 2018). Many studies also suggest that response style differences may be the reason behind the contradictory results regarding the negative correlations between attitudinal measures and achievement at the country level; namely, the attitude-achievement paradox (Buckley, 2009; He & van de Vijver, 2015; Kyllonen et al., 2010).

Another concern relates to self-report measures in general. Self-reports are shown to be vulnerable to certain moderating factors. For instance, Marsh and Parker (1984) demonstrated that the level of socio-economic status and academic performance of the school are strongly related to students' self-concept. This phenomenon, namely the *big-fish-little-pond effect*, suggests that being in a low-ability/low-performing school can increase students' self-concept, especially if they perform above the (low) average of their peers. Furthermore, a study by Hopfenbeck and Maul (2011) discussed another source of incomparability, as they presented evidence that students with lower performances may not provide reflective responses to the questionnaires, so that the relationships between their non-cognitive skills and academic performances may not represent the true strength of the relationships. This may also result from what is known as the Dunning-Kruger effect in the field of psychology. The Dunning-Kruger

effect asserts that people who lack metacognition (or self-monitoring) may not be aware of their lack of skills and thus may overestimate their abilities (Dunning, 2011; Kruger & Dunning, 1999). While response styles can be considered as systematic errors that introduce bias into comparisons, studies that imply a lack of understanding or metacognition may play a role in reducing the quality of self-reports do suggest that unsystematic, random error components may be increased in these cases.

Even though there appears to be a consensus in the literature that construct validity and cross-cultural measurement equivalence cannot be assumed but only be empirically supported, most of the statistical procedures usually make the assumption of measurement invariance (Gustafsson, 2018; van de Vijver & He, 2016). When measurement invariance does not hold and the amount and the distribution of the measurement error differ across groups, the relationships between certain factors and the outcome of interest (at the country level) may appear more heterogeneous than is actually the case.

2.2 Existing Advances in Modeling Data with Measurement Error and Clustering

In this section, a selection of candidate models that properly account for measurement error and/or clustering is described. On the basis of the review of these approaches, the analytical techniques that were conducted in this study were chosen to be most suitable for the characteristics of the data to be analyzed, as well as for the research questions. These are described in more detail in Chapter 3.

2.2.1 Ordinary Linear Regression Model with Adjustments

Ordinary linear regression analysis is commonly used to investigate relationships among variables of interest in cross-sectional studies. However, it has been extensively demonstrated

that errors in the covariates (x) as well as in the criterion variable (y) in a linear regression model lead to incorrect statistical inferences (Carroll et al., 2006; Fuller, 1987; Jones, 1979; Woodhouse et al., 1996). Measurement error that is large and unaccounted for has been shown to attenuate regression coefficients and suggest weaker estimated relationships than the true strength of the relationships.

Errors in the covariates (x) have received relatively more attention in the literature. This is because errors in y and errors in x are treated differently in a linear regression model and, as a result, their impact on the estimated regression coefficients differs accordingly (Berkson, 1950). In a univariate regression, suppose x , the predictor of y , is measured without error. Suppose further that y is measured with error (y^*). The noise in y^* becomes a part of the residual term e in the regression equation and is minimized by the fitted regression line. Therefore, even though the observed relationship between y and x weakens due to the uncertainty in y^* , the estimated regression coefficient is not biased and has the smallest standard error among all linear unbiased estimators according to the Gauss-Markov theorem (Lewis-Beck et al., 2004).

$$y^* = \beta(x) + e \quad (2.4)$$

Now suppose the observed predictor is a measure of the true value of x with error d . In this case, the measurement error d in the predictor blends into both the predicted value of the criterion variable (\hat{y}^*) and the residual term (\hat{e}). As a result, the fitted regression line becomes biased no matter if y is measured with error (y^*) or without error (y) and the estimated regression coefficient ($\hat{\beta}$) is biased toward zero in comparison to the true regression coefficient (β) (Berkson, 1950).

$$\hat{y}^* = \hat{\beta}(x + d) + \hat{\epsilon} \quad (2.5)$$

Note above, measurement errors in both x and y are assumed to be linearly additive. In the field of psychometrics, this is fundamental to Classical Test Theory (CTT) which assumes that the observed score is a linear function of the true score and random error (Lord & Novick, 1968). When the variance of errors is conditional on the true values of x and y respectively, which is a more realistic assumption in many situations, the uncertainty cannot be handled with standard methodology (Buonaccorsi, 1996). Moreover, in multiple regression with more than one independent variable with error, the problems are more difficult to address because the estimated regression coefficients can become biased in unpredictable directions (Carroll et al., 2006).

Errors-in-variables Models (to adjust for measurement error)

Errors-in-variables (EIV) models, which have been mostly utilized in the econometrics literature, offer strategies for both linear and non-linear regression problems to account for the measurement error in independent variables (Anderson, 1984; Fuller, 1987). Classical EIV models assume that the measurement error is additive and its variance is constant (Carroll et al., 2006) similar to the standard linear regression methodology. There are two types of classical EIV models; functional and structural. In functional modeling, the independent variables are assumed to be fixed and non-stochastic. That is, no specific assumptions are made about the distributional structure of unobserved (or true) variables and the distribution of the predictor is not modelled parametrically (Battauz et al., 2011; Carroll et al., 2006). In structural modeling, on the other hand, the independent variables are assumed to be random and the distribution of the true predictor is modelled parametrically by employing likelihood estimation procedures (Anderson, 1984; Carroll et al., 2006).

EIV models are not without limitations. For one, the assumption of the measurement error having constant variance is known to be limiting and not realistic in many cases. Advanced statistical techniques (e.g., Item Response Theory Models) are capable of treating heteroscedastic measurement error properly. Secondly, although they are more flexible in comparison to functional models, structural models raise a concern in regard to the assumptions made about the distribution of the random independent variables (Carroll et al., 2006). Furthermore, Carroll et al. (2006) emphasize that correcting for measurement error bias often increases the variance of the estimated coefficients and leads to wider confidence intervals.

Validity Generalization (to adjust for measurement error)

Hunter and Schmidt (2004) show in their work on validity generalization (VG) that methodological and statistical artefacts across different studies, such as the error of measurement and sampling error, can alter the estimates of the correlations. They use the word “artefacts” to refer to the imperfections in studies that are not properties of nature but artefactual or man-made (Hunter & Schmidt, 2004). In particular, Hunter and Schmidt argue that heterogeneity in the observed correlations gathered from different situations increases as the reliability of the measure decreases. As a solution, they propose corrections for unreliability of the measurement instruments while conducting meta-analysis (Hunter & Schmidt, 2004). Schmidt and Hunter’s VG approach is seen as an attempt to integrate meta-analytic procedures with psychometric techniques to deal with measurement error (DeShon, 2003). On the other hand, such measurement error corrections arguably have many difficulties and limitations that are discussed extensively in Murphy’s book (2003) that offers a critical review on VG. One of the main arguments behind the criticism of VG is that the psychometric theory behind it is CTT (DeShon, 2003). Similar to EIV models, in VG, one way to correct the correlation estimates for attenuation

due to measurement error is to use the reliability information which is assumed to be constant along the ability scale in the CTT framework and a limited way to treat the error of measurement (DeShon, 2003).

Dummy Variables (to adjust for clustering)

Clustering in the data gathered from cross-sectional studies such as ILSAs can be taken into account in an ordinary regression model by including a set of dummy variables as the indicators of the groups. For example, suppose one is interested in examining the relationship between x and y in a dataset gathered from five countries with sufficient sample sizes. Suppose further that between-country variance explains a considerable amount of the variance in y which has to be taken into account. Four dummy variables can be added to the ordinary regression equation (Equation 2.6) and differences in country means can be examined in terms of the degree to which they differ from the reference country mean.

$$y = \beta_1(x) + \beta_2(\text{country 2}) + \beta_3(\text{country 3}) + \beta_4(\text{country 4}) + \beta_5(\text{country 5}) + e \quad (2.6)$$

One of the disadvantages of this approach is that the number of dummy variables and parameters increase substantially as the number of clusters increases. For example, if one were to examine a relationship within a country using a large dataset such as from an ILSA database and take the clustering by schools into account, it would lead to an increase in the number of parameters to be estimated by more than a hundred. Although it may be convenient with a smaller number of clusters, the increase in the number of parameters can become particularly problematic with more complex models and smaller within-group sample sizes. Moreover, ordinary regression models with dummy variables for groups are structured in a way that the

regression coefficients for groups are treated as fixed effects, which is why these models are sometimes called “fixed effects regression models” (Allison, 2009). Hence, any random effect resulting from group characteristics is confounded in the dummy variables and cannot be estimated separately (Allison, 2009).

2.2.2 Latent Variable Models

Item Response Theory

One of the major advances in the field of psychometrics, Item Response Theory (IRT) is a probabilistic modeling framework in which a person’s ability is assumed to be an unobserved latent variable and is estimated based on both individuals’ responses to the items in an instrument, as well as item characteristics such as difficulty and discrimination (Lord & Novick, 1968). In the IRT framework, measurement error is quantifiable and is not forced to be constant along the scale. Additionally, under certain conditions, IRT-based person ability estimates are independent of the set of items that were administered, which allows response data gathered from incomplete designs (e.g., matrix sampled cognitive measures used in ILSAs) to be used to estimate person abilities on a common scale (Fox & Glas, 2001). Many other features and advantages of IRT are extensively discussed in the literature (de Ayala, 2009; Edelen & Reeve, 2007).

In the usual IRT framework, the probability of observing a response pattern (\vec{x}) conditional on the individual’s underlying latent trait θ is the product of the conditional probabilities of observing responses to each item given θ . That is,

$$P(\vec{x}|\theta) = \prod_{i=1}^I p_i(x = x_i|\theta) \tag{2.7}$$

There are various IRT models that have been developed with different specifications. The simplest model for dichotomous responses, the one parameter logistic (1PL) model, assumes that items are characterized by their difficulty (also called ‘location’) only. This model can be used when the underlying latent trait is assumed to be continuous and unidimensional (de Ayala, 2009). In particular, for such models (Rasch, 1960), item and person parameters can be conditioned on observed total scores which are the sufficient statistics (Fisher, 1922). In IRT models, the non-linear relationship of an unobserved latent variable and the probability of a response is usually modelled by using a logistic regression model. For example, in 1PL IRT models, the conditional probability of a correct response given individual’s underlying unidimensional latent trait θ to a dichotomous item i ($x \in \{0, 1\}$) with difficulty parameter δ_i is specified as follows.

$$p_i(x = 1|\theta, \delta_i) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)} \quad (2.8)$$

Depending on the application context, such as instrument characteristics and model-data fit considerations (de Ayala, 2009), extended IRT models that involve other item parameters (e.g., discrimination, lower-asymptote, and/or upper- asymptote) can be used. In the two-parameter logistic (2PL) model (Birnbbaum, 1968), the conditional probability of a correct response to a dichotomous item with difficulty parameter δ_i and discrimination parameter α_i is defined as follows.

$$p_i(x = 1|\theta, \alpha_i, \delta_i) = \frac{\exp[\alpha_i(\theta - \delta_i)]}{1 + \exp[\alpha_i(\theta - \delta_i)]} \quad (2.9)$$

For polytomous items, models have been generalized depending on whether item response options are assumed to be ordered or unordered using ordinal/nominal logistic regression models (de Ayala, 2009). Common models that can be employed for polytomous responses are Partial Credit Model (PCM; Masters, 1982), Rating Scale Model (RSM; Andrich, 1978), Generalized Partial Credit Model (GPCM; Muraki, 1997), Graded Response Model (GRM; Samejima, 1997), and Nominal Response Model (NRM; Bock, 1972). For example, GPCM is simply an extension to the 2PL IRT model for polytomous items ($x \in \{0, \dots, m_i\}$) and probability of selecting k^{th} category over the previous category is defined as follows.

$$p_i(x = k|\theta, \alpha_i, \delta_{ik}) = \frac{\exp [\sum_{h=1}^k \alpha_i(\theta - \delta_{ih})]}{1 + \sum_{y=1}^{m_i} \exp [\sum_{h=1}^y \alpha_i(\theta - \delta_{ih})]}, \quad (2.10)$$

where δ_{ih} is the difficulty parameter of the transition from the $(h - 1)^{th}$ category to h^{th} category and m_i is the total number of categories. The GPCM is flexible because it can be applied to both dichotomous and polytomous data with varying numbers of categories.

Although these models are based on the assumption that the underlying latent trait is unidimensional, this may not be the reality in some situations. For instance, it would be unrealistic to expect a student to perform well in a science test without being proficient in reading. In other words, students' answers to a science problem in a test presumably depend on their proficiency both in science and in reading. For situations in which it is hypothesized that there is more than one underlying latent trait that influences respondents' answers to the items, a multidimensional IRT (MIRT) model would be more suitable (de Ayala, 2009). Similar to the case with unidimensional IRT models, various MIRT models have been developed with certain specifications in the estimation procedures (de Ayala, 2009; Reckase, 1997; van der Linden,

2016). MIRT models also allow investigating the correlations among multiple latent traits by taking the measurement error in the response data into account (Reckase, 2009).

As the person ability estimate θ is extended to incorporate multiple dimensions, a reparametrization of the model definition is helpful. With multiple dimensions, θ as well as the item discrimination parameter α are no longer scalars but vectors, which does not allow the subtraction of θ from the scalar location parameters as defined in the component $\alpha_i(\theta - \delta_i)$. Therefore, a reparametrization as $\alpha_i\theta - \alpha_i\delta_i$, in which the interaction term $-\alpha_i\delta_i$ is a scalar, allows for convenience in the definition (Reckase, 2009). The interaction term is usually called the intercept parameter (β_i) and the reparameterized version of the model is usually called “the slope/intercept form”. As an example, a multidimensional extension of the GPCM in the slope/intercept form is written as follows. Note that the sign of the intercept is reversed.

$$p_i(x = k | \vec{\theta}, \vec{\alpha}_i, \beta_{ik}) = \frac{\exp [k\vec{\alpha}_i\vec{\theta} - \sum_{h=1}^k \beta_{ih}]}{1 + \sum_{y=1}^{m_i} \exp [y\vec{\alpha}_i\vec{\theta} - \sum_{h=1}^y \beta_{ih}]} \quad (2.11)$$

Many of the MIRT models including the multidimensional extension of GPCM presented above are compensatory models because a low ability on one dimension does not necessarily imply a low probability of correct response (Reckase, 2009). That is, a low ability on one dimension can be compensated by the ability on the other dimensions and the individual can still respond to the item correctly. Non-compensatory models, on the other hand, require a certain level of ability on all dimensions. Estimates of the parameters of non-compensatory models are likely to be unstable due to this dependency on adjacent latent traits (Chalmers et al., 2016).

Structural Equation Modeling

Often, modeling measurement of the latent variables is in itself the main interest especially when a new scale is being developed. More often, the relationships among latent variables are of importance particularly in studies exploring causality. In the 1920s, Sewall Wright took the first step towards modeling structural dependencies among a set of variables with his work on path analysis (Matsueda, 2011; Schumacker & Lomax, 2016). In path analysis, direct and indirect relationships between independent variables (labelled ‘exogenous’ in this framework) and the dependent variables (named ‘endogenous’ in this framework) can be explored as shown in the example below.

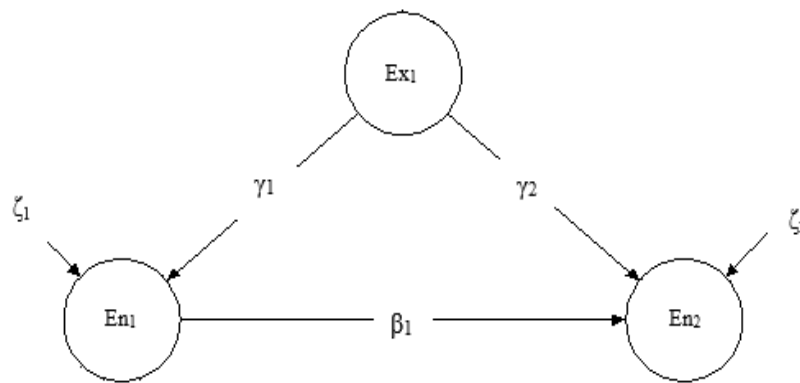


Figure 2.1. A simple path analysis example

The path diagram above demonstrates structural relationships among one exogenous variable Ex_1 and two endogenous variables En_1 and En_2 . The parameters γ_1 and γ_2 represent the regression coefficients predicting the two endogenous variables with the exogenous variable, and β_1 is the regression coefficient representing the relationship between the two endogenous variables in this model. Note that, β_1 represents both the direct effect of En_1 on En_2 and the indirect effect of Ex_1 on En_2 through En_1 . Although the term ‘in/direct effect’ is used in the SEM

framework to distinguish separate dependencies among variables, it does not necessarily indicate causal relationships but correlational relationships. The parameters ζ_1 and ζ_2 represent the errors in each dependent variable.

It may be helpful to consider path analysis as a single-indicator model in which each latent variable is measured by only one observed (also called ‘manifest’) variable. When there are many manifest variables measuring latent variables in a model, measurement models can be added to the model and both the latent variables and the structural relationships among them can be examined simultaneously within the same modeling framework. This framework is termed Structural Equation Modeling (SEM). In the SEM framework, the measurement part of the model (i.e., Confirmatory Factor Analysis; CFA) is usually defined by using linear links between observed variables and underlying latent variables, while the structural part allows for estimating the relationships between latent variables. CFA was fully developed by Karl Jöreskog in the 1960s (Schumacker & Lomax, 2016). The combination of CFA and path analysis into a single, coherent framework (i.e., SEM) was based on the work of Jöreskog (1973), Keesling (1972), and Wiley (1973). Although there are certain differences between a CFA model and an IRT model (Reise et al., 1993), a specific form of IRT model is obtained when link function that connects the expected value of the latent variable with the manifest variables in a linear manner is the logit link (Rabe-Hesketh, Skrondal, & Pickles, 2004).

For example, building on the path analysis example given previously (Figure 2.1), manifest variables measuring independent and dependent variables can be added to the model as shown in Figure 2.2. In this example, there are 11 manifest and 3 latent variables in the model. Although there are a total of 31 parameters to be estimated in this model, there are many other parameters that could have been included. For example, the error terms that are attached to the

observed variables x_1 , x_2 , and x_3 could have been allowed to be correlated. In practical applications, model specifications should be made as the model is defined and modified based on underlying hypotheses and model-data fit. The structural part framed in the center of the path diagram is the same as the previous path analysis example (Figure 2.1) and it is combined with measurement models for the three latent variables in the SEM example. The links 1 through 8 represent the factor loadings. Although it is not mandatory, the first manifest variable linked to each latent variable is fixed in this specified model and, hence, not estimated. Error terms are attached to each manifest variable, shown with arrows numbered 18 through 28, and can be allowed to be correlated as shown with springs numbered 12 through 17.

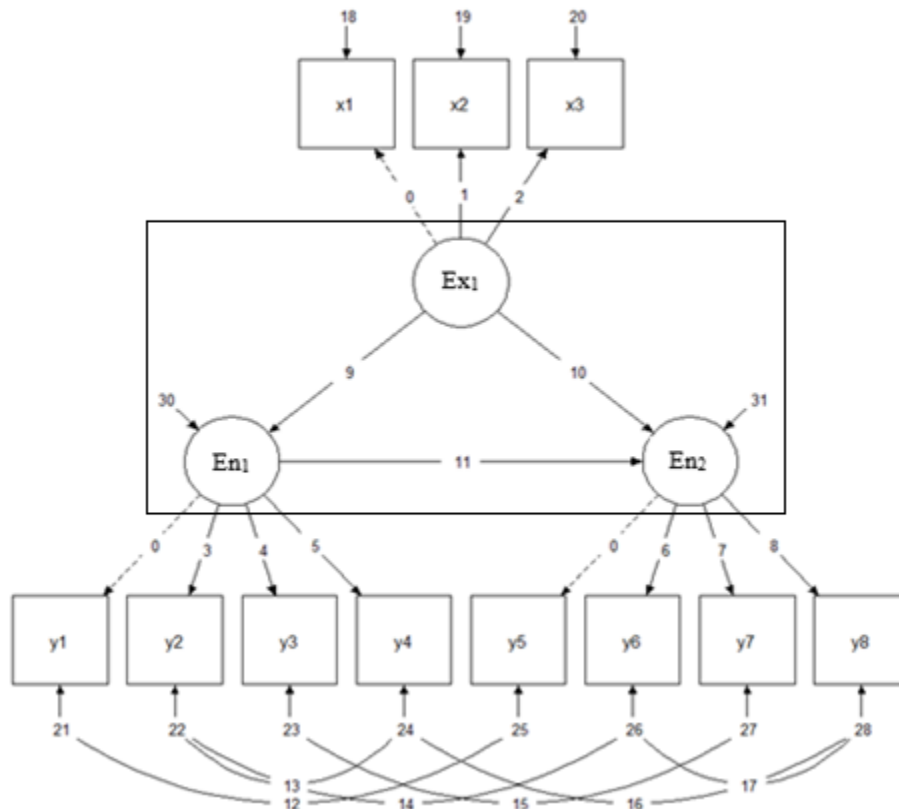


Figure 2.2. A full SEM example

Mixture Item Response Theory Models

In estimating the parameters of standard IRT models, it is assumed that all the individuals in the sample belong to one population; that is, the same item parameters apply to all the respondents. This assumption can be challenged under certain conditions. For instance, there may well be consistent variations in the responses across subgroups of students who solve problems employing different strategies and cognitive processes in a particular testing setting. In order to properly take into account such subgroups within a population, a new line of research that combines Latent Class Analysis (LCA) with IRT models arose towards the end of 1980s (Rost, 1990).

In classical Latent Class Analysis (LCA), it is assumed that the underlying latent trait has a discrete distribution and not a continuous distribution as in the IRT framework. Therefore, individuals' responses are modeled with a discrete latent variable (i.e., latent class variable) that is represented by the absence or presence of each of a set of skills or traits, based on their (unobserved) membership in one class of a family of mutually exclusive classes (Lazarsfeld & Henry, 1968; von Davier, 2009). Even though LCA does not assume an ordering of the latent classes, located latent class models can be used when it is assumed that the underlying latent trait comprises an ordered set of distinct classes each with a fixed latent ability level (von Davier, 2009; von Davier & Yamamoto, 2004). This definition of the ordered latent classes is somewhat similar to the definition of latent traits in the IRT framework as a monotonic increase along the ability levels is assumed in both approaches (von Davier, 2009; von Davier & Yamamoto, 2004).

The first examples of models incorporating LCA into IRT combined the Rasch model with LCA by employing mixture distributions (Mislevy & Verhelst, 1990; Rost, 1990). Mixture distributions allow responses of the individuals from different unknown subpopulations to be

modelled by different item parameter sets (von Davier, 2009; von Davier & Yamamoto, 2004). In the Mixed Rasch Model (Rost, 1990; von Davier & Rost, 2016), the latent structure is defined by a continuous latent variable θ which represents the individual's underlying latent trait and a discrete latent variable g which represents the individual's underlying latent class. In comparison to standard Rasch model, item difficulty parameter (δ) is specific to each latent class (g) in the mixture-distribution Rasch model and the conditional probability of a correct response to a dichotomous item given the individual's underlying continuous latent trait θ and latent class g is written as follows.

$$P_i(x_i = 1|\theta, g) = \sum_g \pi_g \frac{\exp(\theta - \delta_{ig})}{1 + \exp(\theta - \delta_{ig})}, \quad (2.12)$$

where π_g is the proportion of the latent class within the population (also called mixing proportion or class size).

Note that the underlying latent class variable is unknown in mixture-distribution IRT (MixIRT) models in contrast to multi-group IRT models in which the grouping variable is observed (von Davier & Rost, 2007). That is, MixIRT models recognize and account for groups of individuals with systematically different item responses within a population although it is unknown to which group an individual belongs to and what exactly leads to those systematically different responses. Mixture models split individuals into latent classes based on maximizing within-class homogeneity. This approach has been shown to be helpful in many settings ranging from identifying personality styles to differential item functioning (DIF) (Sen & Cohen, 2019). Mixed Rasch Models have also been extended to incorporate polytomous item responses (Rost & von Davier, 1995a) and the GPCM has been extended with mixture distributions (von Davier & Yamamoto, 2004).

Similar to the standard IRT framework, multidimensionality has been considered in the MixIRT framework. It has been shown that multidimensionality can affect the number of latent classes extracted from Mixed Rasch Models and the resulting latent classes can be misleading if multidimensionality is not accounted for (Jang et al., 2018). There have been many studies exploring multidimensional extensions of MixIRT models for different purposes. For instance, De Boeck et al. (2011) employed MixIRT models with a secondary dimension to explain DIF between latent classes. Rost and von Davier (1995b) employed MixIRT models to test the dimensionality of the latent traits. Recently, Jeon (2019) proposed a multidimensional extension to confirmatory MixIRT models in which prior knowledge of the items parameters and the number of latent classes are assumed.

Furthermore, von Davier (2007) proposed a general multidimensional mixture IRT model by utilizing mixture distributions within a diagnostic modeling framework. Diagnostic modeling has been a rapidly developing research area (von Davier, DiBello, & Yamamoto, 2006). Similar to LCA, diagnostic models employ discrete latent traits; however, individuals' behaviors are described in terms of multiple discrete latent abilities instead of one single discrete latent ability. In a sense, diagnostic models can be considered as multiple classification located (or ordered) latent class models with the restriction on the number of possible latent classes (Maris, 1999; Rupp et al., 2010; von Davier, 2009). Moreover, diagnostic models are IRT-based logistic regression models and, since they are by definition multidimensional (Rupp et al., 2010), they can also be considered as multidimensional *discrete* IRT models (von Davier, 2009) as the individual's latent trait is defined by a discrete latent variable.

von Davier proposed the General Diagnostic Model (GDM; von Davier, 2005a) that unifies many of the diagnostic modeling approaches as well as some common IRT models as

special cases within the same framework. GDM uses a logistic model as do many other psychometric models for modeling the non-linear relationship of an unobserved latent variable to the probability of a correct response to an item. von Davier (2007) also demonstrated in his work that mixture-distribution multidimensional IRT models can be developed in this unified framework. Mixture GDM (MGDM) extends GDM by employing mixture distributions and, hence, offers a unique way to incorporate both multidimensionality and mixture distributions into the IRT modeling framework. In comparison to the mixture Rasch model where the latent structure is defined by a combination of one discrete latent class variable and one continuous latent trait, MGDM employs a latent class variable and multiple discrete latent variables representing individuals' underlying latent traits for multiple dimensions. Within this structure, when the underlying latent trait is assumed to be continuous instead of discrete, a general multidimensional mixture IRT model can be developed (von Davier, 2007). As a multidimensional IRT-based model, MGDM is a compensatory model, so that a low value on one dimension does not necessarily imply a low probability of a correct response. Deriving from multidimensional GPCM (Equation 2.11), the conditional probability of k^{th} category over the previous category in a polytomous item ($x \in \{0, \dots, m_i\}$) given the individual's underlying latent trait on d dimensions ($\vec{\theta} = (\theta_1 \dots \theta_d)$) and underlying latent class (g) can be defined as shown below.

$$p_i(x = k | \vec{\theta}, \vec{\alpha}_i, \beta_{ik}, g) = \frac{\exp [k \vec{\alpha}_{ig} \vec{q}_i \vec{\theta} - \sum_{h=1}^k \beta_{ihg}]}{1 + \sum_{y=1}^{m_i} \exp [y \vec{\alpha}_{ig} \vec{q}_i \vec{\theta} - \sum_{h=1}^y \beta_{ihg}]} \quad (2.13)$$

where β_{ihg} denote class-specific item intercept parameters of the transition from the $(h - 1)^{th}$ category to h^{th} category of item i in latent class g , $\vec{\alpha}_{ig}$ ($\vec{\alpha}_{ig} = (\alpha_{i1g} \dots \alpha_{idg})$) denote class-

specific item slope parameters for d dimensions for item i in latent class g , and \vec{q}_i ($\vec{q}_i = (q_{i1} \dots q_{id})$) denote the binary entries representing whether the item calls on a certain skill, namely design matrix or Q-Matrix (Rupp et al., 2010). For example, a total of I items measuring a D -dimensional construct would require an $I \times D$ matrix with binary entries.

2.2.3 Multilevel Models

The modeling approaches discussed in the previous section offer effective strategies to minimize inferential biases as much as possible by accounting for the measurement error and systemic heterogeneity within the population. In many cases, employing these techniques should result in improved parameter estimates. However, with regard to a proper ILSA data analysis, these models only treat part of the problem. The complex sampling design used in large-scale assessment studies induces a hierarchical data structure, causing certain complications in statistical analysis.

One of the issues with the nested structure of data is that it results in correlated errors in regression analyses and leads to a violation of the assumption of independence between observations that justifies the use of ordinary least squares (OLS) methods (Fox & Glas, 2001; Goldstein & Spiegelhalter, 1996). One consequence is that using OLS-based standard errors in this situation (correlated errors) leads to underestimation of the standard errors of the regression coefficients. Furthermore, if the nested structure of the data is ignored, then any relationship of interest is forced to be represented by one regression line even though there may well be differences in regressions across schools or regions. That is, if the fitted regression lines are allowed to be free for each school, they are likely to have different intercepts and/or slopes for each school. Not accounting for the nested structure of the data may hide the differences in the relationships across schools and generate misleading results. Especially in non-experimental

studies like ILSAs, it is likely there are confounding variables that can reverse the direction of the correlations when the data are aggregated (Cohen, 1986). Correlations seen at the within-group level can be reversed when the groups are combined and lead to counter-intuitive statistical findings (i.e., Simpson's paradox) (Cohen, 1986; Goltz & Smith, 2010; Simpson, 1951; Thompson, 2006).

As the complex sampling strategies are commonly used in many studies, statistical models were extended to account for the dependencies among observations. A few of the existing modeling techniques are discussed in this section.

Multilevel Linear Models

In order to take the clustered nature of the observed data into account, multilevel linear modeling (MLM) is the commonly preferred statistical method. A MLM is an extension of standard multiple regression accounting for dependency in the data due to clustering (Ravand, 2015). MLMs allow separate models for each level in the data (e.g., school- and student-level) and the residuals are separately examined in this structure (Aitkin & Longford, 1986). Thus, the issue that observations within clusters may be more similar than between clusters is properly treated. As seen in the null (or unconditional) two-level linear model below, which does not have any covariates and is equivalent to the one-way ANOVA with random effects, the residual term e_i in a simple OLS regression (Equation 2.4) is split into two terms separately estimated at each level.

$$\text{Level-1 Model: } Y_{ij} = \beta_{0j} + r_{ij} \quad (2.14)$$

$$\text{Level-2 Model: } \beta_{0j} = \gamma_{00} + u_{0j} , \quad (2.15)$$

where $r_{ij} = N(0, \sigma^2)$ and $u_{0j} = N(0, \tau)$. r_{ij} is the residual term representing the deviation of individual ij 's observed Y_{ij} from β_{0j} which is the group mean of the outcome variable in group j , and u_{0j} is the residual term representing the deviation of group j 's mean from γ_{00} which is the grand mean of the outcome for the population of groups.

The null model also allows for the calculation of the intra-class correlation (ICC) as well as the design effect and the effective sample size as discussed earlier. Estimated individual-level variance and group-level variance are used to infer the degree of dependence among the observations in terms of the outcome of interest as shown below.

$$ICC = \frac{\text{Variance between groups } (\tau)}{\text{Total variance } (\tau + \sigma^2)} \quad (2.16)$$

In addition, the relationship between the predictors and the outcome is allowed to vary from group to group and this variation among groups can be modelled by using group-level variables in a multilevel model. In the example below, which is usually called as “intercepts- and slopes-as-outcomes model”, the variation among intercepts, as well as the relationship between X_{ij} and Y_{ij} from group to group is predicted by a group-level variable W_j .

$$\text{Level-1 Model: } Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij}) + r_{ij} \quad (2.17)$$

$$\text{Level-2 Model: } \beta_{0j} = \gamma_{00} + \gamma_{01}(W_j) + u_{0j} \quad (2.18)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(W_j) + u_{1j} \quad (2.19)$$

Models can easily get very complicated and computationally expensive when many components are allowed to vary from group to group and are predicted by certain group-level variables. Therefore, depending on the application context, models should be specified carefully.

For instance, even though there is an underlying hypothesis for a variation between groups, model estimates and fit should be analyzed to determine whether the observed data support the a priori specifications of the model; if not, the model should be re-specified accordingly. However, it should be noted that searching through data for the best fitting model leads to the issue of *selective inference* and should be addressed by employing proper statistical procedures (Benjamini et al., 2009).

Although it provides a powerful solution to deal with the problems caused by the clustered data, MLMs are still limited because measurement error in the criterion variable or the predictors in the model is not treated properly (Adams et al., 1997; Battauz et al., 2011; Fox, 2007; Fox & Glas, 2001; Fox & Glas, 2003; Pastor, 2003). The effects of measurement error on the accuracy of estimates have been widely studied in single-level regression analysis (Carroll et al., 2006; Fuller, 1987; Goldstein, 1979) and similar effects were found in multilevel models (Woodhouse et al., 1996).

Multilevel Latent Variable Models

Multilevel Item Response Theory Models

In order to take into account both the measurement error and the clustering in the analysis of ILSA data, one common and straightforward method is estimating latent variables by employing standard IRT models as a separate procedure and using these IRT estimates in a multilevel modeling framework (Adams et al., 1997; Kamata, 1998; Pastor, 2003; Schofield, 2015; Sulis & Toland, 2017). However, a simultaneous estimation procedure, instead of a two-step approach, has been shown to provide more accurate estimates (Adams et al., 1997; Fox, 2007; Kamata, 1998). As one of the advantages, the bias that is caused by the two-step approach can be eliminated with direct estimation of the population parameters (e.g., the regression

coefficient of the student ability score on the predictor) from the item responses (Mislevy, 1984). Additionally, the use of student-level predictors in the same model can increase the accuracy of both item parameter estimates (Mislevy, 1987) and student ability estimates (Adams et al., 1997).

Earlier efforts towards accounting for both measurement error and clustering in a simultaneous estimation procedure utilized the standard IRT model as a multilevel model by considering the item response model to be a within-student model while allowing item parameters to vary across units (Adams et al., 1997). These methods mostly employed the Rasch model (Rasch, 1960) and incorporated only varying person ability estimates (i.e., two-level formulation). Kamata's Hierarchical Generalized Linear Model (HGLM) has offered a three-level extension by which variation among groups can also be estimated (Kamata, 1998). Over the years, there have been many different specifications of the multilevel formulation of the IRT models for different purposes (Kamata, 2001). For instance, Fox and Glas (2001) carried out Bayesian estimation by using Gibbs sampling and extended the multilevel IRT model by embedding the two-parameter IRT model into a two-level latent regression model.

Multilevel IRT (MLIRT) models are advantageous for many reasons when analyzing ILSA data. In comparison to multilevel linear models, MLIRT models allow the measurement error to be heteroscedastic, as it is defined locally. Therefore, this family of models should provide more accurate estimates of the relationships between variables by incorporating the errors in the measurement of underlying latent traits into the model, as well as by avoiding misleading correlation estimates resulting from ignoring the clustered nature of the data.

Although they are advantageous, increasing the number of parameters to be estimated, along with simultaneous estimation of the parameters for different levels in the model, brings

computational challenges. In particular, multidimensional extensions of MLIRT models appear to be a challenging, albeit growing, research area due to computational difficulties. For instance, Lu and Bolt (2015) emphasized the computational challenges they faced in their study. Instead of a simultaneous estimation of a multilevel multidimensional IRT (MLMIRT) model, they followed a two-stage approach by, first, fitting a multidimensional IRT model, and then using the resulting item parameters to fit a multilevel IRT model using PISA 2006 data. Additionally, they used the average of a set of plausible values in their models for simplicity even though it is well-established in the literature that plausible values are intended to be used as a set (von Davier, Gonzalez, & Mislevy, 2009).

In recent years, an increasing number of software packages that improve the estimation of these models have been developed. Specifically, using Markov Chain Monte Carlo (MCMC) algorithms have made the estimation of various complex models possible and, as a result, many software packages with Bayesian approaches have emerged (Junker et al., 2016; Rue et al., 2017). A comprehensive review of publicly available software packages for Bayesian multilevel modeling can be found in Mai and Zhang's (2018) paper. According to their paper, WinBUGS (Lunn et al., 2000) stands out as the most commonly used software package for Bayesian multilevel modeling. Examples of programming MCMC algorithms manually are also found in the literature (e.g., De Jong & Steenkamp, 2010) as they may be preferred to gain more flexibility to estimate more complex models (Junker et al., 2016). More recently, Zhang et al. (2019) proposed a multidimensional extension to Fox and Glas (2001) and Kamata (2001)'s MLIRT models to investigate correlations between multiple latent variables and covariates as the sources of the between- and within-cluster variations using a series of BUGS software

packages. In contrast to Fox and Glas' model, their proposed model was developed for dichotomous item response data only and has not been extended to polytomous IRT models.

Although they allow for considerable flexibility in estimating multilevel IRT models, Bayesian estimation procedures are known to place great demands on the computer processing power and take longer time to run (Mai & Zhang, 2018; Rue et al., 2017). Especially due to the use of MCMC algorithms, computing time increases with the model complexity and larger sample sizes (Mai & Zhang, 2018). Moreover, when the data is extremely sparse (e.g., cognitive data at the item level in ILSAs), Bayesian methods can be very challenging and may not converge for complex models (Mun et al., 2019). Bayesian methods also require carefully specified priors, which can be difficult for the variance-covariance parameters in mixed models (Browne & Draper, 2006; Jeon & Rabe-Hesketh, 2012).

Due to their practical infeasibility, software packages employing MCMC and Bayesian estimation methods are not usually preferred by the users (Junker et al., 2016). Zhang et al. (2019) noted that more efficient Bayesian algorithms and easy-to-use software packages are needed as the Bayesian estimation method faces real challenges when the number of items or the sample size is substantially increased. In order to soften the impact of MCMC algorithms on the computing time, MultiBUGS (Goudie et al., 2017) is a recently developed software package that builds on the existing algorithms in WinBUGS but runs independent Markov chains in parallel. Finally, R-INLA (Rue et al., 2017) employs an algorithm that approximates Bayesian inference in a computationally faster way but it is relatively limited in terms of its flexibility in comparison to other statistical packages implementing MCMC algorithms (Mai & Zhang, 2018).

Multilevel Structural Equation Models

Dependencies of the observations in the data from a clustered setting were addressed also in the SEM framework with different techniques and procedures. Muthén (1994) provided an introduction to some techniques, yet limited to a subset of models, for multilevel SEM. Using existing software packages that are particularly used for SEM such as LISREL, Muthén (1994) demonstrated that separate models for within- and between-group covariance matrices can be used to incorporate nested structure of the data. However, several limitations are listed by other authors regarding the use of separate models for within- and between-group covariance matrices (Rabe-Hesketh, Skrondal, & Pickles, 2004).

Rabe-Hesketh, Skrondal, and Pickles (2004) proposed a unifying framework that combines MLM and SEM, namely Generalized Linear Latent and Mixed Models (GLLAMM). In GLLAMM, lower-level latent variables are allowed to be regressed on latent and observed variables at the same or higher-levels. GLLAMM provides flexibility to model both dichotomous and polytomous response types. An example path diagram of a multilevel SEM model with a latent dependent variable and a latent covariate at level 2 can be seen in Figure 2.3.

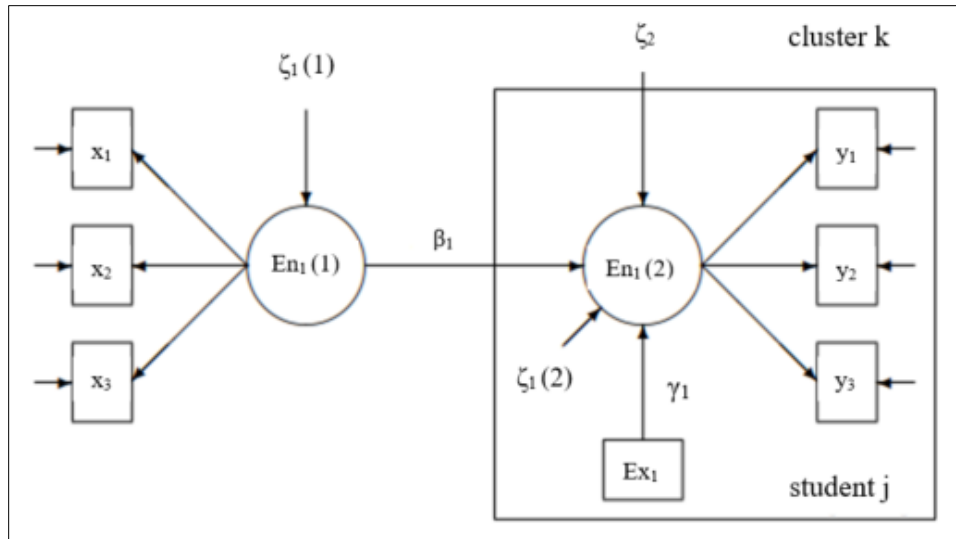


Figure 2.3. A Multilevel SEM example (Rabe-Hesketh, Skrondal, & Zheng, 2012)

Computational challenges similar to estimating multilevel IRT models are seen in the multilevel SEM literature. The authors of the GLLAMM framework (Rabe-Hesketh, Skrondal, & Pickles, 2004) developed a maximum likelihood approach to estimate the models and this procedure is implemented in the *gllamm* software package which is available in the statistical software program Stata (StataCorp, 2019). Jeon and Rabe-Hesketh (2012) improved this estimation procedure further by proposing a profile-likelihood approach for estimating complex models with maximum likelihood. This approach overcomes some of the limitations in existing software packages including *gllamm* which cannot fit models with crossed random effects. The profile-likelihood method was implemented by using the *lme4* package (Bates et al., 2015) in R (R Core Team, 2019). This procedure employs a Laplace approximation similar to the R-INLA package and, hence, is computationally faster. However, it may still require a very long runtime—depending on the research question and its complexity. For instance, the computation time for the model that was given by the authors as an example of the profile-likelihood method was about 3 days (Jeon & Rabe-Hesketh, 2012).

Furthermore, Muthén and Asparouhov (2016) proposed a series of multilevel SEMs with multidimensional extensions. They implemented both Weighted Least-Squares and Bayesian estimation methods in the Mplus software program (Muthén & Muthén, 2012) which is commonly used for SEMs and offers mixture modeling as well. A similar model, namely Mixed Effects Structural Equations (MESE) which combines SEM and IRT was developed by Schofield (2015). In particular to the MESE model, a prior for the criterion variable can be defined based on the other covariates in the structural model. It employs an MCMC algorithm and WinBUGS can be used to estimate MESE models (Schofield, 2015; Schofield et al., 2015). In the MESE framework, a variety of models are covered such as factor analysis, multidimensional IRT, and cognitive diagnostic modeling.

Multilevel Mixture Item Response Theory Models

Mixture IRT models have also been extended to multilevel models to account for the clustered nature of the data. It has been shown that multilevel mixture IRT models (MLMixIRT) can be obtained by a variety of ways; incorporating mixture distributions into multilevel IRT, extending mixture IRT models to a multilevel structure, and combining IRT into multilevel latent class models (Cho, 2007). Among these, multilevel extensions of mixture IRT models appear to be more common in the literature (e.g., Cho & Cohen, 2010; Finch & Finch, 2013; Gnaldi et al., 2016). Vermunt (2003; 2004) extended multilevel LCA and mixture models to multilevel settings by employing Maximum Likelihood (ML) estimation and expectation-maximization (EM) algorithm. He then implemented this approach (2008) in Latent GOLD software (Vermunt & Magidson, 2005) and noted that this model can also be implemented with a Bayesian MCMC algorithm using WinBUGS software. However, he pointed out the challenges similar to those were discussed in regards to Bayesian MLIRT models.

In order to incorporate multidimensionality within the MLMixIRT framework, De Jong and Steenkamp (2010) expanded their previous work on the DIF analysis (De Jong et al., 2007) which was based on Fox and Glas' (2001) MLIRT model employing Bayesian estimation. They implemented this method with an MCMC algorithm that they programmed manually and a relatively complex model with 50,000 iterations took about 24 hours. In a similar study (Finch & Finch, 2013), a multilevel multidimensional extension of the Mixed Rasch model was implemented with ML estimation using Mplus (Muthén & Muthén, 2012).

Building on Vermunt's work (2003; 2004), von Davier (2010) developed a multilevel extension to the MGDM (von Davier, 2007), namely Hierarchical Mixture General Diagnostic Models (HMGDM). HMGDM takes into account the clustering that is induced by the survey design. That is, in addition to the underlying latent class indicator, HMGDM introduces *an observed clustering variable (s)* representing sampling units in a study such as schools or classrooms. As mentioned earlier in this paper, although GDM is developed within the diagnostic modeling framework and assumes multiple discrete latent variables representing individual's underlying latent traits for multiple dimensions, the HMGDM can also be extended to multilevel multidimensional mixture IRT (MLMixMIRT) model when the underlying latent traits for multiple dimensions are assumed to be continuous instead of discrete. If it is assumed that individuals' observed responses to the items depend on their underlying latent classes (g) and their underlying latent traits only ($\vec{\theta}$), then the probability of an observed response vector can be defined as,

$$P(\vec{x}) = \sum_g P(g) \sum_{\vec{\theta}} P(\vec{\theta}|g) P(\vec{x}|\vec{\theta}, g), \quad (2.20)$$

where the marginal probability of the latent class variable is defined as

$$P(g) = \sum_s P(s) \sum_g P(g|s). \quad (2.21)$$

Note that the observed clustering indicator does not depend on the individual's underlying latent class. However, the latent class variable does depend on the clustering variable in this framework and this can be seen as adding complexity to the model (von Davier, 2010). In the case where a separate latent class distribution is assumed for each cluster, this would substantially increase the number of parameters to be estimated. That said, in many contexts including educational settings, it is more reasonable and feasible to assume that item parameters do not differ from cluster to cluster (von Davier, 2010).

HMGDM was implemented in the *mdltn* software package developed by von Davier (2005b). *mdltn* employs marginal maximum likelihood (MML) estimation and expectation-maximization (EM) algorithm to estimate mixture-distribution IRT models. EM algorithm is commonly preferred for mixture IRT models due to its flexibility with incomplete data (von Davier & Rost, 2016). It enables fitting a wide range of models including multidimensional IRT, mixture-distribution IRT, multilevel mixture-distribution IRT, and multilevel multidimensional mixture IRT models. It is suitable for large-scale data analyses and can handle sparse data with dichotomous and polytomous responses. A comparison of *mdltn* and other software packages that are used for similar diagnostic models can be found in Sen and Terzi's paper (2019).

It is expected that the choice of software package is critical in this study due to the computational challenges when estimating the very many parameters of multilevel multidimensional IRT models. Because of its flexibility in handling large and sparse data along

with modelling complex multilevel multidimensional IRT models, *mdltm* is used in this study to fit multidimensional IRT and multilevel multidimensional mixture IRT models. Similar studies conducted by the Educational Testing Service for PISA data analyses (von Davier, Yamamoto, et al., 2019) supports this choice. Mixture IRT models account for underlying subpopulations that are neither directly observed nor captured in the item response data. Therefore, by using multilevel multidimensional mixture IRT models, within-country heterogeneities in students' performances that might result from factors such as varying response styles can also be examined for the purposes of this study.

2.3 Relationship in Focus: Math Self-efficacy and Math Achievement

To illustrate and ameliorate the impact of both clustering and measurement error on the correlations reported between background data and proficiency measures in ILSAs, this dissertation offers a comprehensive investigation of the operating characteristics of competing modeling techniques. To this end, the Programme for International Student Assessment (PISA) 2012 database is utilized and the analyses focus on the relationship between PISA math achievement and math self-efficacy.

The PISA 2012 Assessment

PISA is a cross-sectional study conducted every three years under the auspices of the Organisation for Economic Co-operation and Development (OECD), with the aim of collecting internationally comparable information about 15-year-old students in participating countries or jurisdictions. Both member and non-member countries participate in PISA and the number of participating countries continues to increase since its first administration in 2000 in 32 countries. There were 79 participating countries and economies in the seventh cycle of PISA in 2018

(OECD, 2019b). PISA is designed to monitor educational outcomes and its emphasis is on knowledge and skills that are considered fundamental for students' adult life (OECD, 2017). Reading, mathematics, and science have been the major content domains in PISA, although the PISA's scope has expanded over time. For instance, financial literacy was offered as an optional assessment for the first time in PISA 2012. In PISA 2018, many other optional questionnaires were offered including global competence, well-being, educational trajectory, and familiarity with information and communications technology (OECD, 2019a). This study employs data from the fifth cycle of the PISA administered in 2012. Although this was not the most recent data available in the PISA international database by the time this study was proposed, it was the last cycle in which math was the major domain.

In PISA 2012, 34 OECD countries and 31 partner countries/economies participated and these are referred to as "countries" throughout this study. PISA follows an age-based sampling strategy and targets 15-year-old students enrolled in both school-based and work-based educational programs. In order to ensure the representativeness of the samples and comparability of the information at the country level, between 4,500 and 10,000 students from at least 150 schools are typically tested in each country. Although target cluster size is set to be 35 students within each sampled school, this may vary in some countries (OECD, 2014). A two-stage stratified sampling design was used in all participating countries except the Russian Federation. Eligible schools with 15-year-old students were systematically sampled from a comprehensive national sampling frame in the first stage by assigning a higher probability of being selected to schools with larger numbers of eligible students (i.e., Probability Proportional to Size (PPS) sampling). In the second stage, students were selected from a list of all eligible students within the sampled schools. In the Russian Federation, three-stage sampling was used with the added

sampling stage of geographical areas prior to sampling of schools and students (OECD, 2014). Targeted and achieved response rates by country are reported in the *PISA 2012 Technical Report* (2014).

Math Assessment

In PISA 2012, the major domain was mathematics literacy to which two-thirds of the two-hour testing time was allocated (OECD, 2014). Although math literacy is a part of the assessment framework in every cycle of PISA, a new formal definition was developed for PISA 2012, as math was the major domain. Math literacy was defined in *PISA 2012 Assessment and Analytical Framework* (OECD, 2013a) as follows:

“Mathematical literacy is an individual’s 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” (p. 25).

As seen from the depth of the construct definition, it is an ambitious goal to accurately assess students’ math literacy in full. In addition, PISA is a low-stakes assessment for students and this potentially reduces students’ engagement with the test and consequently may undermine the quality of information gathered (Akyol et al., 2018; Ulitzsch et al., 2019). In order to achieve necessary construct representation while keeping the testing time short enough to reduce the burden on students, PISA uses a matrix-sampling design by which each student is administered only a subset of the total available items (Gonzalez & Rutkowski, 2009). With a balanced incomplete block design, seven item blocks for the major domain, and three item blocks each for

minor domains, were developed for the assessment of reading, math, and science in PISA 2012 and these were systematically rotated to form two-hour (four item blocks) standard test booklets. Three of the seven item blocks measuring math proficiency were the same blocks that were administered in the previous PISA cycle in 2009. In total, there were 109 items measuring math proficiency; 11 of them were partial credit (i.e., complex multiple choice or constructed-response items) and 98 of them were dichotomous (i.e., multiple choice or dichotomously scored constructed-response items). Each booklet had at least one math item block so each student was administered at least 30 minutes of math testing. Similar to other ILSA designs, five plausible values are provided in PISA 2012 dataset as the individual-level math performance values and these are incorporated into the analyses conducted in this study.

Math Self-efficacy

With math being the major domain, the PISA 2012 student background questionnaire contained attitudinal measures particularly related to math literacy. This study's main focus is on the relationship between students' math self-efficacy and math performance. The literature suggests that domain-specific self-efficacy measures, such as math self-efficacy, are more closely related than generalized measures of efficacy to the criterion reference of performance and students' responses are less likely to be influenced by frame of reference effects (Marsh, Pekrun, et al., 2019; Marsh, Roche, et al., 1997; Meece et al., 1990). In support of this argument, studies found that math self-efficacy is positively related to math performance both at the within- and between-country levels (Lee, 2009). Therefore, it was hypothesized that the attitude-achievement paradox would not occur when the data were aggregated to the country level and this was considered to provide a more solid ground for this study. In addition, there has been an ongoing debate on inaccuracies observed when one assesses their own performance. In

particular, previous studies suggest that low performers are not good at assessing their own knowledge and skills due to lack of metacognitive competence (Dunning, 2011; Kruger & Dunning, 1999). On the other hand, there have been critiques of this claim, arguing that these findings may result from statistical or methodological artefacts such as measurement error (Ehrlinger et al., 2008; Krueger & Mueller, 2002). Because this study investigates the impact of measurement error and clustering on correlations reported between background data and proficiency scales in ILSAs, the relationship between math self-efficacy and math achievement was chosen to be the focus of this study -- with the hope of informing this discussion.

Self-efficacy is defined in literature as one's belief in regard to their ability to successfully accomplish desired outcomes (Bandura, 1977). It has been a focus of interest in many educational achievement studies as a strong predictor of students' academic performance (Hackett & Betz, 1989; Pajares & Miller, 1994; Schunk, 1989). PISA's math self-efficacy measure was developed and used for the first time in its 2003 administration and the same measure was again used in PISA 2012 after re-evaluation of its psychometric qualities (OECD, 2014). The scale contains eight mathematical tasks for which students are asked how confident they feel in solving them. Students' confidence levels are evaluated by using a four-point Likert scale response format, ranging from "very confident" to "not at all confident".

As an indicator for internal consistency of the scale, Cronbach's alpha values were documented for the math self-efficacy scale in the *PISA 2012 Technical Report* and the values demonstrated only minor variation across participating countries, ranging from 0.78 to 0.91 with a median of 0.83 (OECD, 2014). In addition, international item parameters (difficulty and category thresholds) were documented in the PISA technical reports. However, as also noted by others previously (Pepper, 2020), the item threshold parameters were not presented in full for all

three categories for the items in the *PISA 2012 Technical Report*. It is well-documented by OECD that the development of the scales used in PISA is supported in various ways such as conducting field studies and evaluating the psychometric quality by using proper statistical modeling approaches. Nonetheless, results from these studies were not published. For example, there is no documentation of item-fit statistics or information on differential item functioning across countries. In their comprehensive evaluation of the validity argument for the math self-efficacy scale used in PISA, Pepper (2020) argues that the degree of lack of publicly available validity evidence for PISA's math self-efficacy scale is concerning and demands attention. In the *PISA 2022 Context Questionnaire Framework* ("Field Trial Version", (OECD, 2019c)), it has been announced that the PISA 2012 math self-efficacy scale will be retained but expanded. That said, concerns raised regarding the validity of the current version of the scale are not mentioned.

Domain-specific Self-related Constructs

For the fifth research question of this study, a composite measure that combines math self-efficacy, math anxiety, and math self-concept is used as the background measure to examine if the accuracy of the correlation estimates can be improved by combining multiple background measures. Among various background variables measured in PISA 2012, these background measures were selected because it has been well-established in the literature that they are closely related constructs (Lee, 2009; Morony et al., 2013).

Indeed, it has been debated that self-concept and self-efficacy, often referred as "self-beliefs", have much in common and, thus, are mistakenly confused constructs in terms of their definition and measurement (Marsh, Pekrun, et al., 2019). Researchers suggest that self-efficacy is more descriptive of one's future success whereas self-concept is how one thinks of their capabilities with respect to others and it is more global (Marsh, Pekrun, et al., 2019; Pajares &

Miller, 1994). Therefore, domain-specific measures of self-concept such as math self-concept are claimed to be more useful to predict future performance. On the other hand, due to its sensitivity to frame of reference effects, it has been shown that the relationship between math self-concept and math achievement could be influenced by the attitude-achievement paradox (Lee, 2009; Marsh, Pekrun, et al., 2019; OECD, 2019c). In the *PISA 2012 Assessment and Analytical Framework* (OECD, 2013a), self-concept was defined as “the overall perception of one’s personal attributes based on continuous self-evaluation” (p. 185).

Math anxiety is commonly discussed along with math self-efficacy and math self-concept due to its close relationship to these constructs, as well as one’s math performance. Although no clear definition for math anxiety is made in the documentation provided for the PISA 2012 administration, it is described as “feelings of helplessness and emotional stress when dealing with mathematics” in the *PISA 2003 Technical Report* (OECD, 2005, p. 291). In the same report, it is also noted that math anxiety is found to be negatively correlated with math achievement but its effect could be indirect when math self-efficacy and math self-concept are taken into account (Ma, 1999; Meece et al., 1990). In particular, it is argued in the literature that higher levels of anxiety are associated with lower levels of self-efficacy (Bandura, 1997).

Similar to the math self-efficacy scale, the math self-concept and math anxiety scales were first developed and used in PISA 2003, and the same measures were again used in PISA 2012 after re-evaluation of their psychometric qualities (OECD, 2014). Each measure included five items with four-point Likert scale response formats, ranging from “strongly agree” to “strongly disagree”. Based on the PISA 2012 documentation, the variation in the internal consistencies of the math self-concept and math anxiety scales across participating countries were larger than those for math self-efficacy. For the math self-concept scale, Cronbach’s alpha

values ranged from 0.70 to 0.92, with a median of 0.88. The math anxiety scale typically exhibited smaller Cronbach's alpha values, ranging from 0.51 to 0.88 with a median of 0.82. International item parameters (difficulty and category thresholds) were documented in the PISA technical reports. However, as for the math self-efficacy scale, the item threshold parameters were not presented in full for all three categories for the items in the *PISA 2012 Technical Report*. There was neither documentation of item-fit statistics nor information on differential item functioning across countries provided.

Chapter 3: Methodology

3.1 Data

For this study, cognitive-, student-, and school-level data files contained in the PISA 2012 international database (OECD, 2020) were employed. Even though computer-based administration of the math assessment was optional in PISA 2012, this study only employed the paper-based assessment data files. The cognitive data file contained each student's responses to the items in the main survey that measured mathematics, reading, and science. Cognitive data was provided in both its raw/original form and scored form. It should be noted that according to the official scoring guidelines, missing responses that were not at the end of the test booklet and were followed by valid responses, were assumed to be skipped by the respondents (i.e., omitted or invalid responses) and were scored as incorrect (OECD, 2014). Although this distinction between 'not reached' and 'omitted/invalid' items has been known to introduce bias¹ (Glas & Pimentel, 2008; Ludlow & O'Leary, 1999; Rose et al., 2017), the scored cognitive data file named *INT_COG12_S_DEC03.txt* was used in this study as it was based on the official scoring guideline of PISA 2012.

The publicly available student data file named *INT_STU12_DEC03.txt* contained variables from the background measures as well as demographic information collected from students. With regard to the background measures, students' original responses to the items, as well as summary indices that were derived through IRT scaling, were contained in this data file. Five plausible values, serving as summary proficiency scores for each student in each domain,

¹ Scoring omitted responses as incorrect responses and treating not-reached items as not administered items has been shown to advantage respondents who answered the same number of items in a test but reached more items, even when they omitted more items (Ludlow & O'Leary, 1999).

were also contained in this file, as were three sets of student sampling weights (replicate, final, and senate).

Except for the analyses that required specialized software packages, all data cleaning, analyses, and visualizations were programmed in R (R Core Team, 2019).

Sample

Cognitive and student data files were merged by country, school ID, and student ID as described in *PISA Data Analysis Manual* (2009). The original data files contained data from 68 countries/economies including region-level data from three states of the United States (Florida, Connecticut, and Massachusetts) and a city of Russian Federation (Perm). Data from these four jurisdictions were removed from the analyses before merging the files. Additionally, among the 64 countries left in the data file, data from Albania were removed due to known problems in the administration of the assessment (Gortazar et al., 2014) and data from Liechtenstein were removed due to small sample size (N=293). Finally, based on the preliminary analyses, Shanghai-China was found to be an outlier with respect to country-mean math proficiency and math self-efficacy within the distribution of all sixty-two countries. Country-mean math proficiencies and math self-efficacy indices were used in this study to examine their relationships with within-country correlation estimates. Given that there are only 60+ countries observed, an outlying country could have a big influence on the results. Therefore, data from Shanghai-China were excluded from the analyses.

The final dataset that contained 468,200 students from 61 participating countries was utilized. The numbers of students and schools participating in PISA 2012 are listed by country in the Appendix (Table A.2).

Variables

Item-level variables

Student response data for four measures were employed in latent variable models. There were, in total, 127 item-level variables:

Math Proficiency (109 variables): Math performance, the sole cognitive outcome, was measured by 109 items in total; 11 were partial credit (i.e., complex multiple choice or constructed-response items) and 98 were dichotomous (i.e., multiple choice or dichotomously scored constructed-response items). The study employed scored data that was provided in the PISA international database so no extra recoding was needed for these variables.

Math Self-efficacy (8 variables): Students' math self-efficacy was measured by eight items. All eight items were reversed-coded so that higher difficulty was associated with higher math self-efficacy. The items are listed below.

Table 3.1. Items measuring math self-efficacy

Stem: How confident do you feel about having to do the following mathematics tasks?
Response categories: Very confident, Confident, Not very confident, Not at all confident

<u>Variable Name</u>	<u>Item</u>
ST37Q01	a) Using a <train timetable> to work out how long it would take to get from one place to another.
ST37Q02	b) Calculating how much cheaper a TV would be after a 30% discount.
ST37Q03	c) Calculating how many square meters of tiles you need to cover a floor.
ST37Q04	d) Understanding graphs presented in newspapers.
ST37Q05	e) Solving an equation like $3x+5 = 17$.
ST37Q06	f) Finding the actual distance between two places on a map with a 1:10 000 scale.
ST37Q07	g) Solving an equation like $2(x+3) = (x + 3) (x - 3)$.
ST37Q08	h) Calculating the petrol consumption rate of a car.

Math Anxiety (5 variables): Students' math anxiety was measured by five items. All five items were reversed-coded so that higher difficulty was associated with higher math anxiety. The items are listed below.

Table 3.2. Items measuring math anxiety

Stem: Thinking about studying mathematics: to what extent do you agree with the following statements?

Response categories: Strongly agree, Agree, Disagree, Strongly disagree

<u>Variable Name</u>	<u>Item</u>
ST42Q01	a) I often worry that it will be difficult for me in mathematics classes.
ST42Q03	c) I get very tense when I have to do mathematics homework.
ST42Q05	e) I get very nervous doing mathematics problems.
ST42Q08	h) I feel helpless when doing a mathematics problem.
ST42Q10	j) I worry that I will get poor <grades> in mathematics.

Math Self-concept (5 variables): Students' math self-concept was measured by five items. All five items except ST42Q02 were reverse-coded so the higher difficulty was associated with higher math self-concept. The items are listed below.

Table 3.3. Items measuring math self-concept

Stem: Thinking about studying mathematics: to what extent do you agree with the following statements?

Response categories: Strongly agree, Agree, Disagree, Strongly disagree

<u>Variable Name</u>	<u>Item</u>
ST42Q02	b) I am just not good at mathematics.
ST42Q04	d) I get good <grades> in mathematics.
ST42Q06	f) I learn mathematics quickly.
ST42Q07	g) I have always believed that mathematics is one of my best subjects.
ST42Q09	i) In my mathematics class, I understand even the most difficult work.

Student-level variables

Summary, individual-level indices for the background measures math self-efficacy (*MATHEFF*), math anxiety (*ANXMAT*), and math self-concept (*SCMAT*) were provided in the form of weighted likelihood estimates (WLEs). WLEs were obtained by employing the Partial Credit Model and, then, transformed to an international metric with an OECD average of zero and an OECD standard deviation of one, in the PISA database. These summary indices for the background measures, as well as five plausible values for cognitive outcome math proficiency (*PV1MATH*, *PV2MATH*, *PV3MATH*, *PV4MATH*, and *PV5MATH*), were employed in linear models in this study. To facilitate comparisons of the summary statistics gathered from different models, these variables were standardized within each country before they were employed in the analytical procedures. Standardization of the variables was done by using the *scale* function available in R. The *scale* function generates z-scores by centering the variable at its mean and scaling it by its standard deviation as shown below.

$$x_z = \frac{x - \text{mean}(x)}{\text{sd}(x)} \quad (3.1)$$

Missing Data

Similar to the design for cognitive items as described in Chapter 2, PISA 2012 introduced a rotation design for the context questionnaire, which brought challenges to this study with respect to within-country sample sizes. In this rotated design, except for the common component that contained demographic questions such as age and gender, items were grouped into three blocks and each item block was administered to only two thirds of the sample in each country (OECD, 2013a). For example, out of three forms created by rotating item blocks, the math self-

efficacy scale was placed in only Forms A and B, and not in Form C. The implication of this rotated design for this study was that there were neither response data nor imputed summary indices of students' math self-efficacy available for one third of the sample. Furthermore, math self-concept and math anxiety were placed in only Form B and C (Appendix, Figure A.1). Therefore, to be able to use all three measures for RQ5 of this study, only one third of the sample could be employed. The use of a rotated design for the context questionnaire has raised methodological concerns about possible biases (von Davier, 2014). Nonetheless, the data that were missing by design were considered as missing at random (MAR), and were listwise deleted from the dataset. Moreover, summary indices in the dataset were provided only for the students who responded to at least one item in the scale. Therefore, there were no summary scale data available for students who did not respond to any items on that scale. The missing cases remaining after listwise deletion of MAR cases (i.e. missing by design) were considered as not missing at random (NMAR). Although listwise deletion of NMAR observations can be problematic², these cases were also removed from the sample. In the final dataset, there were 305,051 students from 61 participating countries. Final sample sizes by each country are listed in the Appendix (Table A.2).

Weighting

The PISA international database provided both student full sample (unscaled) sampling weights (W_FSTUWT) as well as senate (scaled) weights ($senwgt_STU$). Senate weights were computed by normalizing unscaled sampling weights so that students' senate weights summed to 1000 in each country. The use of senate weights was preferred as they could mitigate the impact

² If the probability of being NMAR depends on the value of the criterion variable, the regression coefficient estimates are biased. See Little & Rubin (2002).

of differences among within-country sample sizes on analyses that involved more than one country. As senate weights provided in the PISA database were scaled based on the full sample, new senate weights were computed based on the final samples after the listwise deletion of the missing cases. Within each country, the new senate weights were calculated as follows.

$$new_senwgt_STU = 1000 * \frac{W_FSTUWT}{sum(W_FSTUWT)} \quad (3.2)$$

Although school-level weights were also provided in the database, they were not employed due to their incompatibility with the specialized software packages used in the study.

3.2 Analysis Overview

A step-wise model refinement approach was proposed to better understand the impact of clustering and measurement error on the analysis of PISA data. The step-wise model refinement comprised four different statistical techniques as listed below to address the first three research questions of the study.

Table 3.4. Model refinement steps

		<u>Accounting for Measurement Error</u>	
		No	Yes
<u>Accounting for Clustering</u>	No	Ordinary Regression Model (Conventional)	Multidimensional IRT Model (MIRT)
	Yes	Multilevel Linear Model (MLM)	Multilevel Multidimensional Mixture IRT (MLMixMIRT)

In order to make comparisons between models possible, correlation coefficients were chosen as the summary statistics to be compared at each step because they could easily be

obtained from each of the four modeling techniques that were used. Correlation is a measure of association and does not imply any causal relationships between the variables. That is, in contrast to regression analysis, correlation analysis does not describe a prediction of a dependent variable based on the value of an independent variable. As the main goal of this study was to make comparisons of the summary statistics and not predictive inferences, this was not a limitation.

Moreover, with regard to the comparability of the correlation estimates across various modeling techniques, it should be emphasized that modeling approaches that were used in this study required employing different sets of variables from the PISA data. As mentioned earlier, item response data were employed in latent variable models (i.e., MIRT and MLMixMIRT) whereas summary indices of the background measures and plausible values (PVs) of cognitive outcome were employed in linear models (i.e., OLS and MLM). As discussed in more detail in Chapter 2, PVs are generated by latent regression modeling which incorporates both students' background information and their responses to the items administered to capture uncertainty resulting from the use of matrix sampling of the items and to approximate sub-population characteristics. Because they carry more information than students' responses to the items, they are expected to be slightly different from other estimators of the latent variable such as EAPs or WLEs obtained by employing IRT modeling. Note that the correlation estimates obtained from the latent trait models that were conducted in this study were not based on EAP or WLE point estimates but rather based on the estimated two-dimensional latent distributions. Nevertheless, PVs and EAPs obtained from the MIRT models were compared to ensure that the use of different sources of data did not jeopardize the comparability across models.

In the first step of the model refinement, ordinary linear regression models were fit to the data from each participating country and the resulting standardized regression coefficients (i.e.,

Pearson correlations) were gathered as the conventional correlation estimates. Secondly, a two-dimensional IRT model was fit to the data from each participating country. The resulting estimated correlations between two latent traits were examined. As the next step, a multilevel linear modeling approach was used to take into account the clustering in the PISA data. Finally, in order to take into account both measurement error and clustering, multilevel multidimensional mixture IRT (MLMixMIRT) modeling was employed. With MLMixMIRT modeling, underlying subpopulations that were not directly observed in the item response data were produced based on maximizing within-class homogeneity and between-class heterogeneity, and correlation estimates for the resulting latent classes were examined accordingly, showing to what extent the variability of correlations across countries may depend on clustering of students collected in countries. Again, these models were estimated separately for each participating country in PISA 2012.

The distributions of these country-level correlation estimates were compared across models to ascertain the impact of taking into account one or both of the confounding factors. The changes in the estimates were also examined in relation to (i) the amount of the measurement error, by utilizing within-country reliability estimates and (ii) the degree of clustering, by utilizing intra-class correlations (ICC) within countries. Moreover, it was of interest to examine whether differences in correlation estimates among populations were moderated by group-level performance. This relationship was investigated both at the country-level and at the school-level within countries. Although it was expected that within-country relationships would be more similar (among populations) after the model refinement steps were implemented, population-level performance might still moderate the relationships between math self-efficacy and math performance. It was possible that even after measurement error and clustering were taken into

account, the correlations would be weaker in lower-performing countries. Therefore, this was investigated further by examining the relationship between school-level performance and the correlations between math self-efficacy and math achievement. If stronger correlations were observed in higher performing schools, it could be argued that there were other reasons, such as students' lower levels of meta-cognition (i.e., the Dunning-Kruger effect) that resulted in weaker correlations.

Finally, a supplementary set of analyses employing a composite variable that comprised multiple background indicators were conducted to determine whether using multiple indicators substantially reduced measurement error in the background measure and, hence, the impact of measurement error on correlation estimates. Principal component analysis (PCA) was used to create the composite variable. Although employing a composite indicator as the background measure instead of the math self-efficacy measure alone led to a change in the construct and had a different substantive interpretation, this procedure was carried out as part of a more general methodological investigation.

Phase 1: Conventional Analysis of Within-Country Relationships

In this first phase, the relationship between students' math self-efficacy and math performance was examined by fitting ordinary linear regression models. Because there were five plausible values of math performance provided in the PISA 2012 database, five separate models were fit to the data from each country. Summary indices of students' math self-efficacy were employed as the only predictor of the cognitive outcome and standardized regression coefficients, which are Pearson correlations when the regression model has only one predictor, were estimated.

$$(PV1MATH_z)_i = b * (MATHEFF_z)_i + e_i , \quad (3.3)$$

$$(PV2MATH_z)_i = b * (MATHEFF_z)_i + e_i , \quad (3.4)$$

$$(PV3MATH_z)_i = b * (MATHEFF_z)_i + e_i , \quad (3.5)$$

$$(PV4MATH_z)_i = b * (MATHEFF_z)_i + e_i , \quad (3.6)$$

$$(PV5MATH_z)_i = b * (MATHEFF_z)_i + e_i , \quad (3.7)$$

where $PV(1 - 5)MATH_z_i$ represent the five plausible values of student i 's math representing the performance criterion. These were standardized within each country. $MATHEFF_z_i$ is the standardized measure of student i 's math self-efficacy; b represents the standardized regression coefficient as the correlation estimate; and e_i is the residual term representing the deviation of student i 's observed math performance from their predicted math performance.

In total, sixty-one sets of five regression models were fit in R using senate weights (new_senwgt_STU). In order to use plausible values and take the imputation variance into account, Rubin's (1987) method for pooling parameter estimates was used. Rubin's method was programmed in R. This procedure is also described in the *PISA 2012 Technical Report* (OECD, 2014).

First, the population estimate was calculated by taking the average of the correlation estimates (r) from each of the five regression models as shown below.

$$r = \frac{1}{5} \sum_{m=1}^5 r_m \quad (3.8)$$

Similarly, sampling variance (U) was calculated by taking the average of the variances of the correlation estimates gathered from each of the five regression models.

$$U = \frac{1}{5} \sum_{m=1}^5 U_m \quad (3.9)$$

Using plausible values, imputation variance (B) (i.e., the uncertainty in the estimates due to measurement error) was obtained by calculating the deviations of each correlation estimate from the population estimate as shown below.

$$B = \frac{1}{4} \sum_{m=1}^5 (r_m - r)^2 \quad (3.10)$$

Finally, the total variance (V) of the pooled population estimate was calculated by combining sampling and imputation variances as shown below.

$$V = U + \left(\frac{6}{5}\right) * B \quad (3.11)$$

The distribution of Pearson correlations of the within-country relationships between math self-efficacy and math proficiency was examined. Because conventional correlation estimates did not account for differences in the data gathered from countries with regard to measurement error or clustering, the distribution was expected to display substantial variability. Moreover, it was of interest to examine whether the attitude-achievement paradox held true when the data were aggregated to the country level. According to the correlation estimates reported in the PISA 2012 results (OECD, 2013b), within-country relationships between math proficiency and math self-efficacy were positive in all participating countries. The Pearson correlation between mean

math self-efficacy and mean math proficiency at the country level was evaluated for differences from the results found at the within-country level.

It was also hypothesized that the within-country relationships would be weaker in lower-achieving countries. Students with lower cognitive skills are generally inclined to randomly guess or omit items due to factors such as not understanding the questions or being unfamiliar with the types of questions given in PISA. Therefore, both the cognitive and non-cognitive measures from lower-achieving countries were likely to incorporate substantial uncertainty and conventional estimates of within-country relationships were expected to be attenuated. A scatterplot of the relationship between within-country correlations and country-mean math proficiency was examined to determine whether the expected differences emerge.

Phase 2: Refinement with Multidimensional IRT Modeling

After obtaining the conventional within-country correlation estimates in the first phase, the first step of the model refinement process was performed. Two-dimensional two-parameter logistic IRT models were fit to account for the uncertainty in both cognitive outcome and non-cognitive background measure. The correlations estimated using this multidimensional item response theory (MIRT) approach are hereafter referred to as “MIRT estimates”. The research question to be answered in this phase of the study was:

(RQ1) “When modeling techniques are used to properly account for *measurement error* in the observed data, do the estimates of within-country relationships between math self-efficacy and math achievement display greater homogeneity across countries than the conventional, within-country correlation estimates?”

In total, sixty-one MIRT models were fit by using the *mltm* software package (von Davier, 2005b) which was chosen for this study because it could accommodate both sparse data and large-scale data analyses. Moreover, *mltm* enables fitting multilevel multidimensional mixture IRT models – required for the final step of the model refinement. Two-dimensional IRT models were fit to the item response data in a confirmatory manner; that is, math proficiency items and math self-efficacy items were assigned to two separate dimensions in the estimation. In order to incorporate the multidimensionality within the model, a binary 117 x 2 Q-matrix was developed (Appendix, Table A.3); 8 math self-efficacy items were specified with a value of one under the first dimension and zero under the second dimension, 109 math items were specified as zero under the first dimension and one under the second dimension.

Because both math self-efficacy and math proficiency scales contained polytomous items ($x \in \{0, \dots, m_i\}$), the multidimensional extension of the GPCM, discussed in Chapter 2, was employed to model responses to polytomous items. In this model, the probability of selecting the k^{th} category over the previous category for an individual with underlying two-dimensional trait ($\vec{\theta}$) is defined for item i as follows.

$$p_i(x = k | \vec{\theta}, \vec{\alpha}_i, \beta_{ik}) = \frac{\exp [k\vec{\alpha}_i\vec{\theta} - \sum_{h=1}^k \beta_{ih}]}{1 + \sum_{y=1}^{m_i} \exp [y\vec{\alpha}_i\vec{\theta} - \sum_{h=1}^y \beta_{ih}]} \quad (3.12)$$

where β_{ih} is the intercept parameter of the transition from the $(h - 1)^{th}$ category to h^{th} category, $\vec{\alpha}_i$ is the item slope parameter, and m_i is the total number of categories.

Individuals' underlying latent traits were estimated by using *expected a posteriori* (EAP or *Bayes Mean Estimate*) (Bock & Aitkin, 1981; Bock & Mislevy, 1982) by which the mean of the posterior distribution resulting from the estimation process was used as the final estimate ($\hat{\theta}$).

In comparison to the maximum likelihood estimator (MLE), EAP can provide estimates for all response patterns including all correct or all incorrect. Therefore, finite θ estimates were available for all individuals in the data. Moreover, EAP has been found to be advantageous in comparison to *maximum a posteriori* (MAP or *Bayes Modal Estimate*) because (i) the mean squared error of EAP estimates is smaller and (ii) EAP is more efficient due to its simpler calculation requirements.

Ability estimates gathered from the multidimensional IRT models were standardized to make them comparable with the estimates from the other models employed in the model refinement process. Standardized estimates were obtained from the *mdltn* output. Note that correlations between latent traits were estimated jointly within the multidimensional IRT estimation and were not based on EAP point estimates but rather based on the estimated two-dimensional latent distributions. Therefore, the estimated correlations were corrected for measurement error. These correlations were labelled “skill distribution correlations” in the output.

Because measurement error in the response variables could result in attenuation in the correlation estimates, conventional within-country correlations that were found in Phase 1 were expected to display substantial variation due to the differences in the amount of measurement error across countries. When latent trait models were introduced and measurement error was properly accounted for, it was expected that the distribution of within-country correlations would have smaller variance compared to the variance of the distribution of conventional within-country correlation estimates. Moreover, the relationship between within-country correlations and country-level proficiency was expected to be weakened for MIRT estimates in comparison to conventional estimates. Both cognitive and non-cognitive measures are known to incorporate

substantial uncertainty in the data gathered from especially lower-achieving countries. Therefore, MIRT estimates were expected to be not as strongly associated with country-level proficiencies as conventional linear regression based estimates.

The magnitudes of the changes at the country level were also examined in relation to the magnitudes of the measurement error at the country-level, by using empirical reliabilities. In contrast to Classical Test Theory, a scale's reliability is not constant but conditional on the ability estimate in the IRT framework. However, in order to assess the quality of estimation for the entire scale, empirical reliability indices can be used. In this study, EAP reliability (Adams, 2005) was used to examine the amount of measurement error in the data. EAP reliability was calculated based on the ratio of the variance of the ability estimates (i.e., EAP estimates) to the population variance of the latent variable as shown below.

$$EAP\ Reliability = \frac{var(\hat{\theta})}{var(\hat{\theta}) + mean(\hat{\delta})} \quad (3.13)$$

where $\hat{\delta}$ denotes the square of the estimated measurement error attached to each EAP estimate resulting from the multidimensional IRT modeling.

For each country, two EAP reliability indices were calculated; one for math self-efficacy and one for math proficiency. Similar to the CTT-based reliability indices, the EAP reliability has a range from 0 to 1 and values that are closer to 1 are desirable as they indicate smaller error variance. However, IRT-based reliability indices are expected to be more precise than conventional reliability estimates such as Cronbach's alpha. Country-level EAP reliabilities were examined in relation to the changes in the correlation estimates from Phase 1 to Phase 2. Because

measurement error could be a major cause of attenuation in the correlation estimates, greater changes were expected to be seen for the countries with larger measurement error.

Phase 3: Refinement with Multilevel Linear Modeling

In this phase of the analyses, two-level linear models were fit within each country to estimate the correlation between math performance and math self-efficacy while taking into account the clustering of students in schools. These estimates are hereafter referred to as “MLM estimates”. The research question that was the focus in this phase of the study was:

(RQ2) “When the *clustered nature* of the observed data is taken into account, do the estimates of within-country relationships between math self-efficacy and math achievement have greater homogeneity across countries than displayed by the conventional correlation estimates?”

First, unconditional models were fit with no predictors to examine the variation among schools within each country with respect to the outcome variable and to determine whether there was a need for multilevel modeling.

Unconditional model:

$$\text{Level-1 Model: } (PV1MATH_z)_{ij} = \beta_{0j} + r_{ij} \quad (3.14)$$

$$\text{Level-2 Model: } \beta_{0j} = \gamma_{00} + u_{0j} \quad (3.15)$$

$$\text{Mixed Model: } (PV1MATH_z)_{ij} = \gamma_{00} + u_{0j} + r_{ij} \quad (3.16)$$

In the models above, $(PV1MATH_z)_{ij}$ represents the first standardized plausible value of math performance of student i in school j ; β_{0j} represents the mean math performance in school j ;

γ_{00} represents the average of the school means (or grand mean) of the outcome for the population of schools within the country; u_{0j} is the residual term representing school j 's deviation from the grand mean; and r_{ij} is the residual term representing the deviation of student ij 's observed math performance from the mean math performance in school j . This model was repeated for all five plausible values ($PV(1 - 5)MATH_z_i$) and Rubin's method used in Phase 1 was employed for pooling the estimates for each country.

The unconditional models enabled calculation of the intra-class correlations (ICC) which was the proportion of total variance in students' math performance that could be attributed to the variation among schools.

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma^2}, \quad (3.17)$$

where τ_{00} is the estimated between-school variance and σ^2 is the estimated within-school variance.

As the next step of Phase 3, a second set of two-level linear models in which student math self-efficacy index was added as a covariate at the student-level was fit. In these models, the intercept β_{0j} was allowed to vary across schools but no school-level variable was added to explain this variation. School-to-school variation of the relationship between math performance and math self-efficacy (β_{1j}) was not allowed (i.e., fixed slope).

Conditional Model:

$$\text{Level-1 Model: } (PV1MATH_z)_{ij} = \beta_{0j} + \beta_{1j}(MATHEFF_z)_{ij} + r_{ij} \quad (3.18)$$

$$\text{Level-2 Model: } \beta_{0j} = \gamma_{00} + u_{0j} \quad (3.19)$$

$$\beta_{1j} = \gamma_{10} \quad (3.20)$$

$$\text{Mixed Model: } (PV1MATH_z)_{ij} = \gamma_{00} + \gamma_{10}(MATHEFF_z)_{ij} + u_{0j} + r_{ij}, \quad (3.21)$$

where γ_{10} represents the average school mean (or grand mean) of the correlation between math self-efficacy and math performance. Similar to the unconditional models, conditional models were for all five plausible values ($PV(1 - 5)MATH_z_i$) and Rubin's method was employed for pooling the estimates for each country.

In total, sixty-one sets of unconditional and sixty-one sets of conditional two-level linear models were fit by using the *lme4 package* (Bates et al., 2015) in R, which allowed the use of senate weights. In order to employ plausible values, the same procedure with Rubin's method for pooling estimates used in Phase 1 was used for pooling the MLM correlation estimates (β_{1j}) for each country.

The distribution of standardized regression coefficients as correlation estimates of the within-country relationships between math self-efficacy and math proficiency was examined. It was expected that the distribution had smaller variability compared to the one found in Phase 1. That is, when the clustered nature of the data was not taken into account, the correlation estimates could vary due to the differences among countries with respect to the level of dissimilarity across schools (i.e., school level achievement gaps). Therefore, within-country correlation estimates were expected to be more similar when the clustering of students within schools was accounted for.

Moreover, the question whether within-country correlations were associated with country-level average proficiency was examined with respect to whether this association changed after accounting for the clustering. Taking the clustering into account was likely to lead to less variation among within-country correlations. Consequently, the association between conventional correlation estimates and country-level proficiencies was likely to be weaker. Furthermore, if the correlation estimates differed in comparison to the conventional estimates found in Phase 1, the magnitudes of the changes at the country level were examined in terms of country-level between-school variances (i.e., ICCs). The changes were expected to be greater for countries where the ICCs were larger and the conventional correlation estimates were less accurate.

Phase 4: Refinement with Multilevel Multidimensional Mixture IRT Modeling

In this phase of the analyses, the final step of the model refinement process constituted fitting multilevel two-dimensional mixture IRT models in each country to obtain estimated correlations between math performance and math self-efficacy, while taking account of both the measurement error in the variables and the clustering in the data. In addition, employing mixture-distribution IRT enabled accounting for the possibility of distinct underlying subpopulations within the population (i.e., latent class structure). The estimates gathered from MLMixMIRT models are hereafter referred to as “MLMixMIRT estimates”. The research question that was the focus of this phase was:

(RQ3) “When modeling techniques are used to properly account for *both measurement error and clustering* in the observed data, do the estimates of within-country relationships between math self-efficacy and math achievement show greater homogeneity across countries than that displayed by the conventional correlation

estimates or the ones gathered from the models that account for either measurement error or clustering only?”

In total, sixty-one multilevel two-dimensional mixture IRT models were fit by using the *mdltm* package developed by von Davier (2005b). This package was used in this study to fit multidimensional IRT and multilevel multidimensional mixture IRT models due to its flexibility in handling large, sparse datasets, along with its ability to model complex multilevel multidimensional IRT models.

Although the underlying latent structure is generally unknown in mixture-distribution IRT models, a smaller number of latent classes is preferred so that it is more likely that the latent classes can be meaningfully interpreted. However, optimal decision rules on the number of classes in LCA and mixture IRT models is an unresolved problem (Nylund et al., 2007). If the decision on the number of latent classes cannot be informed by previous research on the topic or the research context, exploratory approaches are adopted. These conduct goodness-of-fit analyses, such as those based on likelihood ratio or model-data fit indices to choose the most parsimonious model (Sen & Terzi, 2019). For this study, constraining the MLMixMIRT models to have two latent classes is a reasonable choice based on previous research (von Davier, 2005a; von Davier, Yamamoto, et al., 2019). Moreover, because the model structure had to be kept the same across all countries in order to have comparable estimates, a smaller number of latent classes was preferred so that small class proportions did not become an issue and resulting latent classes could be interpreted similarly. Given the size of the dataset and the volume of analyses in this study, a larger number of classes was not tested but should be considered in future research.

Within each country, the split into two classes was allowed to vary across clusters (i.e., schools), so that different proportions of students could be in latent class 1 and 2 respectively,

depending on which cluster they were sampled from. This makes the approach a multilevel latent class model (Vermunt, 2008). In addition, for each country, separate sets of item parameters were estimated for each of the two latent classes. That is, students' observed responses were assumed to depend on both their underlying latent trait on the two dimensions ($\vec{\theta} = (\theta_1, \theta_2)$) and underlying latent class ($g \in \{1,2\}$). However, as mentioned in Chapter 2, it was reasonable and more feasible to assume that item parameters did not differ from cluster to cluster and the same latent class distribution could be assumed for all clusters (s) in the data. Therefore, the proportions of students within each latent class were allowed to vary across schools, while the classes that were assumed to exist within schools had the same ability distribution across schools. Note that MLMixMIRT models were fit separately to the data from each country and, thus, class-specific item parameter sets did differ from country to country. As an example, a representation of the multilevel estimation process for one country and one dimension is given below.

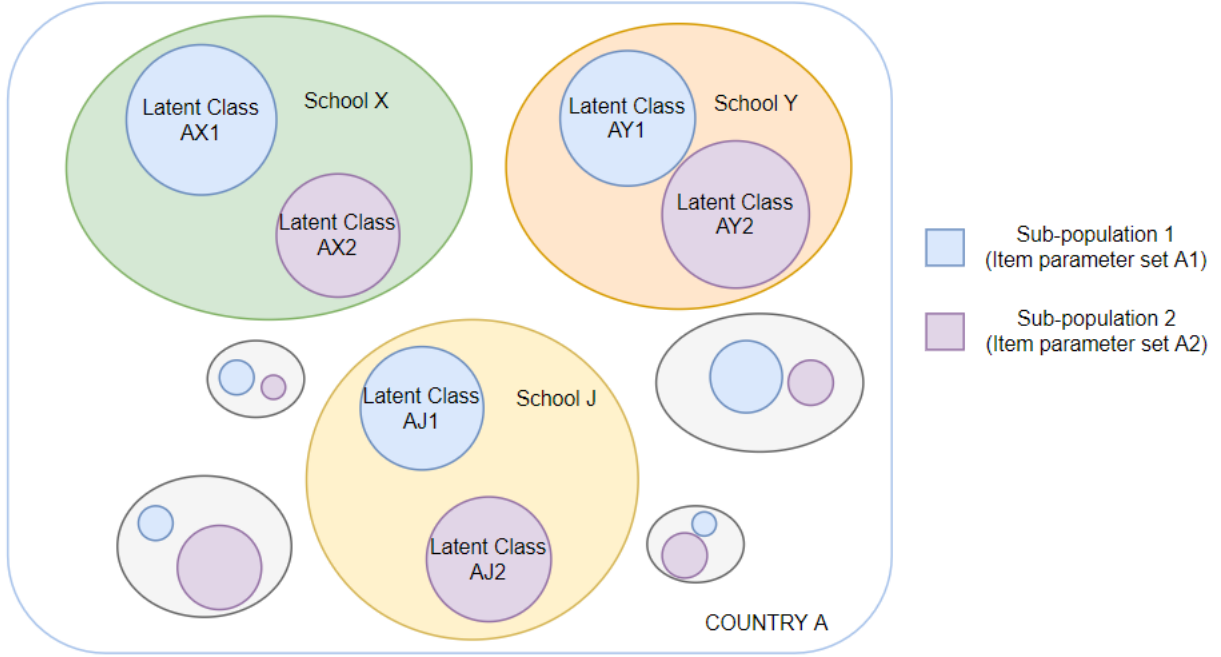


Figure 3.1. Illustration of MLMixMIRT modeling process for one country in a unidimensional setting

The 117 x 2 Q-matrix that was developed in Phase 2 to fit MIRT models was also used to fit MLMixMIRT models (Appendix, Table A.3). Because both math self-efficacy and math proficiency scales had polytomous items ($x \in \{0, \dots, m_i\}$), the multidimensional extension of the GPCM with a mixture distribution was employed to model students' responses. In this model, the conditional probability of k^{th} category over the previous category in a polytomous item ($x \in \{0, \dots, m_i\}$) given the individual's underlying latent trait on two dimensions ($\vec{\theta}$) and underlying latent class (g) can be defined as shown below.

$$p_i(x = k | \vec{\theta}, \vec{\alpha}_i, \beta_{ik}, g) = \frac{\exp [k \vec{\alpha}_{i,g} \vec{q}_i \vec{\theta} - \sum_{h=1}^k \beta_{ihg}]}{1 + \sum_{y=1}^{m_i} \exp [y \vec{\alpha}_{i,g} \vec{q}_i \vec{\theta} - \sum_{h=1}^y \beta_{ihg}]} \quad (3.22)$$

where β_{ihg} denote class-specific item intercept parameters of the transition from the $(h - 1)^{th}$ category to h^{th} category of item i in latent class g , $\vec{\alpha}_{i,g}$ ($\vec{\alpha}_{i,g} = (\alpha_{i1g}, \alpha_{i2g})$) denote class-specific

item slope parameters for two dimensions for item i in latent class g , and \vec{q}_i ($\vec{q}_i = (q_{i1}, q_{i2})$) denote the binary entries representing whether the item calls on a certain skill (obtained from the Q-Matrix).

Note that, the equation above is the same as the one derived for mixture IRT models described in Chapter 2 (Equation 2.13). However, the multilevel modeling structure originates from how the marginal probability of the latent class variable (g) depends on the observed clustering variable (s) as shown below.

$$P(g) = \pi_g = \sum_s P(s) \sum_g P(g|s), \quad (3.23)$$

when it is assumed that a student's observed responses to the items depend on their underlying latent class membership (g) and their underlying latent trait only ($\vec{\theta}$),

$$P(\vec{x}) = \sum_g P(g) \sum_{\vec{\theta}} P(\vec{\theta}|g) P(\vec{x}|\vec{\theta}, g). \quad (3.24)$$

Also note that, $P(g) = \pi_g$ is the proportion of the group (i.e., latent class) within the population (also called mixing proportion or class size). In this study, the model was constrained to have two latent classes ($g \in \{1,2\}$) with respect to each of the two dimensions (i.e., math self-efficacy and math proficiency) within each school.

Similar to the MIRT models, students' underlying latent traits were estimated by using *expected a posteriori* (EAP or *Bayes Mean Estimate*) (Bock & Aitkin, 1981; Bock & Mislevy, 1982); that is, the mean of the posterior distribution resulting from the estimation process was used as the final estimate ($\hat{\theta}$). Standardized ability estimates were obtained directly from the

mdltn output. Moreover, correlations between latent traits, which are estimated jointly within the MLMixMIRT estimation, were also obtained directly from the *mdltn* output. Because there were two latent classes in the model, correlations between math proficiency and math self-efficacy were estimated separately for each latent class. As a result, a pair of correlation estimates were obtained for each country in the dataset.

The distribution (across countries) of correlation estimates, for each class, of the within-country relationships between math self-efficacy and math proficiency was then examined. It was expected that, for each class, the distribution would exhibit smaller variance in comparison to the distributions obtained in the first three phases of the analyses. In other words, when both measurement error and clustering area accounted for, country-specific correlation estimates were expected to be more similar across countries. In addition, the patterns in within-country relationships in relation to country-level proficiencies were examined to ascertain whether they were changed after accounting for both the clustering and the measurement error in the data. It was hypothesized that with more accurate correlation estimates, a weaker association between within-country relationships and country-level averages of proficiency would be found in comparison to those seen in the first three phases. The two latent classes resulting from the estimates of the MLMixMIRT models were expected to offer more insights regarding any patterns observed in within-country relationships, as well as their relationships to country-level proficiencies.

Furthermore, since the MLMixMIRT estimates may differ from the conventional estimates and MLM estimates obtained from models that did not treat measurement error appropriately, the magnitudes of the changes at the country level were examined in terms of the magnitude of the measurement error at the country-level. Similar to the Phase 2 with MIRT

models, EAP reliability indices were used as an indicator of the amount of measurement error in the data. Because there were two latent classes, four EAP reliability indices were calculated; two for math self-efficacy for each latent class and two for math proficiency scale for each latent class. Country-level EAP reliabilities were examined in relation to the changes in the correlation estimates from the conventional estimates obtained in Phase 1 and from the MLM estimates obtained in Phase 2 to the estimates obtained to MLMixMIRT models in Phase 4. Because measurement error could be a source of attenuation affecting within-country correlations, greater changes were expected when the measurement error was larger and properly taken into account. In addition, if MLMixMIRT estimates differed in comparison to conventional estimates and MIRT estimates obtained from models that did not treat the clustering properly, the magnitudes of the changes at the country level were examined in terms of country-level between-school variances (i.e., ICCs) that were gathered in Phase 3. The changes were expected to be greater for the countries where the differences among schools were larger and the conventional correlation estimates and MIRT estimates were less accurate.

Phase 5: Refinement at the School-level Within-countries

Previous phases of the study tackled both measurement error and clustering in the data in a step-wise manner because both of these statistical artefacts are present when the PISA data is analyzed at the country-level. As mentioned in Chapter 2, clustering of the data originates from the PISA study design, whereas various factors can play roles in introducing different amounts and distributions of uncertainty in the data. The impact of measurement error on the correlation estimates was of particular interest for this study and, therefore, it was investigated further in Phase 5.

It was conjectured that both the cognitive and non-cognitive measures from lower-achieving countries were likely to incorporate greater uncertainty and, further, that country-level correlations between math proficiency and math self-efficacy were moderated by country mean math performance levels. Thus, the conventional correlation estimates obtained in Phase 1 between math proficiency and math self-efficacy were expected to be weaker in lower achieving countries. Furthermore, when advanced modeling techniques that properly account for measurement error were employed, it was hypothesized that the attenuation resulting from measurement error in data would be mitigated and correlation estimates would be more similar across countries regardless of countries' math achievement levels. In order to investigate if the relationships between math proficiency and math self-efficacy were moderated by math performance levels, school-level relationships within countries were examined to answer the research question:

(RQ4) “Are the relationship patterns in school-level correlations and school-level proficiencies within countries similar to the patterns seen at the country-level?”

When analyzed at the school-level, the PISA data is no longer clustered and the impact of measurement error on the estimates can be investigated directly. Hence, Phase 1 and Phase 2 of the model refinement process was repeated at the school-level for a selected set of countries exhibiting a wide range of country-level conventional correlations between the two variables of interest. It was expected that conventional correlations obtained from lower-achieving schools would be attenuated due to larger measurement error and, in analogy with country-level analyses, were lower than those in higher-achieving schools. Furthermore, when measurement error was taken into account, it was hypothesized that the correlation estimates would be more similar across schools within countries regardless of schools' math achievement levels.

In order to carry out the selection, countries were ordered based on country-level conventional correlations and two countries were selected from each of the bottom, middle, and top quartiles: Kazakhstan ($r_{OLS} = 0.28$), Jordan ($r_{OLS} = 0.29$), Greece ($r_{OLS} = 0.48$), Netherlands ($r_{OLS} = 0.48$), Portugal ($r_{OLS} = 0.64$), and Chinese Taipei ($r_{OLS} = 0.66$). Because the analyses were conducted at the school-level, schools with sample sizes smaller than 15 students were excluded so that sufficient within-school sample sizes could be ensured to support subsequent modeling procedures. The remaining sample included more than 100 schools in each country and the school sample sizes ranged from 15 to 27 students.

Although the main goal of Phase 5 was to replicate country-level analyses at the school-level, preliminary analyses revealed convergence issues with MIRT modeling at the school-level. These issues could not be resolved as there were too many parameters to be estimated, given the small sample sizes within schools. Therefore, as an alternative strategy, schools within each country were grouped based on their average math performances to conduct further analyses. It was expected that the convergence issues encountered when running complex models could thus be avoided. Although dividing the school sample in each country into a small number of school groups based on their math performances led to range restrictions in math proficiency, this could still address the hypothesis whether the correlations between math proficiency and math self-efficacy were moderated by math proficiency level.

Using the first plausible value of math proficiency (*PVIMATH*), schools within each country were ordered based on average mean math performance and grouped into nine “super-schools”: schools falling into bottom 15% of the distribution were grouped into *super-school 1*, those in the top 15% were grouped into *super-school 9*, and the remaining schools were grouped into seven super-schools, each containing approximately 10% of the sample. In order to ensure

the comparability of average scores across schools, school-mean math performances were calculated by employing a new set of senate weights, which summed to 30 in each school. The number of schools within each super-school ranged from 13 to 32 across countries and the super-school sample sizes ranged from 229 to 698 students. More details regarding the sample sizes by super-schools in each country are given in Appendix, Table A.4.

After grouping schools within each country into nine super-schools, the relationship between students' math self-efficacy and math performance was examined by fitting ordinary linear regression models as described in Phase 1. As shown in equations 3.3 to 3.7, five separate models were fit to the data from each super-school and Rubin's method was employed for pooling the estimates for each super-school. Plausible values and the summary index for math self-efficacy were standardized within each super school before they were employed in the models. A new set of senate weights that summed to 1000 within each super-school were used in these models to make the super-schools comparable with each other. Resulting Pearson correlations between math proficiency and math self-efficacy were examined across super-schools within each country. In addition, patterns observed within countries were compared across the six selected countries, in analogy with Phase 1.

After obtaining the conventional correlation estimates within super-schools by fitting the OLS models, the MIRT models were fit to account for the uncertainty in both the cognitive outcome and the background measure as described in Phase 2. Senate weights that summed to 1000 in each super-school were used in the models. The MIRT models employed the same Q-matrix used in Phase 2 and Phase 4. Resulting correlations between two latent traits were compared across super-schools within each country and the patterns observed within countries

were compared across the six selected countries, in analogy with Phase 2. Finally, MIRT estimates were compared to those obtained from the OLS models.

Phase 6: Refinement with a Composite Background Variable

In this last phase of the analytical procedures, a composite background variable that comprised math-self-efficacy, math anxiety, and math self-concept was employed in order to examine whether using multiple indicators substantially reduced the impact of measurement error on correlation estimates. The research question that animated this phase of the study was:

(RQ5) “When participants’ attitudinal background measure comprises *multiple indicators* (math-self-efficacy, math anxiety, and math self-concept), how different are the changes seen in the correlation estimates when clustering is taken into account compared to those seen when math self-efficacy is used as a single background measure?”

As was mentioned earlier in this chapter, only one third of the sample could be employed in this phase due to the use of rotated design for the context questionnaire in PISA 2012. More details on within-country sample sizes are given in Appendix, Table A.5.

The composite variable was created by conducting principal component analysis (PCA). In PCA, correlated variables are linearly combined into a set of uncorrelated variables (i.e., orthogonal principal components). This way, data can be plausibly summarized by fewer variables that are statistically distinct from one another and account for a substantial fraction of the variance. The first principal component is that linear combination of observed variables with maximal variance. Here, the first principal component resulting from the PCA procedure was used as the composite background measure. Although employing a composite indicator as the background measure instead of the math self-efficacy measure alone implicitly constitutes a

change in the construct with a different substantive interpretation, this procedure was proposed primarily as a methodological investigation.

PCA was conducted within each country using the *FactoMineR* package (Le et al., 2008) in R, which allowed the use of senate weights. Note that a new set of senate weights was calculated since only one third of the sample could be employed in this phase due to the rotated design of the context questionnaires as described previously. In order to obtain a score resulting from the first principal component for each student in the dataset, students' missing responses to each item were replaced by the average response value of the corresponding item³. Phases 1 and 3 were repeated by employing the composite measure in the models as the background variable. Because only one third of the sample could be employed in this phase, phases 1 and 3 were repeated on the same one third of the sample to be able to make direct comparisons between estimates obtained from models employing the composite measure and those obtained from models employing math self-efficacy measure alone. Correlation estimates from the two models were then compared. Moreover, conventional estimates and MLM estimates that were obtained in this phase with the composite measure in the models were compared with each other to examine whether the changes in the estimates from the OLS to MLM were less pronounced when the background measure comprised multiple indicators.

³ This procedure is provided by default in the *FactoMineR* package. The uncertainty is underestimated when missing data is replaced by the mean.

Chapter 4: Results

In this chapter, the results of the analyses are presented. As outlined in Chapter 3, the analyses were undertaken in six phases:

- Phase 1: Conventional Analysis of Within-Country Relationships (Baseline)
- Phase 2: Refinement with Multidimensional IRT Modeling (Step 1)
- Phase 3: Refinement with Multilevel Linear Modeling (Step 2)
- Phase 4: Refinement with Multilevel Multidimensional Mixture IRT Modeling (Step 3)
- Phase 5: Refinement at the School-level Within-countries
- Phase 6: Refinement with a Composite Background Variable

4.1 Phase 1: Conventional Analysis of Within-Country Relationships

In the first phase of the analyses, ordinary regression models were fit to data from each country separately and within-country correlation estimates between math proficiency and math self-efficacy scores available in the public use data were gathered as the baseline for comparisons across models in the model-refinement process. These conventional correlation estimates were found to be positive in all sixty-one countries but showed substantial variation across countries ranging from 0.15 (Colombia) to 0.66 (Chinese Taipei). This suggests that there are countries where a student's math self-efficacy is only weakly associated with their math proficiency. The interquartile range (IQR) of the distribution was 0.216. The distribution of the correlation estimates was observed to be bimodal with two distinct peaks; a group of estimates clustered around 0.3 and a group clustered around 0.5. There were more countries with larger correlation estimates and the overall distribution was negatively skewed. Descriptive statistics

and the histogram of the distribution of the conventional correlation estimates are presented below.

Table 4.1. Descriptive statistics for the within-country conventional correlation estimates from OLS models

<u>Correlations</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
Conventional	0.458	0.133	0.151	0.501	0.664	0.514	-0.682	0.216

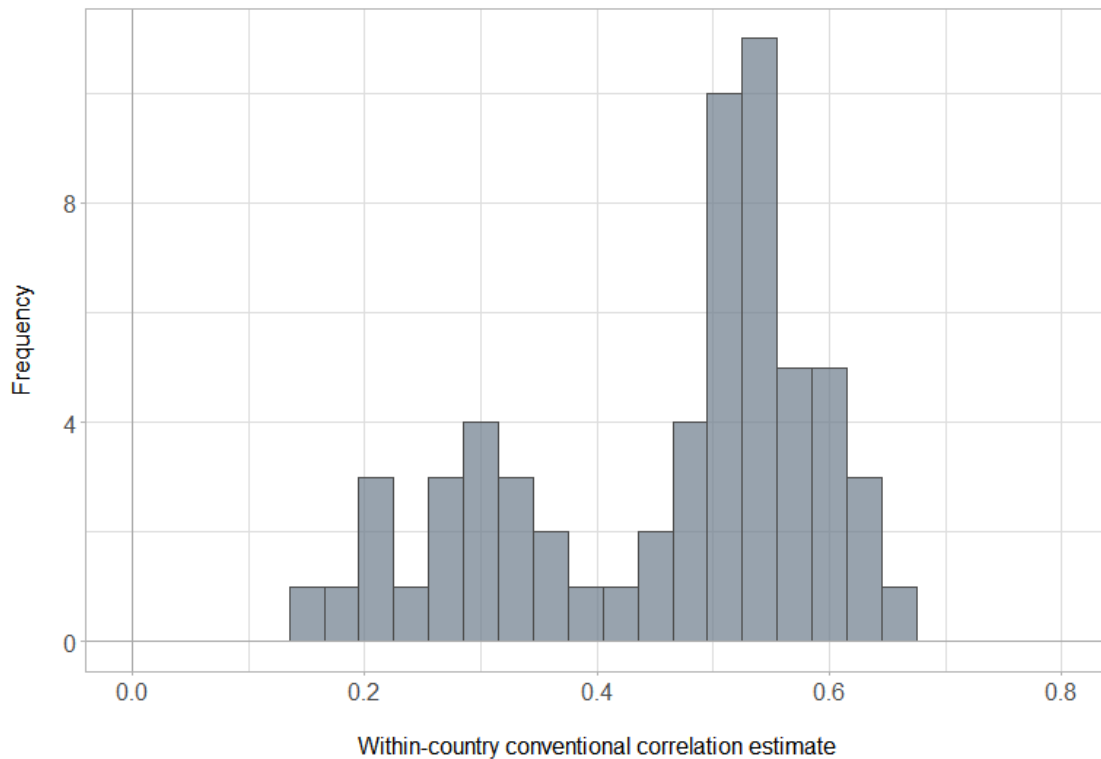


Figure 4.1. Distribution of the within-country conventional correlations

It was also of interest to examine whether the attitude-achievement paradox held true when the data were aggregated to the country level. Since all within-country correlation estimates were found to be positive in this study, a negative correlation between country-mean math proficiencies and country-mean math self-efficacy indices would indicate a reversal in the

relationship. Pearson correlation between country-mean math proficiencies and country-mean math self-efficacy indices was found to be positive ($r=0.55$), which indicates that the attitude-achievement paradox was not found for this relationship in PISA 2012. However, there were outliers as shown in Figure 4.2 below. For instance, Japan and Korea demonstrated low math self-efficacy despite their high math performance. Jordan and Kazakhstan, on the other hand, demonstrated higher math self-efficacy despite their lower math performance.

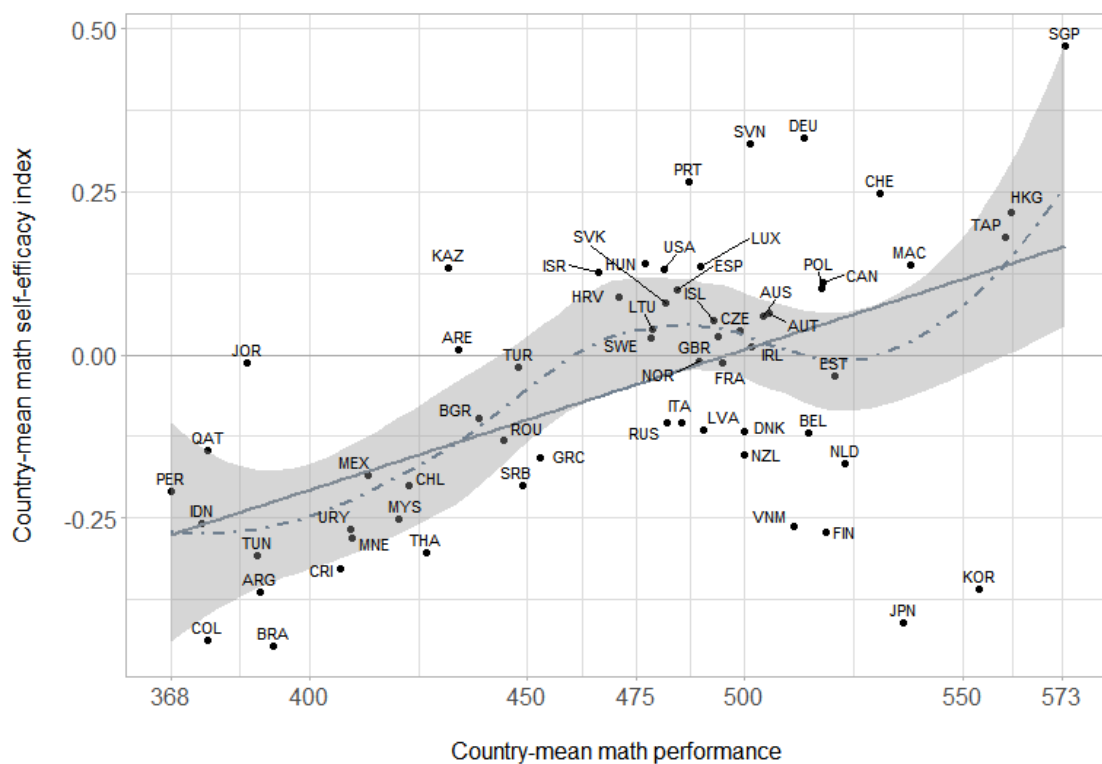


Figure 4.2. Relationship between country-mean math performance and math self-efficacy

When the relationship between conventional correlation estimates and country-mean math performances was examined, the correlations in lower achieving countries were found to be smaller compared to higher achieving countries. In addition, the relationship between correlation estimates and country-mean math proficiencies was approximately linear up to 475 for country-

mean math performance. In higher achieving countries (country-mean math performances above 475), the relationship was weakened, approaching an asymptote of about 0.55. (Figure 4.3).

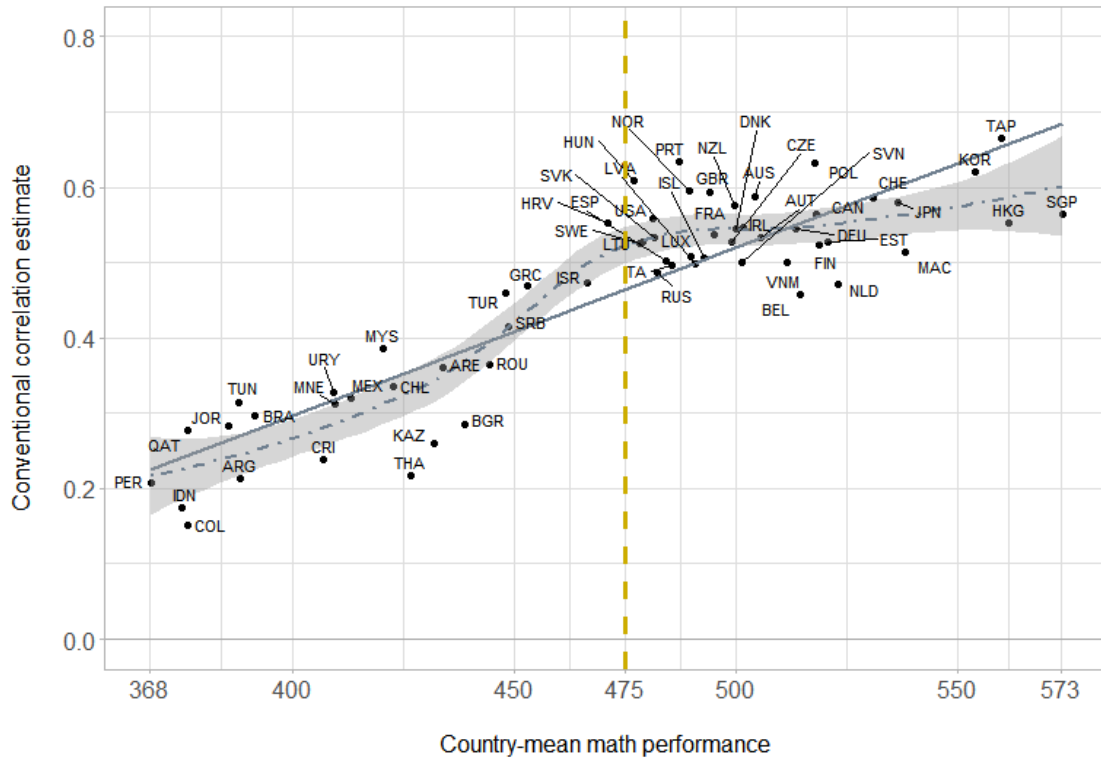


Figure 4.3. Relationship between within-country conventional correlations and country-mean math proficiencies

As seen from the graph above, conventional correlation estimates between math proficiency and math self-efficacy were larger and more similar in countries (e.g., France, Canada, and Finland) with mean-math performance above 475. In order to more closely examine to what extent country-mean math performance moderates the within-country correlations between proficiency and math self-efficacy, a score of 475 was set as a cutoff for country-mean math performance. Pearson correlations between country-mean math performances and within-country correlation estimates between math self-efficacy and math proficiency were examined separately for countries above and below the cutoff.

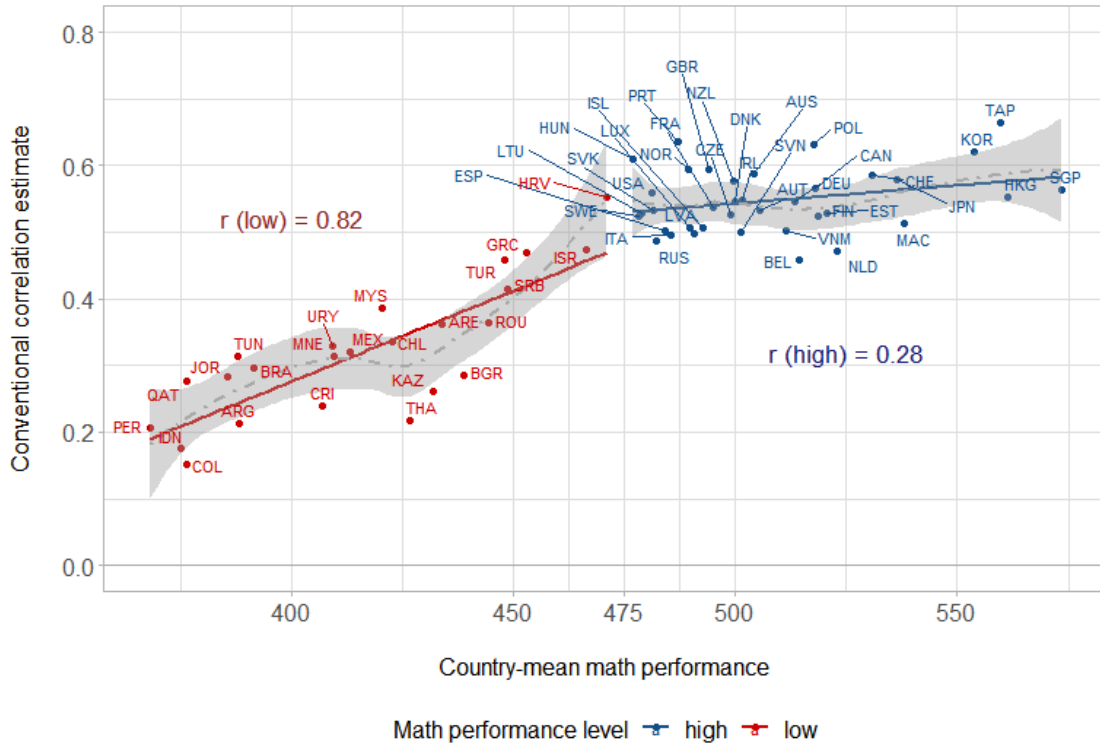


Figure 4.4. Relationship between within-country conventional correlations and country-mean math proficiencies

Figure 4.4 above displays that conventional correlation estimates varied substantially in lower achieving countries and there was a strong correlation ($r_{(low)} = 0.82$) between country-mean math performances and within-country correlation estimates. However, for countries with mean math performance higher than 475, this relationship was considerably weaker ($r_{(high)} = 0.28$).

4.2 Phase 2: Refinement with Multidimensional IRT Modeling

In the second phase of the analyses, multidimensional IRT modeling was used to account for the measurement error in the data and enable estimates of correlations on the latent variable level. Two-dimensional IRT models were fit to the item response data in a confirmatory manner;

that is, math proficiency items and math self-efficacy items were assigned to two separate dimensions in the estimation. As described in Chapter 3, correlations between latent traits are estimated jointly within the multidimensional IRT estimation and this corrects the estimates for measurement error. The *mdltm* software (von Davier, 2005b) provides directly estimated correlations among latent variables and these are named “skill distribution correlations” in the output.

In order to obtain within-country correlation estimates separately for each country in the dataset, a multi-group 2PL MIRT model was fit to the full dataset. To examine and ensure the model-data fit, the model was fit twice. First, item parameters were allowed in the models to differ for each country so that no measurement invariance was assumed. Secondly, item parameters were forced to be equal across countries with an assumption that the scales performed similarly across the countries irrespective of where students’ were from. The two models were then compared to examine if measurement invariance can be assumed for the data. Three model-fit indices were used as the criteria for model selection; Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Likelihood. The model-fit indices as well as the total number of parameters estimated in each model are presented below.

Table 4.2. Model fit indices for the MIRT models

	<u>AIC</u>	<u>BIC</u>	<u>Likelihood</u>	<u>Number of parameters</u>
Country-specific parameters	2650915	2772274	-1312030	13,428
Equal parameters	2620723	2627501	-1309611	750

For all three model-fit indices, smaller values indicate better goodness of fit for the model and the data. As seen in the table above, all three model-fit indices decreased in size only slightly

in the more constrained model with equal item parameters across countries. Although a less constrained model should provide a better fit to the data, model-fit indices penalize complexity to avoid overfitting. The model with country-specific parameters had 13,428 free parameters to be estimated, substantially less parsimonious in comparison to the model with equal parameters. As a result, the model-fit indices suggested an improvement, albeit small, in the model when the equal parameter constraints were imposed. For instance, consistent with this argument, BIC penalizes model complexity more heavily and the decrease in the BIC values were larger in comparison to AIC and Likelihood. Therefore, the multi-group MIRT model was fit to the data by keeping the parameters equal across the countries. Models converged with no issues and estimated item parameters are presented in Appendix, Table A.8.

Overall, within-country correlation estimates from the MIRT models were larger than the conventional estimates. The interquartile range of the MIRT estimates was found to be slightly narrower than the conventional estimates (0.195 and 0.216, respectively). However, correlation estimates still displayed substantial variability across countries ranging from 0.22 (Colombia) to 0.71 (Portugal). Descriptive statistics of the MIRT correlation estimates are presented below.

Table 4.3. Descriptive statistics for within-country correlation estimates from MIRT models

<u>Estimates</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
Conventional	0.458	0.133	0.151	0.501	0.664	0.514	-0.682	0.216
MIRT	0.530	0.130	0.224	0.576	0.712	0.488	-0.758	0.195

As seen in Figure 4.5 below, the histogram of the correlation estimates were shifted to the right demonstrating larger values in comparison to the conventional correlation estimates. On the other hand, the shapes of the two distributions were similar, displaying bimodality and negative

skewness. This indicates that the changes in the estimates from the OLS models to the MIRT models were similar across countries in the dataset.

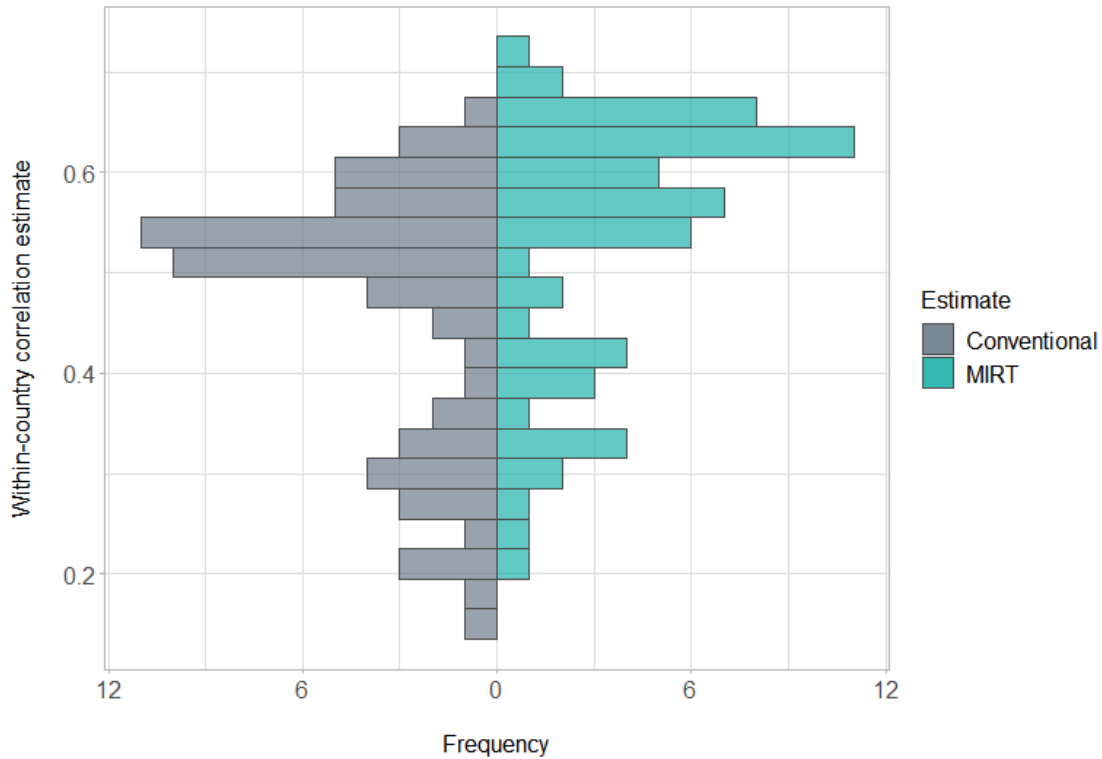


Figure 4.5. Distribution of the within-country conventional correlations and MIRT correlations

In order to ascertain that the patterns of relationships between math proficiency and math self-efficacy were comparable across models regardless of employing different sources of data (i.e., PVs vs. item response data) for the cognitive outcome, EAPs obtained from the MIRT models were compared to the first plausible value of math proficiency. The correlations between the first plausible values and the EAPs ranged from 0.78 to 0.93 with a median of 0.89 across countries. In addition, the association between the changes in the correlation estimates from OLS to MIRT and the correlations between PVs and EAPs was rather weak (-0.16). Therefore, it was

concluded that comparisons across models were not compromised by the use of different sources of data for the cognitive outcome.

The relationship between country-mean math performances and correlation estimates was also similar to the results found in Phase 1. The relationship was still stronger for lower achieving countries but slightly weakened after taking measurement error into account by fitting MIRT models in comparison to conventional estimates (the slopes were 0.82 and 0.77, respectively). In higher achieving countries, the change in the relationship between country-mean math performance and the correlation estimates between math proficiency and math self-efficacy was relatively more noticeable. The relationship was weakened considerably in higher achieving countries in comparison to conventional estimates (the slopes were 0.05 and 0.28, respectively) and the slope was effectively zero.

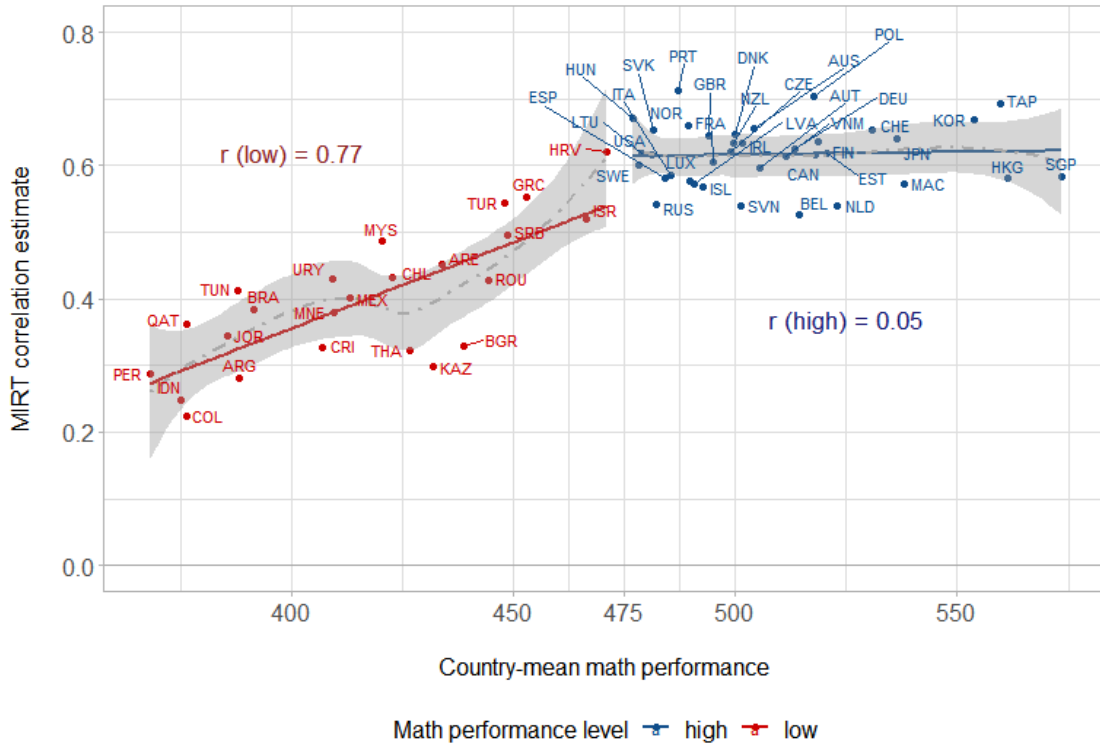


Figure 4.6. Relationship between within-country MIRT correlations and country-mean math proficiencies

With the use of MIRT models, the correlation estimates are corrected for the measurement error in the data so the changes in the estimates from ordinary regression models to MIRT models are expected. It was hypothesized that the changes in the correlation estimates from conventional models to MIRT models would be larger for countries with greater measurement error in the data. Although measurement error is not forced to be constant along the scale and a scale's reliability is conditional on the ability estimate in the IRT framework, indices that could summarize the quality of estimation for the entire scale were needed to make such a comparison. To that end, EAP reliability (Adams, 2005) was used to examine the amount of measurement error in the data. For each country in the dataset, two EAP reliability indices were calculated; one for math self-efficacy and another for math proficiency scale. IRT-based

reliability indices are expected to be more accurate than conventional reliability estimates such as Cronbach's alpha.

As shown in Figure 4.7, the empirical reliability indices for math scale from the MIRT models were found to be considerably heterogeneous ranging from 0.69 (Indonesia) to 0.87 (Chinese Taipei). The empirical reliability indices for the math self-efficacy scale from the MIRT models, on the other hand, were found to be larger and homogenous ranging from 0.77 (Viet Nam) to 0.86 (Korea) (Figure 4.8).

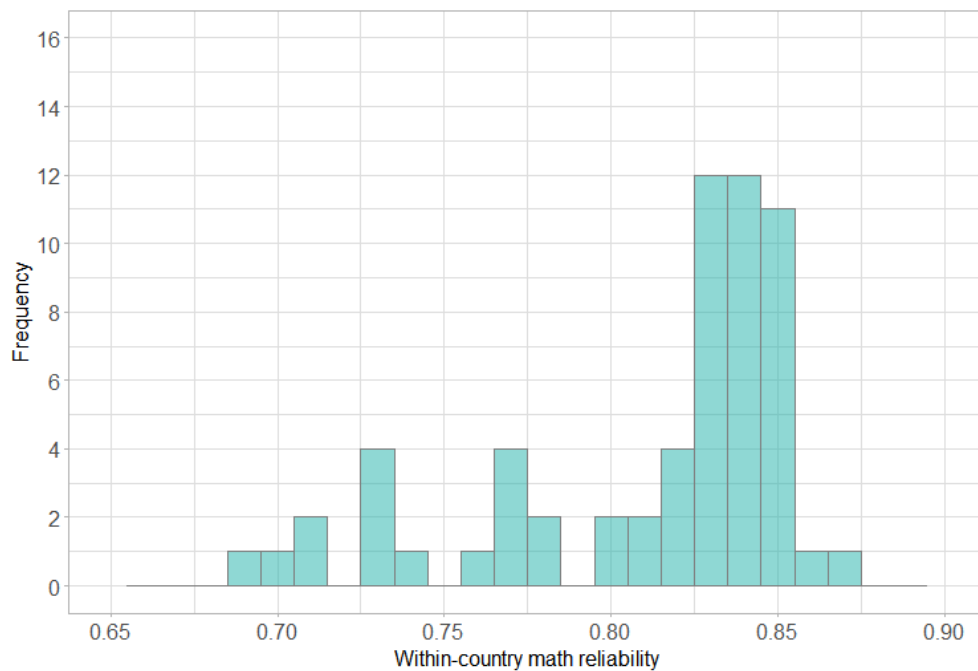


Figure 4.7. EAP reliabilities gathered from the MIRT models for the math scale

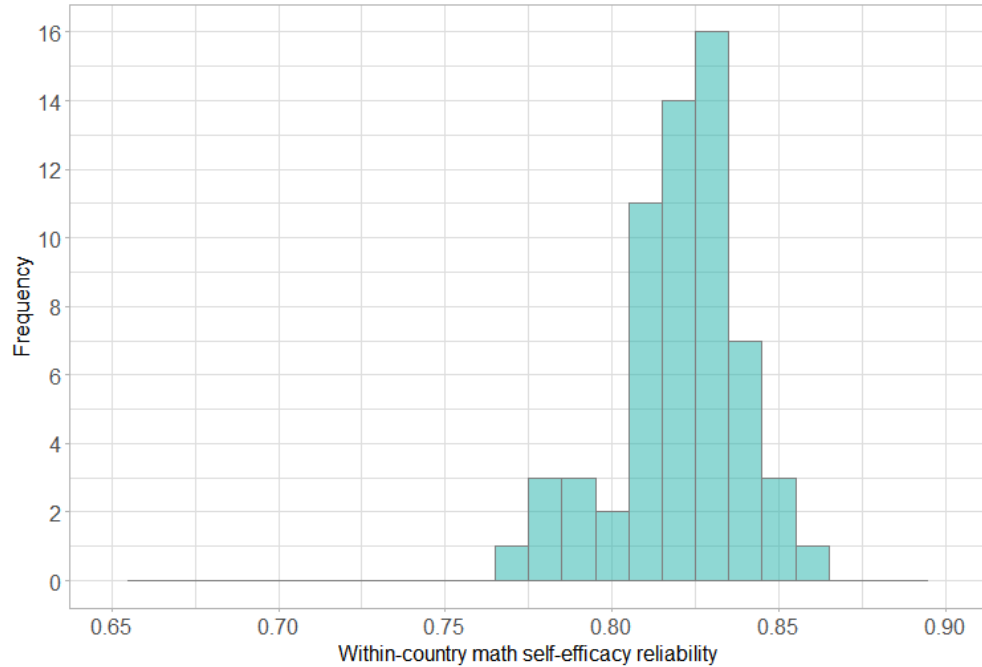


Figure 4.8. EAP reliabilities gathered from the MIRT models for the math self-efficacy scale

The greater heterogeneity in the EAP reliabilities for the math scale suggested that the amount of measurement error in the math proficiency data varied considerably across countries. In order to better understand the changes in the estimates and their relationship to the amount of measurement error in the data, country-level EAP reliabilities for the math scale were examined in relation to the changes in the correlation estimates from Phase 1 to Phase 2. Because measurement error is a major cause of attenuation in the correlation estimates, greater changes were expected to be seen for the countries with larger measurement error and, thus, with lower reliabilities. However, demonstrated in the figure below, there is weak to no association between the changes in the correlation estimates and the EAP reliabilities for the math scale obtained from the MIRT models.

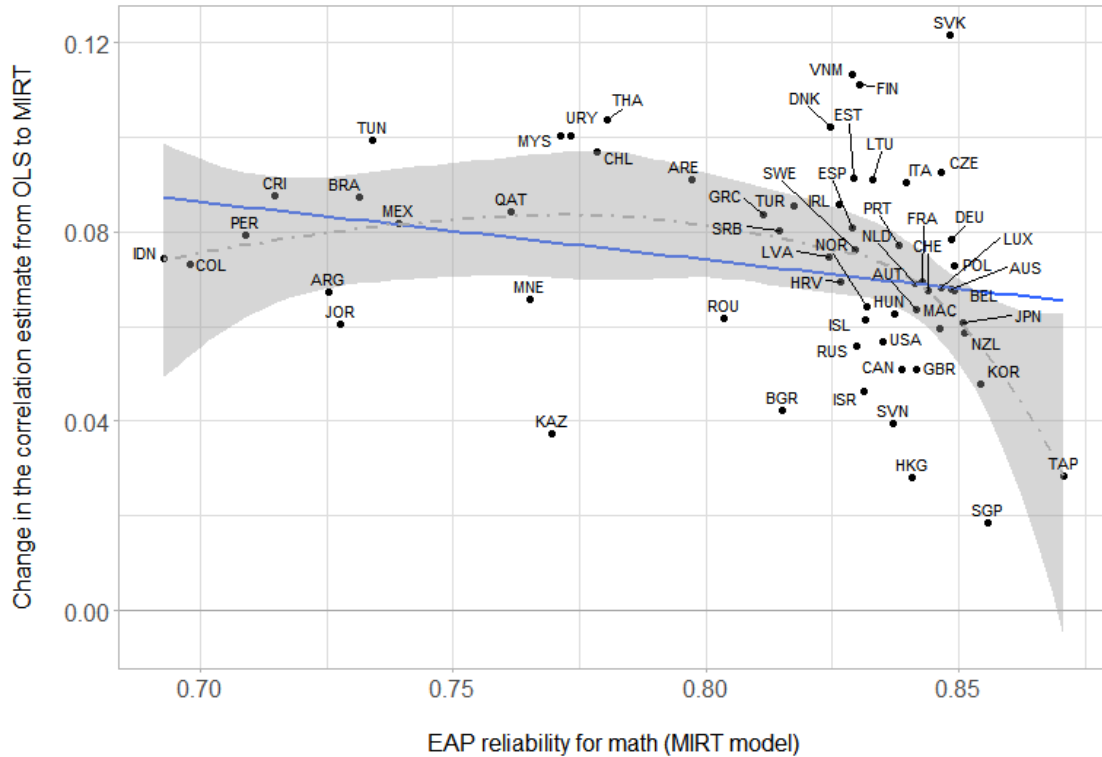


Figure 4.9. Relationship between EAP reliabilities for math proficiency from MIRT models and the changes in the correlation estimates from conventional to MIRT models

Although almost no association was found between EAP reliabilities for math proficiency measure and the changes in the correlation estimates between math proficiency and math self-efficacy from conventional analyses to MIRT model, it was expected that math proficiency scale carried larger amount of measurement error in lower achieving countries due to reasons such as higher tendencies of straightlining behavior. Therefore, the EAP reliabilities for the math proficiency scale were investigated in relation to country-mean math proficiencies. The distribution below demonstrates that the EAP reliabilities for the math proficiency scale for countries with average math proficiency score below 475 were much weaker and more heterogeneous. The math scale carried a smaller amount of measurement error in higher achieving countries.

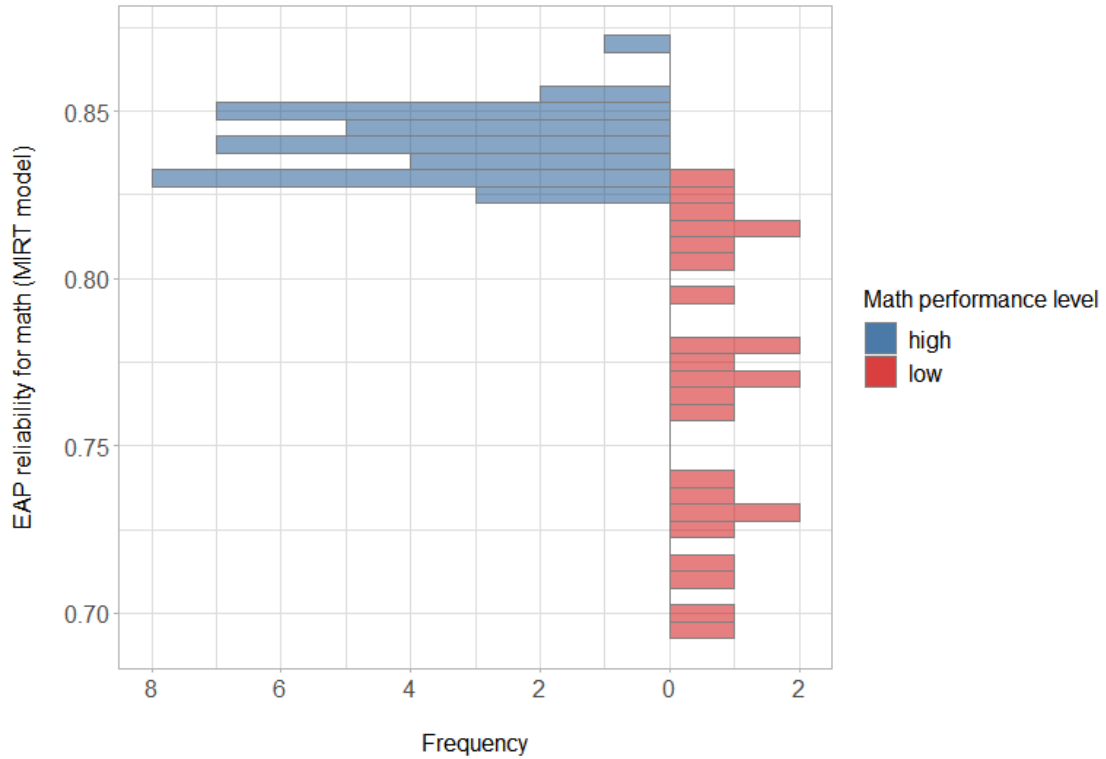


Figure 4.10. Distribution of the EAP reliabilities gathered from the MIRT models for the math scale by country math performance level

The scatterplot below also demonstrates that there was a strong relationship between within-country EAP reliability indices for math proficiency scale gathered from the MIRT models and country-mean math proficiencies in lower achieving countries ($r_{\text{low}} = 0.92$). The strong relationship between empirical reliability estimates and math proficiency in lower achieving countries was found to be positive and linear. That is, the reliability of the math proficiency scale was lower for more poorly performing countries. A similar, but somewhat weaker, relationship was observed in higher achieving countries ($r_{\text{high}} = 0.58$). It should also be noted that EAP reliabilities for the math proficiency measure were more homogeneously distributed in higher achieving countries, ranging between 0.83 (Latvia) to 0.87 (Chinese Taipei).

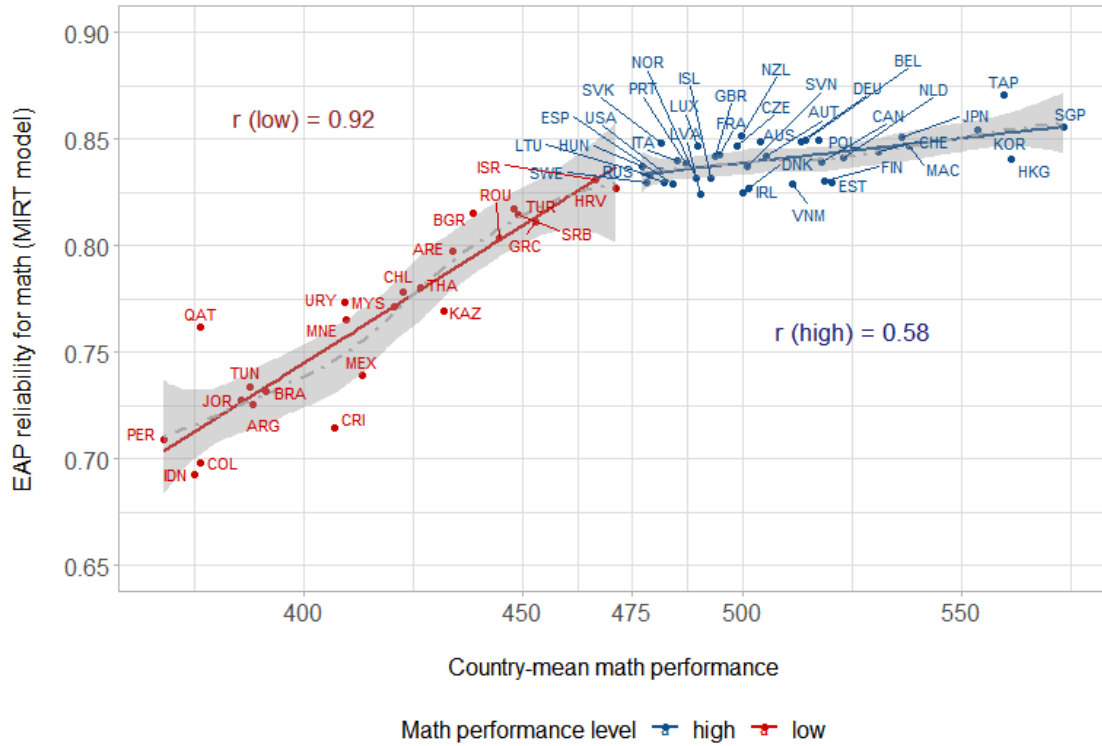


Figure 4.11. Relationship between empirical reliability indices for math proficiency from MIRT models and country-mean math proficiencies

Even though the EAP reliabilities for the math self-efficacy scale did not vary substantially across the countries, a similar investigation was followed to examine whether there was a relationship between the reliabilities and country-mean math performance. In contrast to the EAP reliabilities for the math scale, the distribution of the EAP reliabilities for the math self-efficacy scale was more heterogeneous in higher achieving countries.

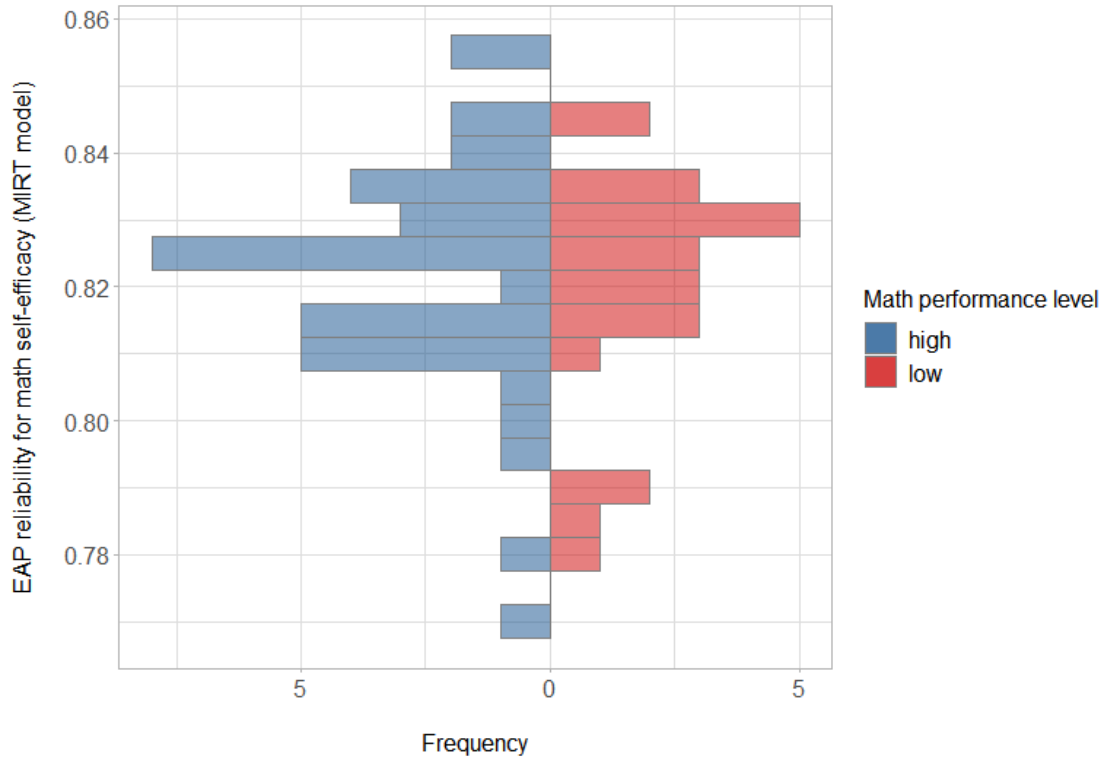


Figure 4.12. Distribution of the EAP reliabilities gathered from the MIRT models for the math self-efficacy scale by country math performance level

As presented in the scatterplot below (Figure 4.13), the relationship between within-country EAP reliabilities for math self-efficacy scale and country-mean math proficiencies in lower achieving countries was found to be positive but weaker ($r_{(low)} = 0.24$) and somewhat non-linear. In higher achieving countries, math self-efficacy scale reliabilities were much more heterogeneous in comparison to the math scale reliabilities. Note that Singapore, which showed one of the highest reliabilities for the math scale, demonstrated the second lowest reliability (0.78) for math self-efficacy scale among all the countries in the dataset after Vietnam (0.77). With Singapore having the highest country-mean math proficiency across all countries but showing the second lowest reliability for math self-efficacy scale, the relationship between within-country math self-efficacy reliabilities and country-mean math proficiencies was found to be negative but effectively zero ($r_{(high)} = -0.02$) in higher achieving countries.

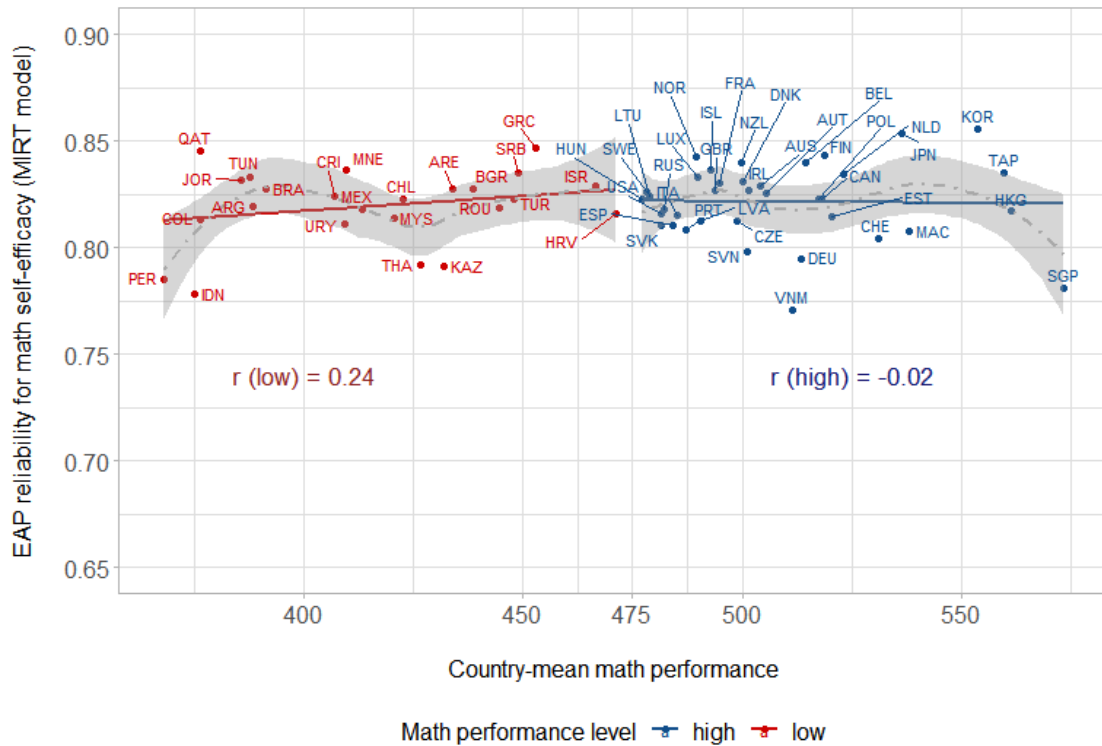


Figure 4.13. Relationship between empirical reliability indices for math self-efficacy from MIRT models and country-mean math proficiencies

Within-country EAP reliabilities for the math self-efficacy scale were also examined with respect to their association with county-mean math self-efficacy indices. Presented in Figure 4.14, a strong and negative relationship was observed between country-mean math self-efficacy and within-country math self-efficacy scale reliabilities in higher achieving countries ($r_{(\text{high})} = -0.55$). For example, in Japan and Korea, students' showed the lowest level of math self-efficacy (-0.41 and -0.36, respectively) across all the countries participated in PISA 2012 but showed the highest empirical reliabilities for the math self-efficacy scale (0.85 and 0.86, respectively). These two countries were among the higher performing countries at math with mean math proficiency scores of 536 and 554 respectively. Consistent with these results, students from Singapore showed the highest level of math self-efficacy (0.50) but the EAP reliability from the MIRT

model was the second lowest (0.78) among all countries in the dataset, same as Indonesia (0.78) after Vietnam (0.77). In countries with lower performance in math, the pattern of relationship between math self-efficacy reliabilities and mean math self-efficacy indices was almost flat but slightly positive (0.04).

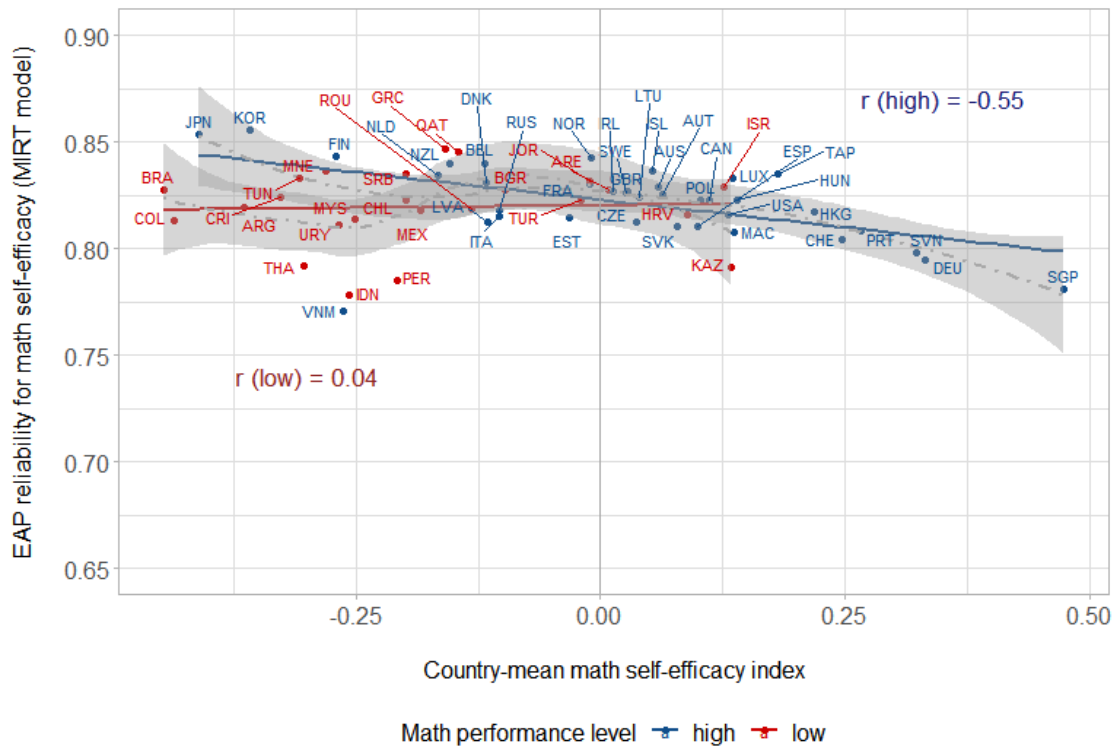


Figure 4.14. Relationship between empirical reliability indices for math self-efficacy from MIRT models and country-mean math self-efficacy indices

In order to further investigate potential factors behind the differences in MIRT correlation estimates across countries, correlation estimates were plotted against empirical reliabilities obtained from the MIRT models. Figure 4.15 below displays the relationship between within-country reliabilities for the math scale and the MIRT correlation estimates. Both MIRT estimates and empirical reliabilities for math scale were smaller for lower achieving countries in comparison to higher achieving countries, and MIRT estimates were typically larger when the

estimated reliabilities for the math scale were greater ($r_{(low)} = 0.78$). Although the distribution of MIRT estimates was more homogenous in higher achieving countries and the range of the reliabilities was narrower, a positive, albeit weaker, relationship was observed between the correlation estimates and the reliabilities for the math scale ($r_{(high)} = 0.25$).

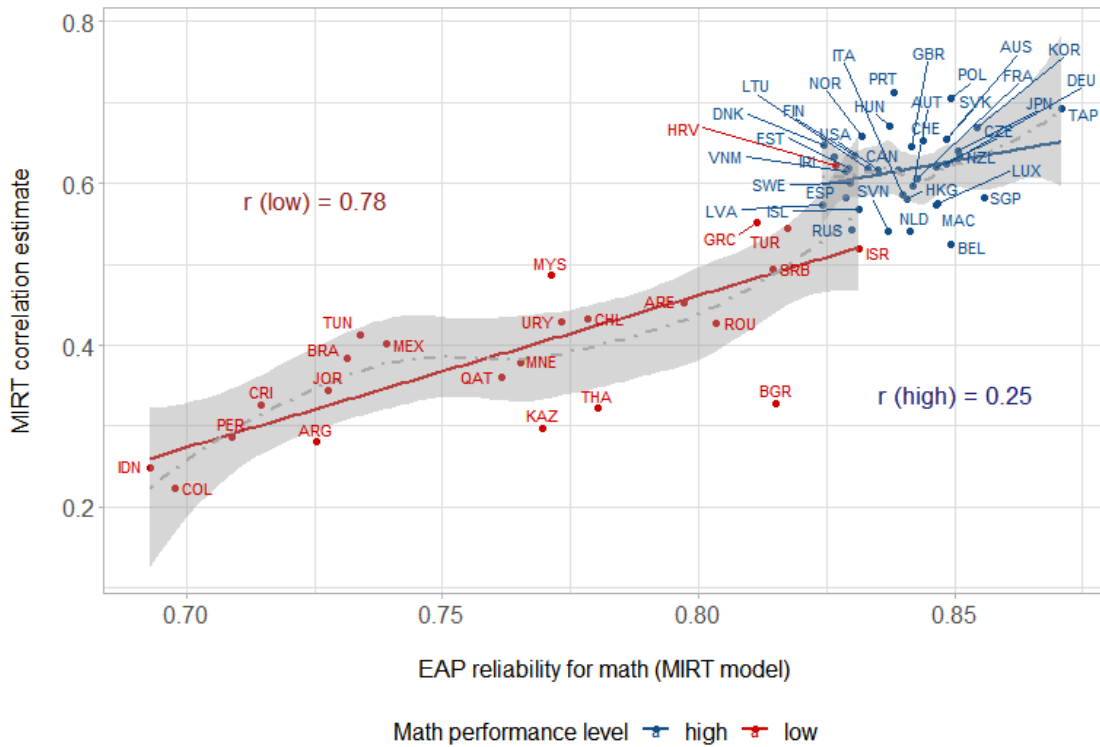


Figure 4.15. Relationship between empirical reliability indices for math from MIRT models and MIRT correlation estimates

The relationship between the MIRT correlation estimates and the reliabilities for the math self-efficacy scale was similar but weaker for both higher and lower achieving countries. Although the distributions of the reliabilities for the math self-efficacy scale covered almost the same range in both higher and lower achieving countries, the MIRT estimates for the lower achieving countries were typically larger when the estimated reliabilities for the math self-

efficacy scale were greater ($r_{(low)} = 0.47$). The slope was almost flat in higher achieving countries but still positive ($r_{(high)} = 0.16$).

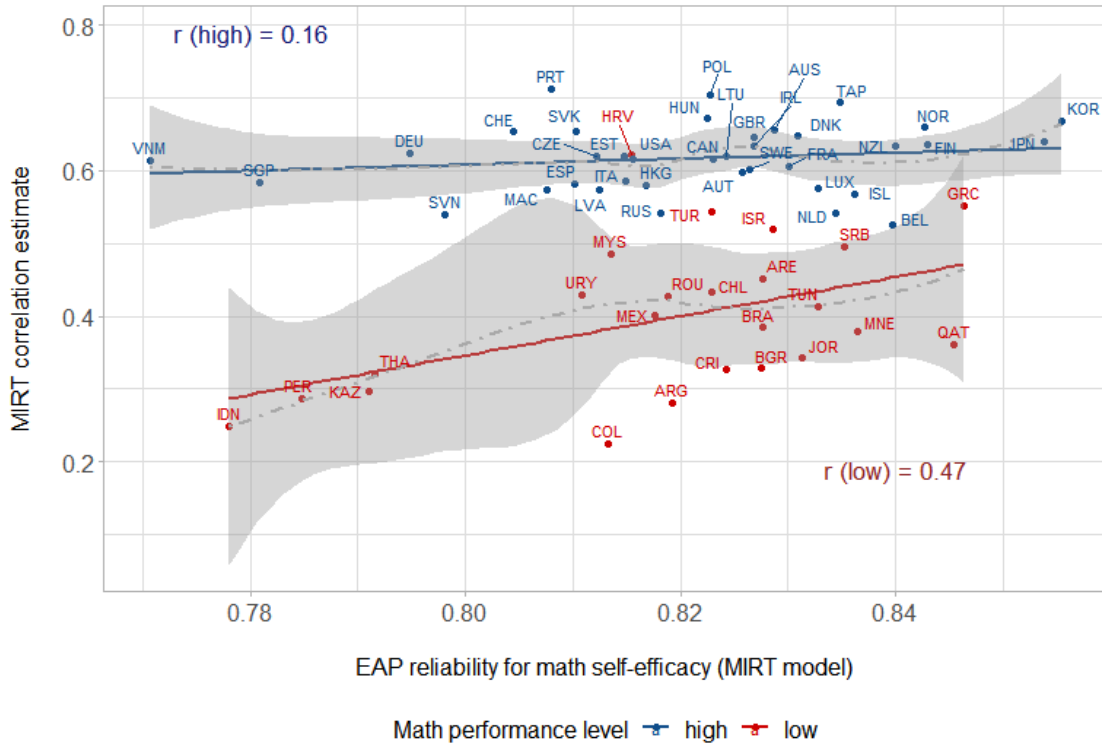


Figure 4.16. Relationship between empirical reliability indices for math self-efficacy from MIRT models and MIRT correlation estimates

4.3 Phase 3: Refinement with Multilevel Linear Modeling

In the next phase of the model refinement process, clustering of the data resulting from the PISA survey design was taken into account by employing two-level linear modeling. When data are collected from individuals who are clustered in groups by design or by the nature of the research context, it is critical to account for this nested design in statistical analyses. Although the PISA research design suggests such a hierarchical structure in the data due to the multi-stage sampling technique used to obtain representative samples of students from sampled schools in

participating countries, one has to look for evidence of nesting in the data first to ensure that multilevel modeling is an appropriate choice. To this end, sixty-one sets of unconditional two-level linear models were fit to the data (one for each country) to investigate the proportion of total variance in students' math performance that could be attributed to the variability among schools within the country (i.e., intra-class correlation or ICC). Note that no predictors were included in unconditional models but separate models for each level in the data (e.g., school- and student-level) allowed for examination of the residuals separately at each level.

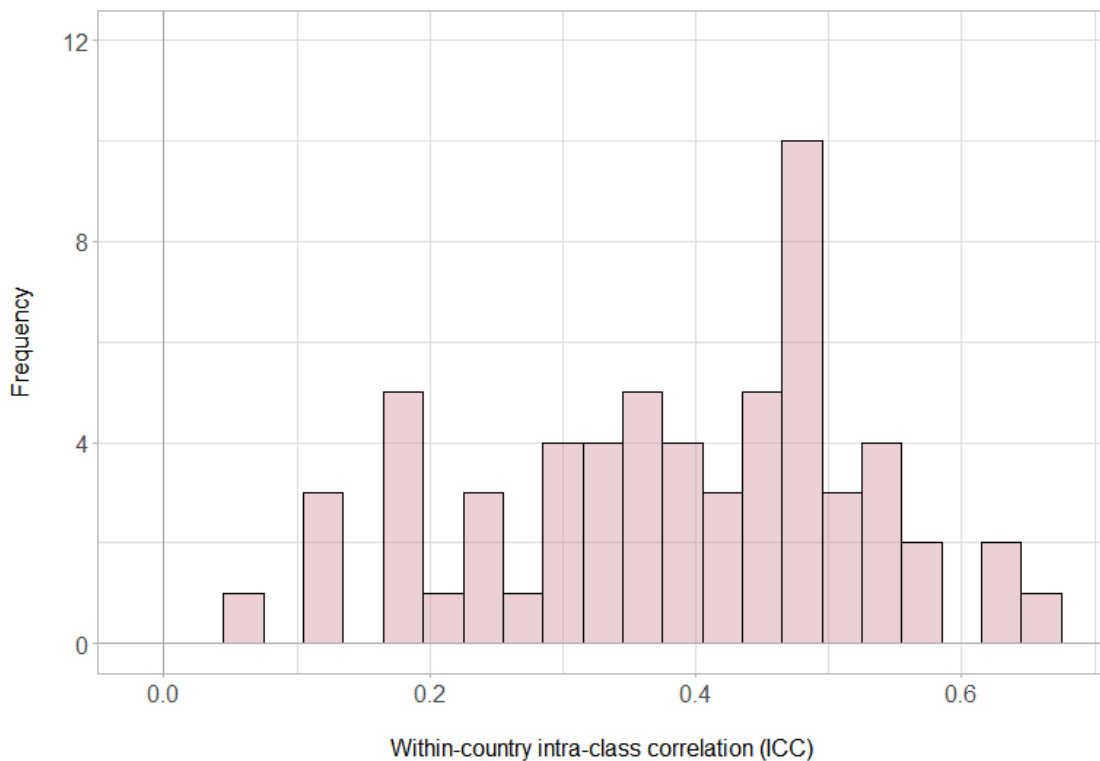


Figure 4.17. Distribution of the within-country ICCs for the math performance outcome

As shown in Figure 4.17, within-country ICCs varied substantially across participating countries. Finland displayed the smallest intra-class correlation of 0.07, indicating that only 7% of the variance among students' math performance could be explained by school-related factors instead of individual differences. Data from Hungary showed the highest ICC of 0.65 which

indicates that 65% of the variance in students' math performance were due to the factors related to schools and not students' individual differences. Overall, the ICCs were larger than 15% in most of the countries. Therefore, multilevel modeling was an appropriate model choice.

For the next step of this phase, a second set of two-level linear models in which student math self-efficacy index was added as a covariate at the student-level was fit to the data from each of the sixty-one participating countries. The distribution of standardized regression coefficients as correlation estimates of the within-country relationships between math self-efficacy and math proficiency were examined.

Overall, within-country correlation estimates from the MLM models were smaller than both the conventional estimates and the MIRT estimates. The interquartile range of the MLM estimates was found to be slightly larger than the one obtained from the conventional estimates (0.224). Correlation estimates still showed substantial variability across the countries.

Descriptive statistics of the MLM correlation estimates are presented below.

Table 4.4. Descriptive statistics for within-country correlation estimates from MLM models

<u>Estimates</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
Conventional	0.458	0.133	0.151	0.501	0.664	0.514	-0.682	0.216
MIRT	0.530	0.130	0.224	0.576	0.712	0.488	-0.758	0.195
MLM	0.350	0.132	0.105	0.371	0.583	0.477	-0.106	0.224

As seen in the histogram below, the correlation estimates were shifted to the left displaying smaller values in comparison to the conventional correlation estimates. Note that this is the reverse of the shift observed with the MIRT models. Moreover, the shape of the distribution was slightly different. For example, the bimodality that was seen in Figure 4.1

almost disappeared and the distribution was not skewed towards larger values as seen in Figure 4.18.

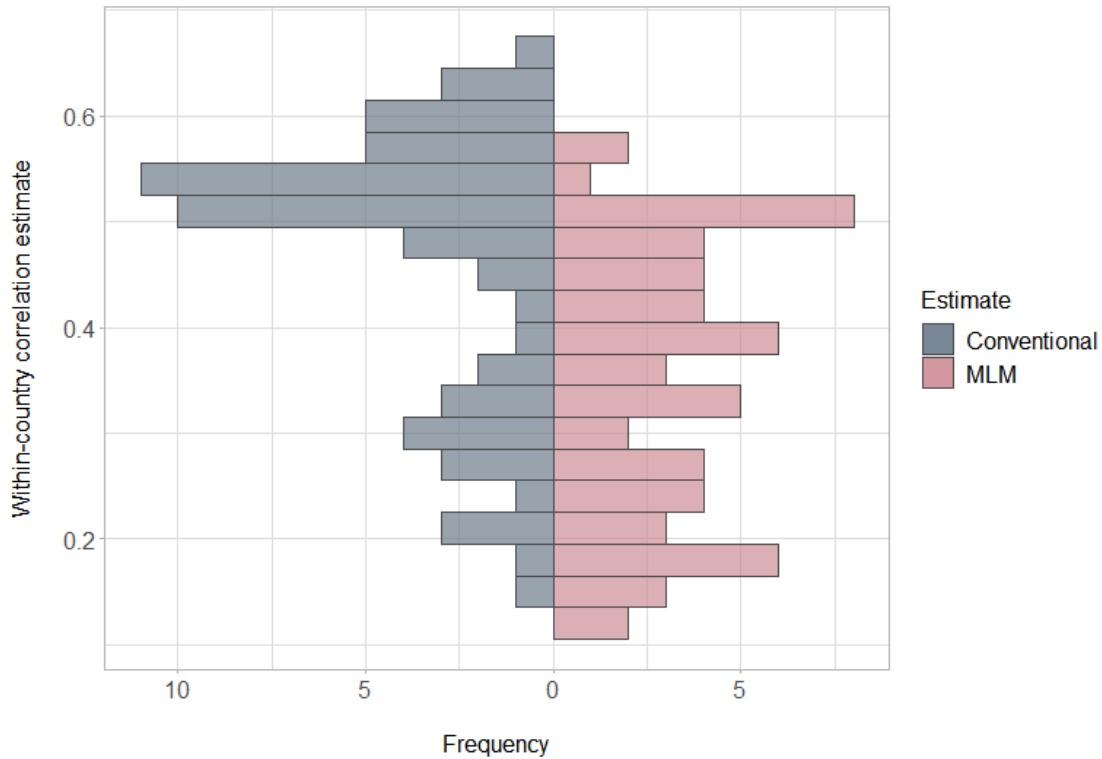


Figure 4.18. Distribution of the within-country conventional and MLM correlations

Similar to the previous phases of model refinement, the relationship between the correlation estimates and country-mean math proficiencies was examined. The relationship was still stronger for lower achieving countries ($r_{(low)} = 0.71$) but slightly weakened in comparison to conventional and MIRT estimates (the slopes were 0.82 and 0.77, respectively). In higher achieving countries, the slope of the relationship with country-mean math performance was effectively zero ($r_{(low)} = -0.04$), similar to the one for the MIRT estimates.

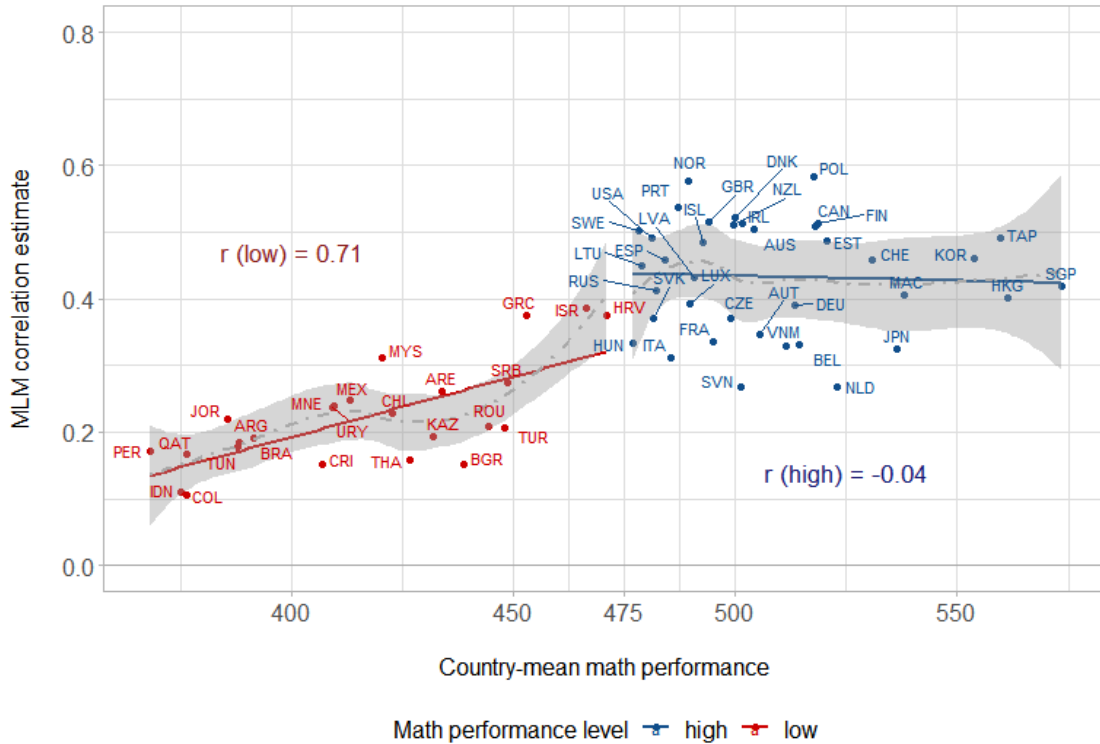


Figure 4.19. Relationship between within-country MLM correlations and country-mean math proficiencies

With multilevel linear modeling, the changes in the correlation estimates were due to taking the clustering in the data into account. Therefore, it was reasonable to expect a strong relationship between within-country ICCs and the changes in the estimates from conventional analyses to MLM. As plotted below, the changes were larger for countries with large ICCs. Furthermore, the relationship was approximately linear with a negative slope (-0.79). That is, the correlation estimates got smaller as the ICC got larger, as is to be expected. For example, the correlation between students' math performance and math self-efficacy almost stayed the same for Finland, as it had the smallest ICC across all countries in the dataset. By contrast, Hungary, with the largest ICC, had the largest change in the correlation estimate from conventional

analyses to MLM (a decrease of 0.27). This suggests that for students from Hungary, the school a student attended made a big difference in their math performance.

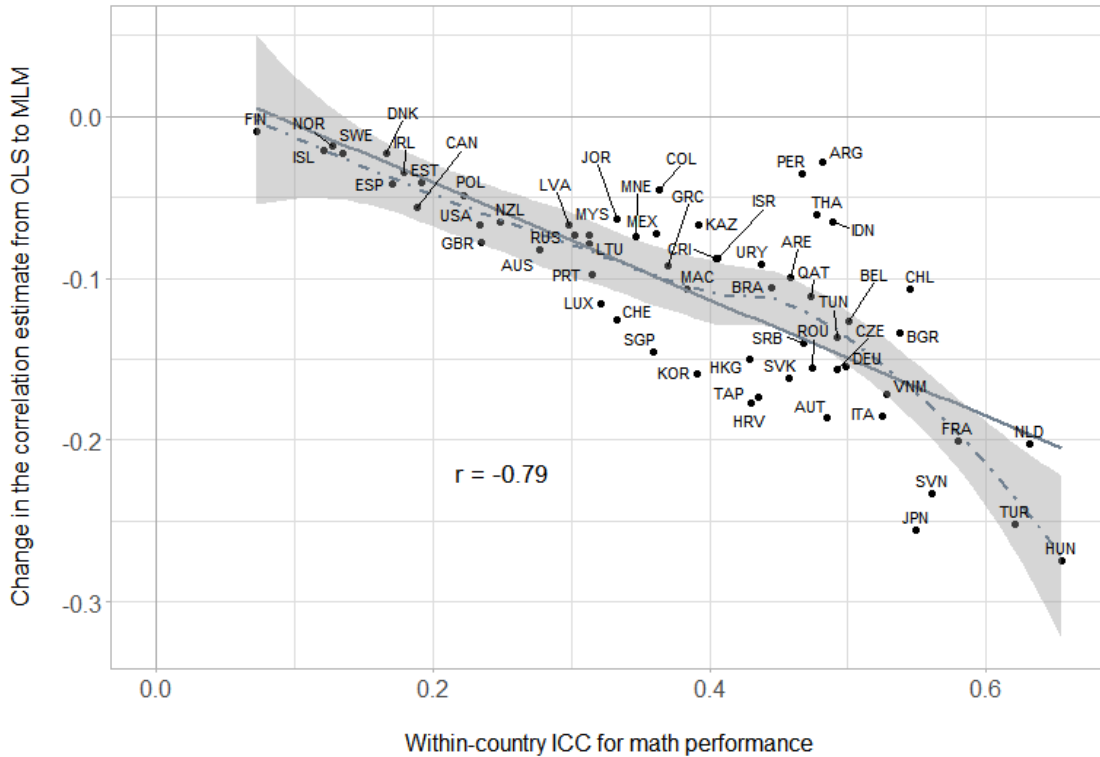


Figure 4.20. Relationship between changes in the correlation estimates from OLS to MLM and within-country intra-class correlations for math proficiency outcome

4.4 Phase 4: Refinement with Multilevel Multidimensional Mixture IRT Modeling

As the final step of the model refinement process, multilevel two-dimensional mixture IRT models were fit to the data from each country in the dataset to account for both measurement error and clustering. Similar to the MIRT models, the models were conducted in a confirmatory manner to incorporate the multidimensionality. The 117 x 2 Q-matrix that was developed in Phase 2 to fit the MIRT models was used to assign each item to one of the two dimensions. The models were constrained to have two latent classes and no measurement

invariance was assumed between the two latent classes. Moreover, the models were fit separately for each country and, thus, no measurement invariance was assumed across countries either. As a result, class-specific item parameter sets were estimated separately and allowed to differ from country to country. Similar to the MIRT models, students' underlying latent traits were estimated by using *expected a posteriori* (EAP) and standardized ability estimates were obtained directly from the *mdltm* output. There were no convergence issues.

In mixture-distribution IRT modeling, the models classify individuals into the pre-defined number of latent classes based on maximizing within-class homogeneity and between-class heterogeneity in their responses to the items. That is, similarities and differences among individuals are taken into account without knowing what leads to those differences. In cases where the researcher has an underlying theory for the research context, the models can be interpreted in a confirmatory way to examine whether the data supports the hypotheses regarding the underlying subpopulations. A similar approach was followed in this study. Although the underlying reasons were not known, it was hypothesized that the students would be split into two latent classes based primarily on their math performance level. The results from the MLMixMIRT models supported this hypothesis. Consistently for each country, the models yielded two latent classes that differed from each other by substantial differences in mean EAP math ability estimates.

Table 4.5. Descriptive statistics for the within-country mean EAP math ability estimates by latent classes

<u>Latent Class</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
High Math Class	-0.080	0.465	-1.087	0.007	0.944	2.031	-0.207	0.619
Low Math Class	-0.972	0.564	-2.467	-0.914	0.246	2.714	-0.127	0.746

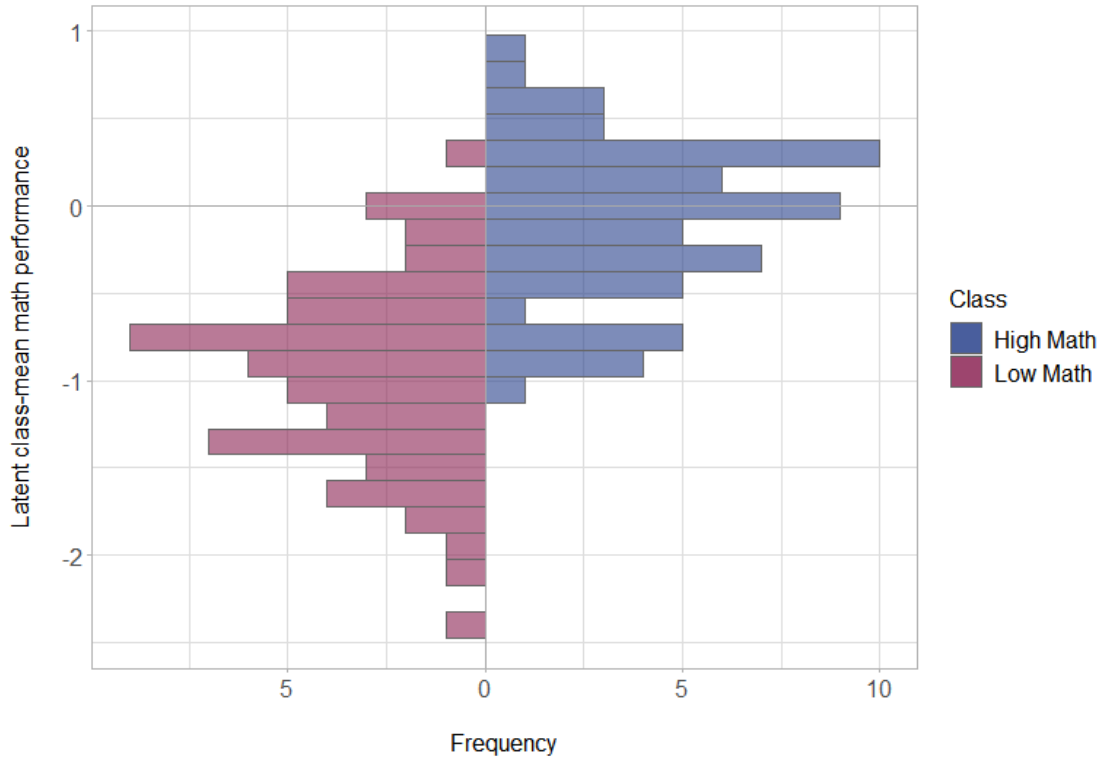


Figure 4.21. Distribution of within-country mean math performance by latent class

As can be seen from the descriptive statistics given in Table 4.5, one of the latent classes from the MLMixMIRT models displayed consistently lower math performance ranging from -2.47 (Hungary) to 0.25 (Chinese Taipei) than the other latent class with ranged from -1.09 (Indonesia) to 0.94 (Singapore) within each country. Moreover, the latent classes with higher mean math performance were more homogenous than the latent class with lower math performance estimates. Figure 4.21 demonstrates the distribution of the two latent classes by mean math ability estimates. Although the two distributions overlapped, the distinction between two latent classes was clear. Based on these results, two estimated latent classes from the MLMixMIRT models were labelled as “High Math Class” and “Low Math Class” (or “MLMixMIRT_High” and “MLMixMIRT_Low” when results from different models were

compared) and the rest of the results from the analyses of the MLMixMIRT models were interpreted accordingly.

The correlations between math proficiency and math self-efficacy are estimated jointly within the MLMixMIRT estimation. Because there were two latent classes defined in the model, correlations between the two latent traits were estimated separately for each latent class within this modeling framework. As a result, two correlation estimates were obtained for each country. The distributions of the correlation estimates for both latent classes were examined in comparison to those that were obtained from other models employed in previous stages of the study. The correlations between math proficiency and math self-efficacy varied substantially for the Low Math Class across participating countries. On the other hand, the distribution of the correlation estimates was much more homogeneous for the High Math Class. Descriptive statistics of the MLMixMIRT correlation estimates for each latent class, as well as the histograms for the distributions, are presented below along with those that were obtained in previous stages of the model refinement process.

Table 4.6. Descriptive statistics for within-country correlation estimates from MLMixMIRT models

<u>Estimates</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
Conventional	0.458	0.133	0.151	0.501	0.664	0.514	-0.682	0.216
MIRT	0.530	0.130	0.224	0.576	0.712	0.488	-0.758	0.195
MLM	0.350	0.132	0.105	0.371	0.583	0.477	-0.106	0.224
MLMixMIRT								
<i>High Math Class</i>	0.560	0.089	0.319	0.571	0.733	0.414	-0.431	0.109
<i>Low Math Class</i>	0.247	0.221	-0.106	0.254	0.700	0.806	0.108	0.368

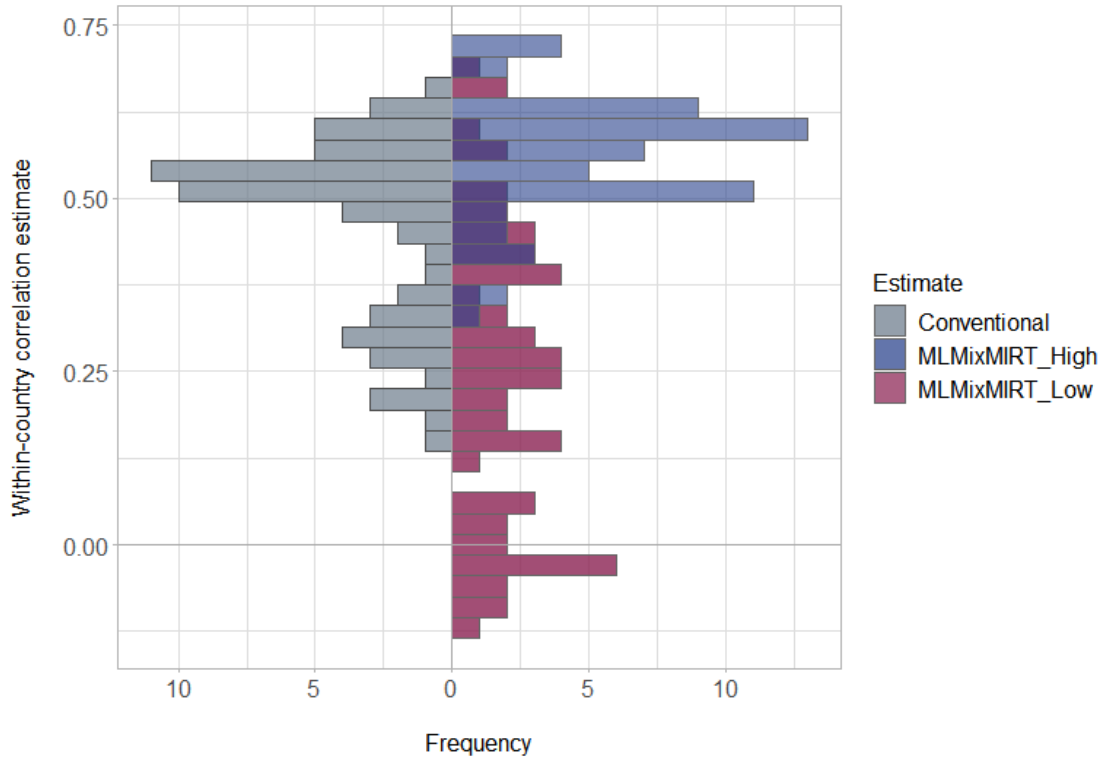


Figure 4.22. Distribution of the within-country conventional and MLMixMIRT correlations

The descriptive statistics suggest that correlation estimates for the High Math Class were considerably larger than those obtained for the Low Math Class. In several countries such as Bulgaria and Peru, students' math performance was found to be negatively correlated with their math self-efficacy level in the Low Math Class (the correlation estimates were -0.11 and -0.10, respectively). However, in certain countries, the correlation for the Low Math Class was as large as those obtained for the High Math Class in other countries. For instance, in Chinese Taipei, the correlation estimates for both latent classes were very similar (0.73 for High Math Class and 0.70 for Low Math Class) and were the highest across all countries. Note that the mean math ability estimate for the Low Math Class was almost equal to the one for the High Math Class in Chinese Taipei (0.25 and 0.27, respectively) and was the highest across all countries.

The correlation estimates for the High Math Class resulting from the MLMixMIRT models were substantially more homogeneous and larger (mean = 0.56) in comparison to those obtained from OLS, MIRT and MLM models. On the other hand, the correlation estimates for the Low Math Class varied substantially and were generally smaller (mean = 0.25) in comparison to the models employed in previous stages of the model refinement process.

Similar to the previous phases in the model refinement, the relationship between the correlation estimates and country-mean math proficiencies was examined. Different from the previous phases, the relationship with country-mean math proficiencies were examined separately for each latent class instead of using the cutoff for the mean math proficiency score of 475. Because the MLMixMIRT models yielded two latent classes and these latent classes were interpreted as high math proficiency and low math proficiency classes, it was considered to be more appropriate and simpler for the interpretation to examine the association between math performance level and the correlation estimates between math self-efficacy and math proficiency separately for each of the latent classes.

As plotted in the figure below, the relationship between country-mean math performance and the correlation estimates was approximately linear for both classes, but much stronger for the Low Math Class in comparison to the High Math Class (the slopes were 0.85 and 0.32, respectively).

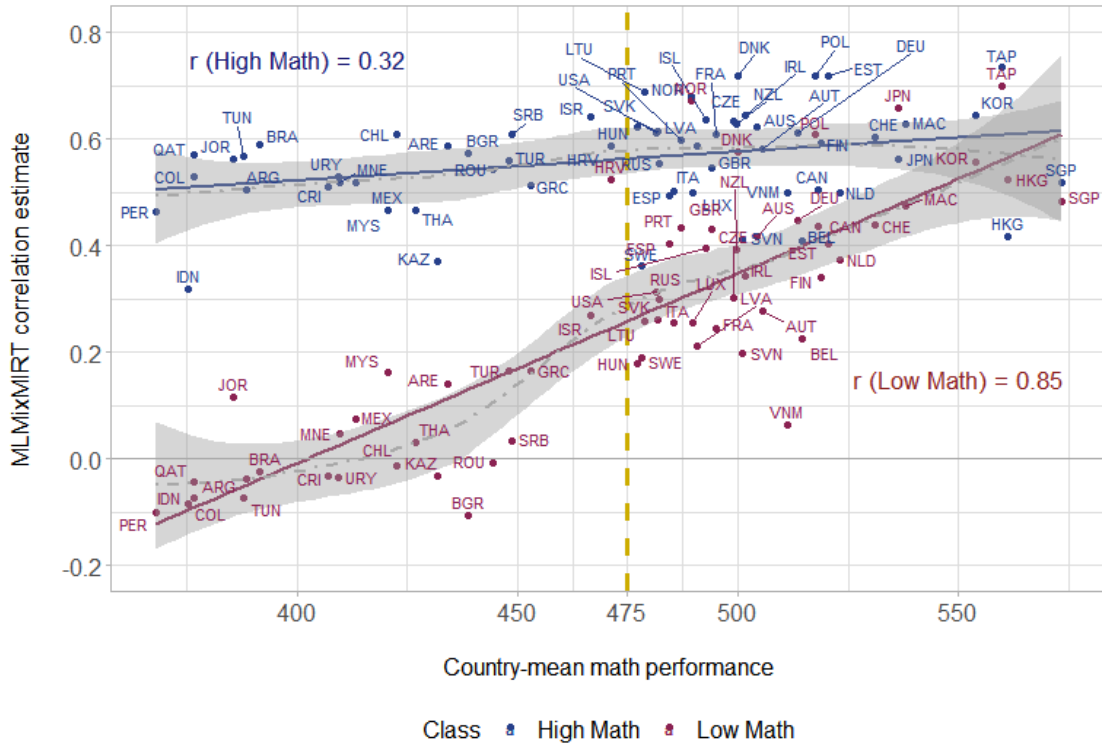


Figure 4.23. Relationship between within-country MLMixMIRT correlations and country-mean math performance

In order to make a comparison that is more analogous to the previous phases, the relationship between country mean math performances and the correlation estimates obtained from the MLMixMIRT models was examined more closely for each latent class by using the math proficiency score of 475 as a cutoff for country math performance level. The table below summarizes the relationships observed across all models employed in this study along with the sample sizes.

Table 4.7. Pearson correlation between within-country correlation estimates and country-level math performance

<u>Estimates</u>	<u>Math performance level based on the cutoff of 475</u>		<u>Overall (N=61)</u>
	<u>Low (N=24)</u>	<u>High (N=37)</u>	
Conventional	0.82	0.28	0.88
MIRT	0.77	0.05	0.85
MLM	0.71	-0.04	0.76
MLMixMIRT			
<i>High Math Class</i>	0.36	0.02	0.32
<i>Low Math Class</i>	0.68	0.56	0.85

The table above clearly demonstrates that when countries' math performance levels (based on 475 cutoff) were ignored, correlation estimates from the OLS, MLM, and MIRT models were strongly associated with country-mean math performance (the slopes were 0.88, 0.85, and 0.76, respectively). However, when examined more closely by splitting the countries into two groups based on their average math performance (below or above 475), it was observed that the relationship was different for low- and high-achieving countries (e.g., the slopes were 0.82 and 0.28, respectively, based on conventional analyses). For the first three models, this grouping had to be based on average math performance scores by using 475 as the cutoff because all students were assumed to belong to the same population (i.e., no subpopulations and no mixture distributions). When MLMixMIRT was employed, the results were consistent. Even in higher achieving countries, the Low Math Class yielded a stronger relationship between mean math performance and correlation estimates in comparison to the High Math Class (the slopes were 0.56 and 0.02, respectively).

When MLMixMIRT estimates were examined in relation to class-mean math ability estimates rather than overall country-mean math proficiencies, a somewhat similar pattern was observed. Displayed in Figure 4.24 below, class-mean math ability estimates were positively correlated with MLMixMIRT estimates for Low Math Class and the relationship was stronger and approximately linear ($r_{(\text{Low Math})} = 0.73$). In High Math Class, however, the relationship was almost curvilinear; the correlation estimates peaked around the class mean math proficiency of zero but were smaller on either side of the distribution.

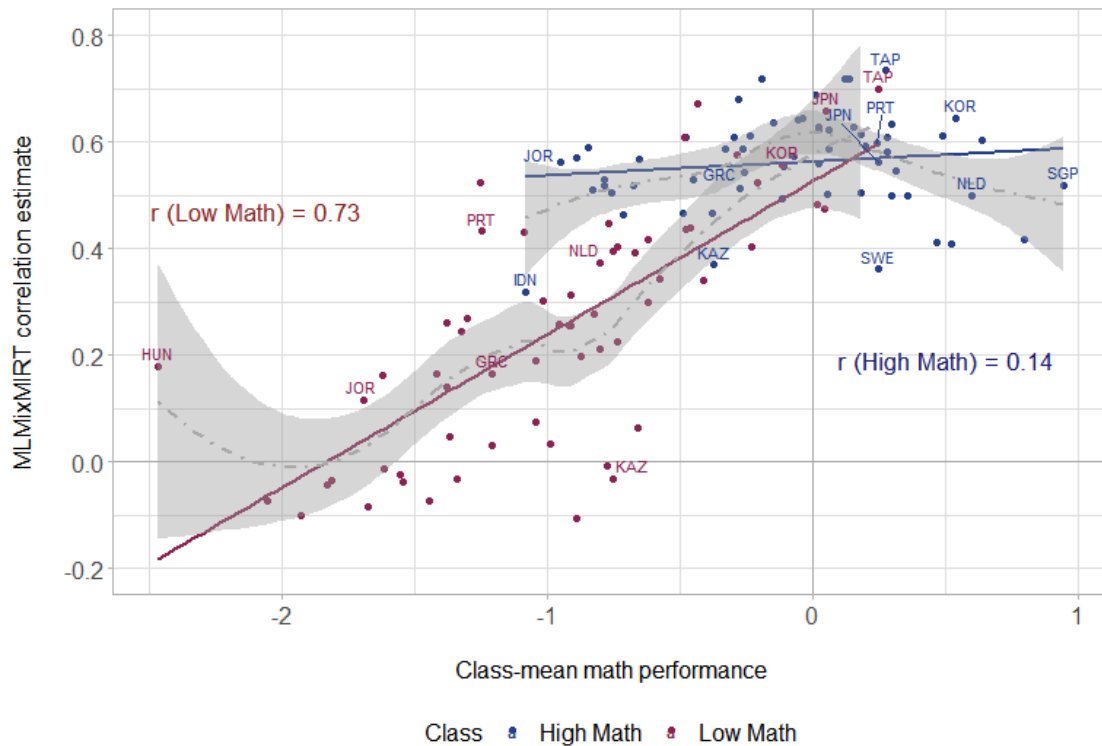


Figure 4.24. Relationship between within-country MLMixMIRT correlations and class-mean math performance

In the MLMixMIRT models, the correlation estimates are corrected for both the measurement error and the clustering in the data. Similar to the MIRT models, the EAP reliability indices obtained from the MLMixMIRT models were examined. As demonstrated

below, EAP reliabilities for the math scale were larger and more homogeneous for the High Math Class than those for the Low Math Class. However, the High Math Class for Colombia and Sweden showed very low EAP reliabilities (0.08 and 0.28, respectively).

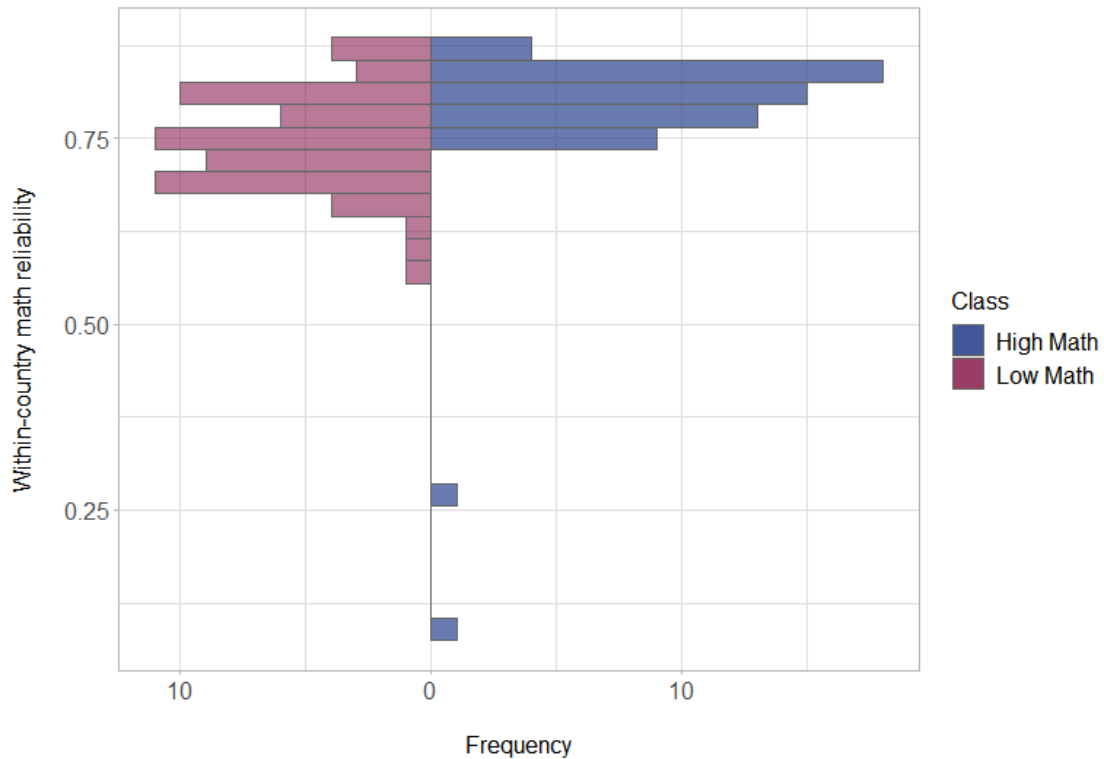


Figure 4.25. Distribution of within-country EAP reliabilities for math scale from the MLMixMIRT models by latent class

For the math self-efficacy scale, the empirical reliabilities obtained from the MLMixMIRT models demonstrated a different pattern. In contrast to the math scale, math self-efficacy scale reliabilities for the Low Math Class were larger and less homogeneous than those that were obtained for the High Math Class, which were relatively smaller and more homogeneous.

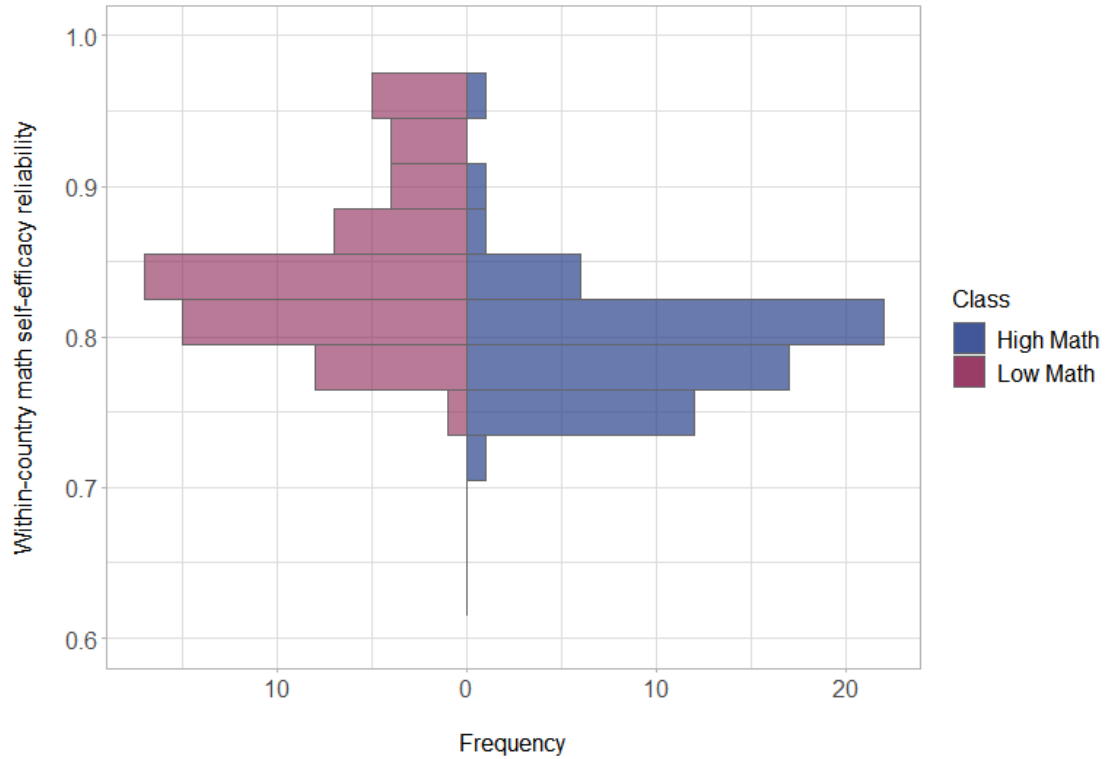


Figure 4.26. Distribution of within-country EAP reliabilities for math self-efficacy scale from the MLMixMIRT models by latent class

In order to better understand the differences between the two latent classes with respect to the amount of measurement error, EAP reliabilities for each latent class were examined in relation to country-mean math performance level. Figure 4.27 below demonstrates that, overall, the relationship between country-mean math performance and the EAP reliabilities for the math scale was linear for both latent classes but stronger for the Low Math Class in comparison to the High Math Class (the slopes were 0.84 and 0.40, respectively). However, this difference between the Low Math and High Math classes was more apparent for countries with mean math proficiency scores lower than 475. This suggests that in lower achieving countries, the difference in the amount of measurement error in students' responses to the math assessment was larger between the Low Math Class and High Math Class. Note that Colombia and Sweden were

excluded in Figure 4.27 below due to having uncommonly low EAP reliabilities for the High Math Class.

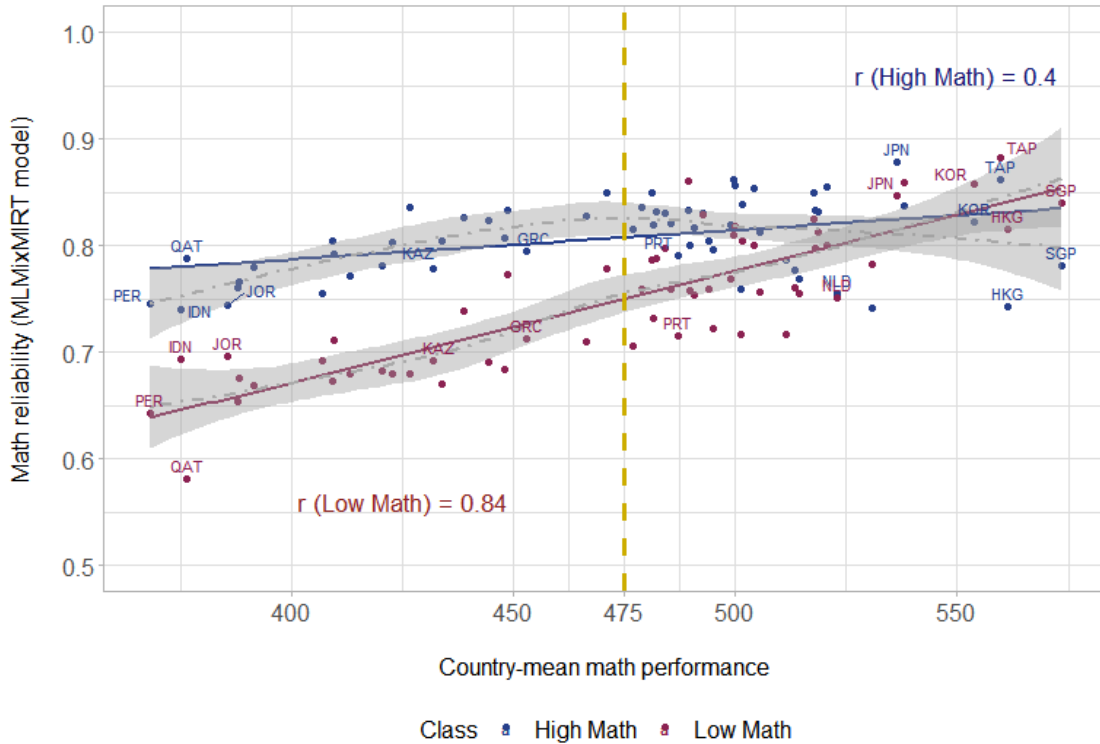


Figure 4.27. Relationship between within-country math reliabilities from the MLMixMIRT models and country-mean math performance (COL and SWE are not included)

The relationship between reliabilities for the math scale was also examined with respect to class-mean math ability estimates obtained from the MLMixMIRT models. Displayed in Figure 4.28 below, a positive and approximately linear relationship was observed in Low Math Class ($r_{\text{Low Math}} = 0.85$). In High Math Class, however, the relationship was curvilinear; the reliabilities for the math scale peaked around the class-mean math proficiency of zero but were smaller on either side of the distribution.

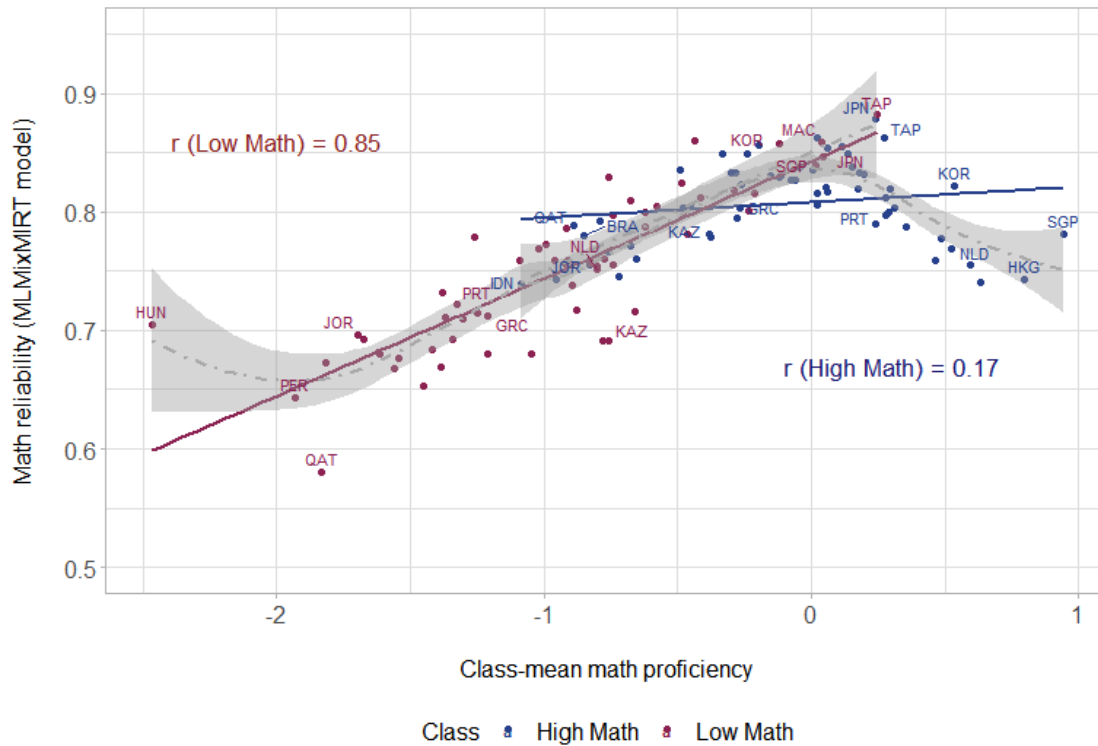


Figure 4.28. Relationship between within-country math reliabilities from the MLMixMIRT models and class-mean math performance (COL and SWE are not included)

Similar to the findings for the math scale, the EAP reliabilities for the math self-efficacy scale were less heterogeneous for the Low Math Class in comparison to the High Math Class. However, in contrast to those obtained for the math scale, the reliabilities for the math self-efficacy scale were larger for the Low Math Class. That is, the math self-efficacy scale demonstrated lower reliabilities in High Math Class. Moreover, when examined in relation to country-mean math performance, a negative relationship was observed in the Low Math Class ($r_{\text{Low Math}} = -0.35$). Figure 4.29 below demonstrates that the math self-efficacy reliabilities were weaker as the country-mean math proficiencies increased for the Low Math Class. For the High Math Class, however, the slope was positive ($r_{\text{High Math}} = 0.40$). Similar to the math scale

reliabilities, the difference between the two latent classes was more apparent for lower achieving countries with a country-mean math performance below 475.

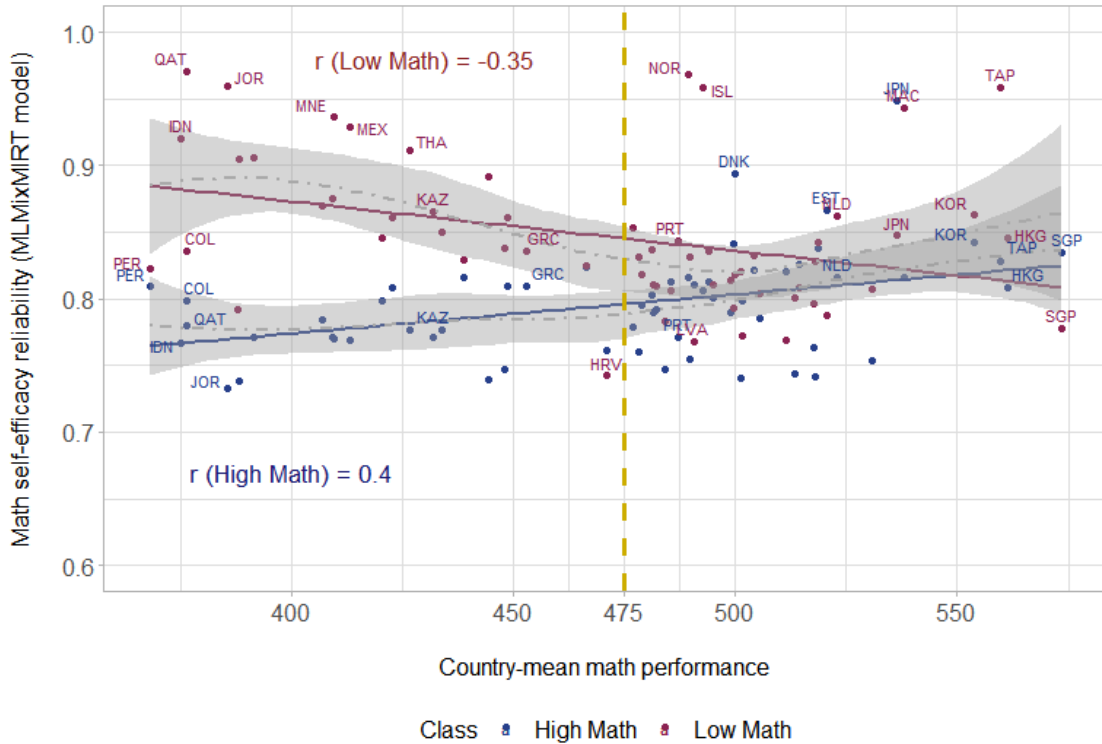


Figure 4.29. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and country-mean math performance

When the estimated reliabilities for the math self-efficacy scale were examined in relation to class-mean math proficiency estimates, a somewhat similar pattern was observed. Although they were smaller in magnitude, the reliabilities were negatively correlated with class-mean math proficiencies in Low Math Class but positively correlated in High Math Class (the slopes were -0.14 and 0.26, respectively).

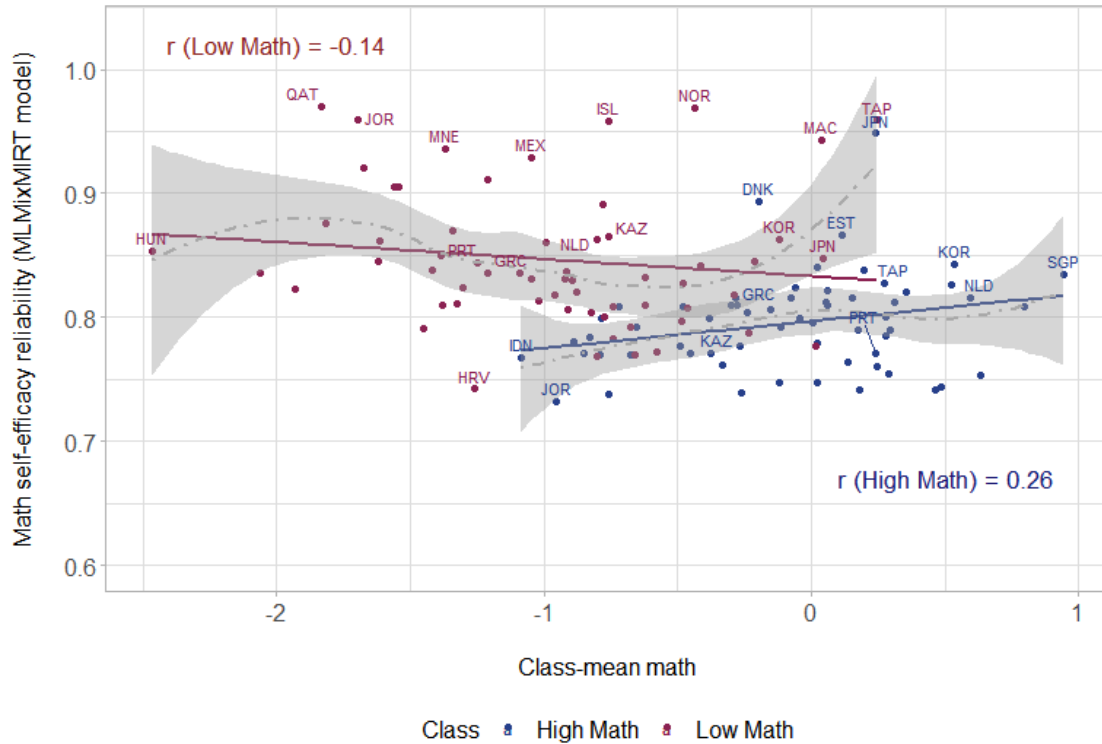


Figure 4.30. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and class-mean math performance

Within-country EAP reliabilities for the math self-efficacy scale were also examined with respect to their association with county-mean math self-efficacy indices. As presented below, a relatively weak and negative relationship was found for both Low Math and High Math classes (the slopes were -0.24 and -0.23, respectively). This suggests that the reliability of the math self-efficacy was lower for countries with higher math self-efficacy index on average. Note that the EAP reliabilities for the math self-efficacy scale were more heterogeneous for the Low Math Class. Moreover, although the reliabilities were weaker for the High Math Class in general, this was not a consistent pattern across all countries. For example, for Singapore (0.47), which exhibited the highest average math self-efficacy index across countries, the scale reliability was weaker for the Low Math Class in comparison to the High Math Class (0.78 and 0.84, respectively).

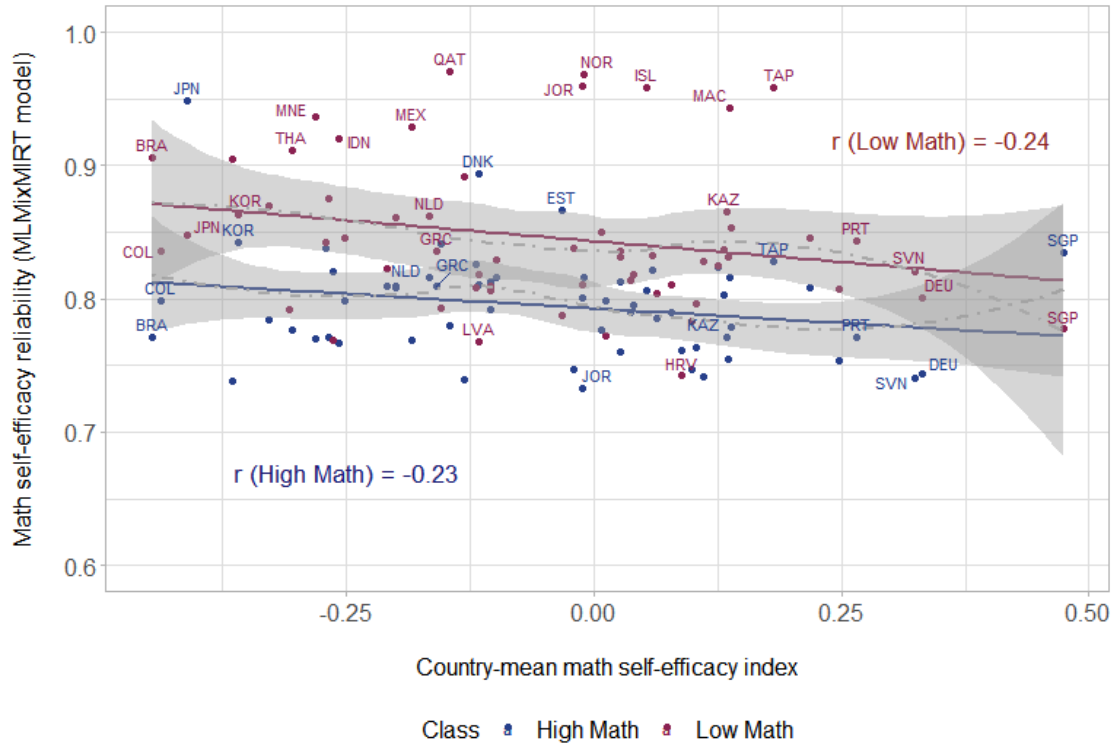


Figure 4.31. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and country-mean math self-efficacy index

In order to better understand the differences among math self-efficacy reliabilities across countries, math self-efficacy reliabilities were further examined in relation to class-mean math self-efficacy estimates instead of overall country-mean math self-efficacy indices. The distribution of class-mean math self-efficacy estimates are given in Figure 4.32. In contrast to class-mean math proficiency estimates, both latent classes demonstrated average math self-efficacies that were mostly greater than zero. Moreover, even though most of the cases were grouped around class-mean of zero, there were several cases with rather large averages in both latent classes. For example, class-mean math self-efficacy for the High Math Class for Singapore was 4.12. Although the Low Math Class demonstrated slightly lower averages, there were cases such as Norway, which had an average of 4.72 math self-efficacy for the Low Math Class.

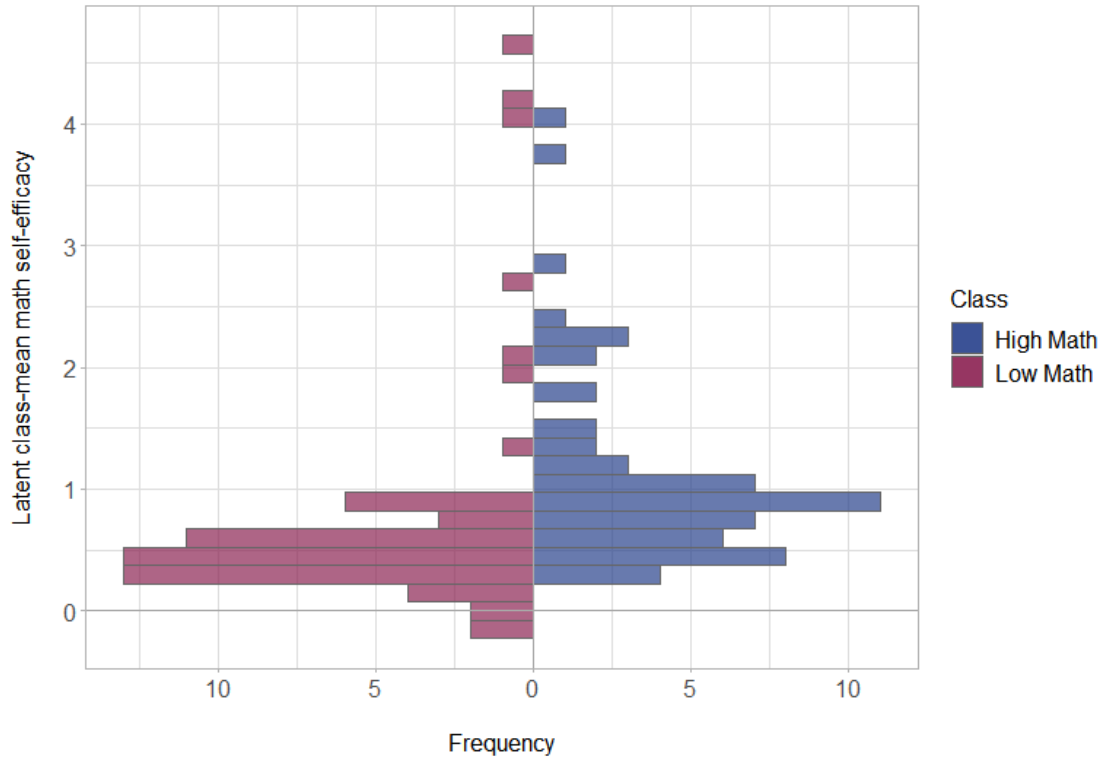


Figure 4.32. Distribution of within-country mean math self-efficacy by latent class

Figure 4.33 below displays the relationship between class-mean math self-efficacies and empirical reliabilities for the math self-efficacy scale. Due to having uncommonly large class-mean math self-efficacy estimates, nine countries (Hong Kong-China, Iceland, Jordan, Japan, Macao-China, Norway, Qatar, Singapore, and Chinese Taipei) with a class-mean math self-efficacy that was larger than 2.5 for the High Math Class and larger than 2 for the Low Math Class were excluded. Note that empirical reliabilities for the countries that were excluded from the scatterplot were larger than 0.80. For both latent classes, a positive relationship was observed between class-mean math self-efficacies and empirical reliabilities. The relationship was stronger for the Low Math Class than the High Math Class (the slopes were 0.10 and 0.33, respectively).

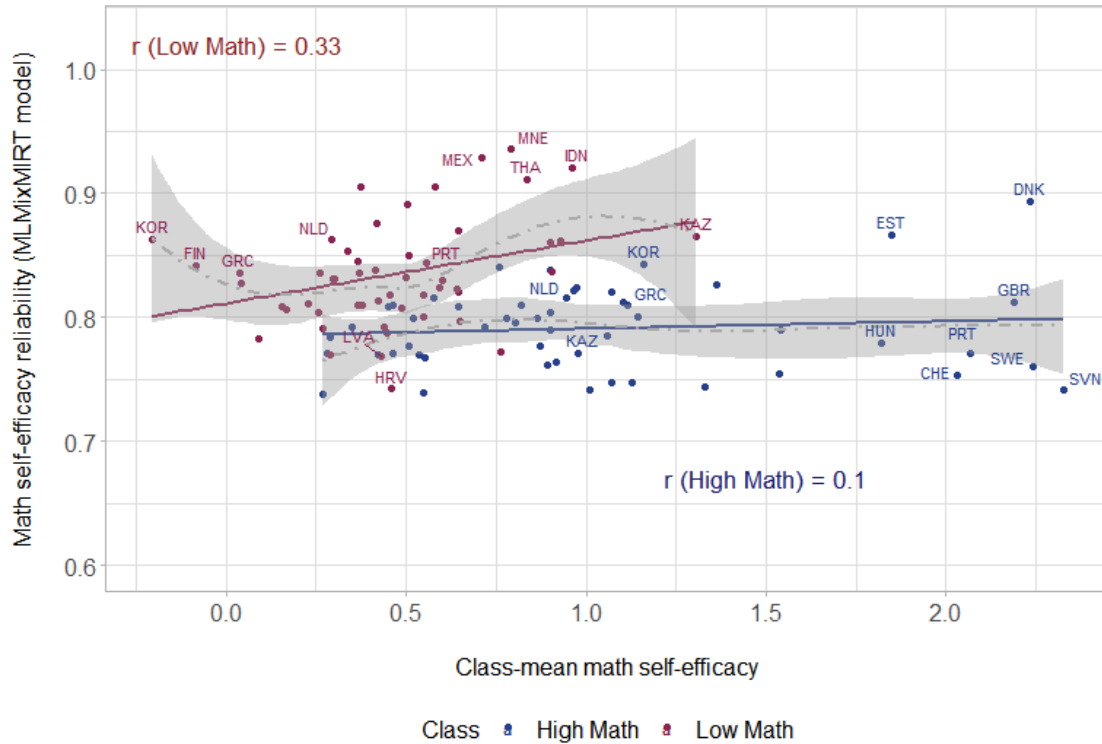


Figure 4.33. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and class-mean math self-efficacy (across 52 countries – outliers excluded)

It was hypothesized that when the reliabilities were lower, signaling a larger amount of measurement error, the correlation estimates would be attenuated. In order to test this hypothesis, the MLMixMIRT estimates were plotted against the reliability estimates for both math and math self-efficacy scales. Displayed in Figure 4.34, a strong and positive relationship was observed for both High and Low Math classes (the slopes were 0.64 and 0.87, respectively). Moreover, the relationships were linear; the MLMixMIRT estimates were larger as the math scale reliabilities increased.

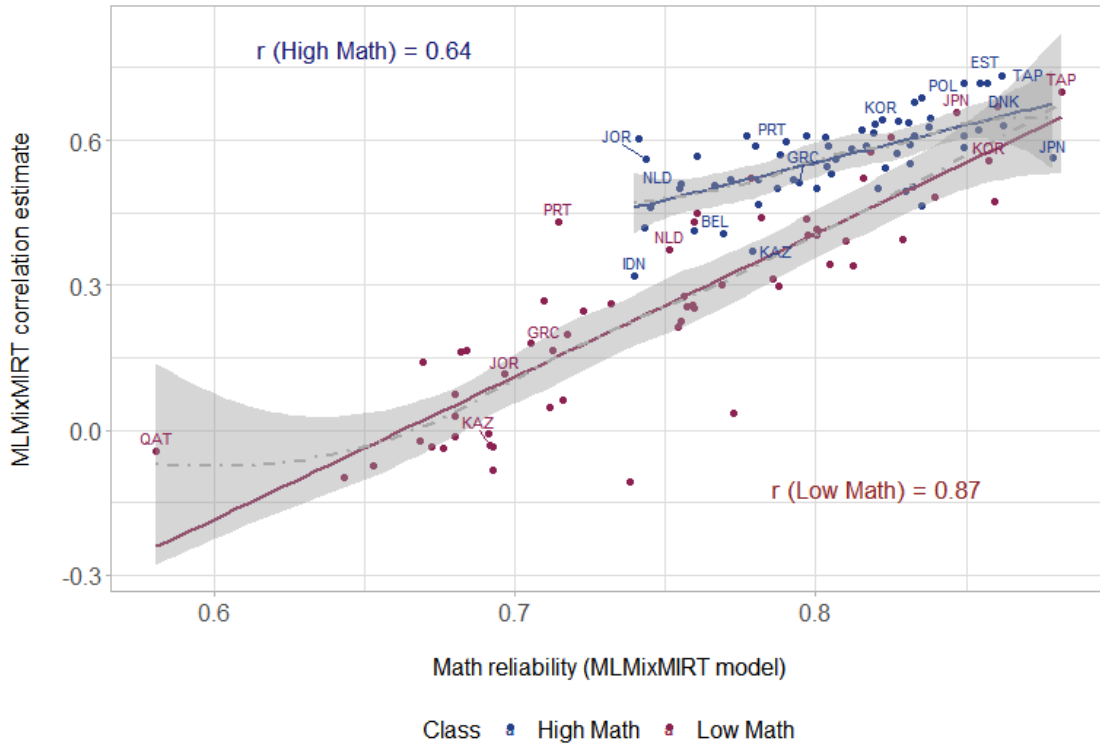


Figure 4.34. Relationship between within-country math reliabilities from the MLMixMIRT models and MLMixMIRT correlation estimates (COL and SWE are not included)

Figure 4.35, on the other hand, does not display a particularly clear pattern. Especially for the Low Math Class, many countries with the math self-efficacy scale reliabilities that were as large as 0.90 demonstrated rather small correlation estimates. There were also countries such as Croatia, which exhibited a moderately larger correlation between math proficiency and math self-efficacy even though the math self-efficacy scale reliability was the smallest in the Low Math Class. For the High Math Class, the slope was moderate but positive.

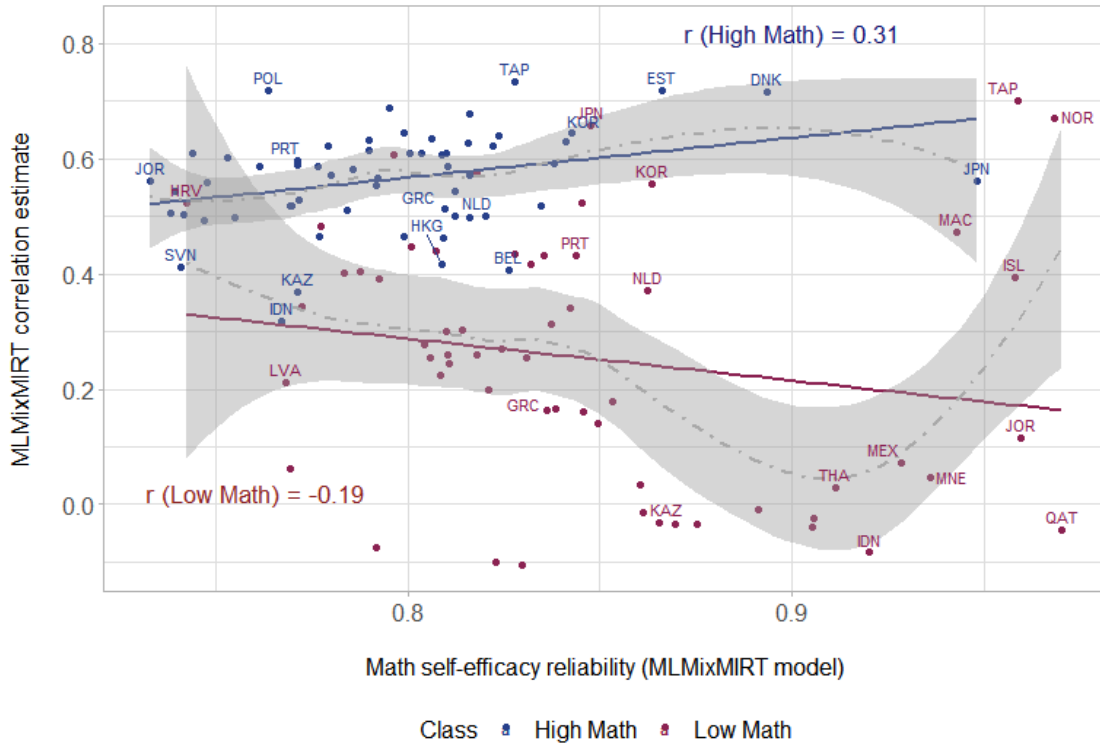


Figure 4.35. Relationship between within-country math self-efficacy reliabilities from the MLMixMIRT models and MLMixMIRT correlation estimates

4.5 Phase 5: Refinement at the School-level Within-countries

In order to investigate further the impact of measurement error on the correlation estimates and to what extent math proficiency may moderate the relationship between math proficiency and math self-efficacy, Phase 1 and Phase 2 of the model refinement process were repeated at the school-level for a set of countries exhibiting a wide range of country-level conventional correlations. To this end, six countries were selected: Kazakhstan ($r_{OLS} = 0.28$), Jordan ($r_{OLS} = 0.29$), Greece ($r_{OLS} = 0.48$), Netherlands ($r_{OLS} = 0.48$), Portugal ($r_{OLS} = 0.64$), and Chinese Taipei ($r_{OLS} = 0.66$). Due to convergence issues with MIRT modeling at the school-level, an alternative strategy had to be followed. As described in Chapter 3, schools within each selected country were grouped into nine “super-schools” that were ordered based on mean math

proficiency. Schools falling into the bottom 15% of the distribution were grouped into *super-school 1*, those in the top 15% were grouped into *super-school 9*, and the remaining schools were grouped into seven super-schools, each containing approximately 10% of the sample. .

First, within each country, ordinary regression models were fit to data from each super-school and within-super-school correlation estimates between math proficiency and math self-efficacy were obtained as the baseline. Descriptive statistics for the distributions of conventional correlations for each country are presented below.

Table 4.8. Descriptive statistics for the within-super-school conventional correlation estimates from OLS models

<u>Country</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
Kazakhstan	0.24	0.07	0.10	0.26	0.33	0.22	-0.44	0.09
Jordan	0.27	0.08	0.09	0.30	0.34	0.25	-1.16	0.06
Greece	0.42	0.06	0.31	0.42	0.49	0.19	-0.48	0.07
Netherlands	0.42	0.08	0.24	0.44	0.49	0.25	-1.31	0.06
Portugal	0.58	0.05	0.46	0.60	0.62	0.17	-1.46	0.03
Chinese Taipei	0.55	0.09	0.35	0.56	0.66	0.31	-0.87	0.07

Countries with larger correlations at the country level exhibited larger correlations at the super-school level. For example, conventional correlations for Kazakhstan ranged from 0.10 to 0.33 whereas they ranged from 0.35 to 0.66 for Chinese Taipei. Moreover, the maximum correlation estimate for the super-schools was close to the country-level correlation estimates for the top four countries with larger correlation estimates at the country-level. For instance, country-level conventional estimates were 0.66 and 0.64 for Chinese Taipei and Portugal and the maximum correlation estimates for the super-schools were 0.66 and 0.62, respectively. Although they differed in magnitude, the level of variation in the conventional estimates across super-

schools was generally similar across countries. By contrast, for Portugal, the estimates were substantially less dispersed than those for the other five countries. However, even though country-level conventional correlations were very similar for Chinese Taipei and Portugal (0.66 and 0.64, respectively), the estimates were more heterogeneous for Chinese Taipei. The histograms of the distributions of the conventional correlation estimates are presented below.

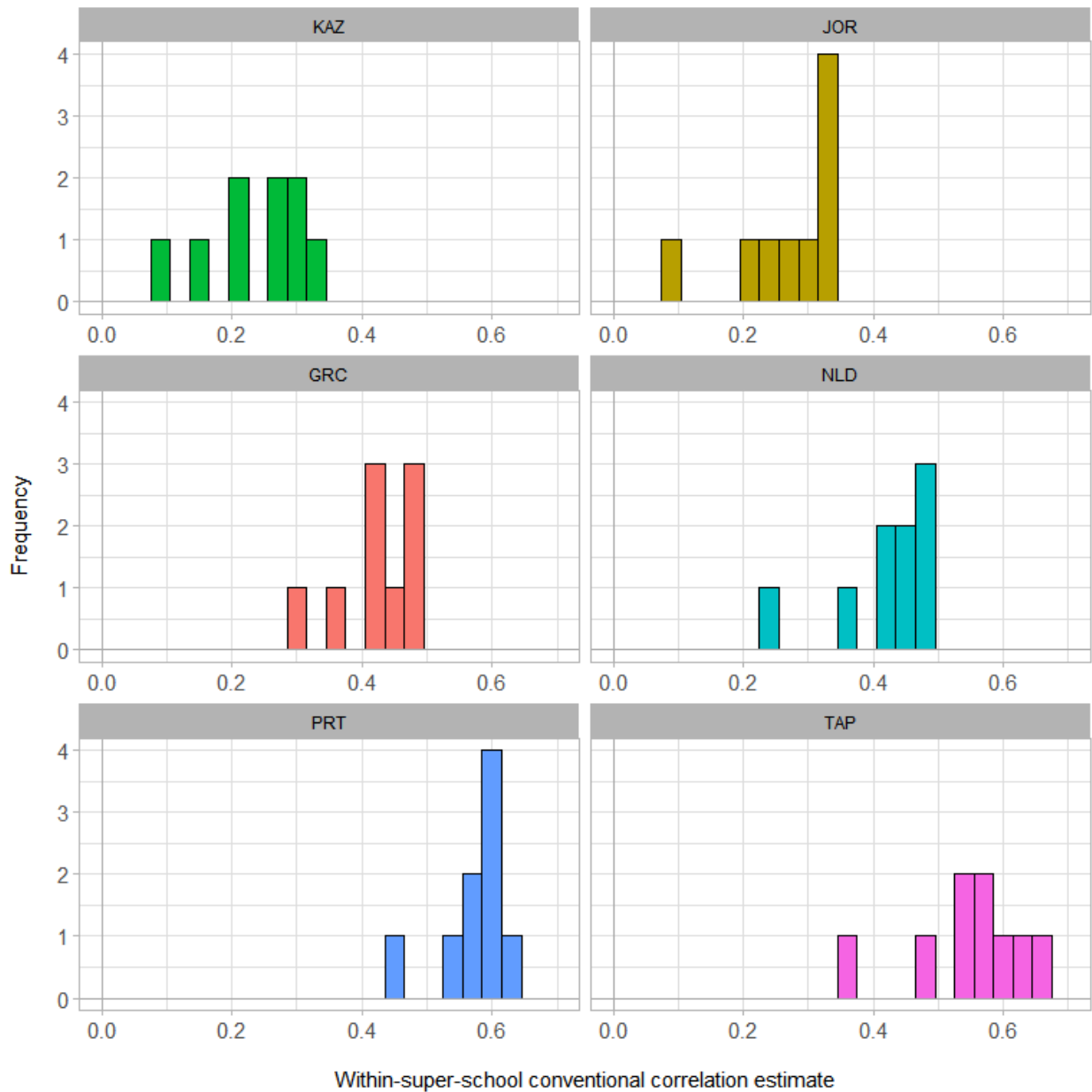


Figure 4.36. Distribution of within-super-school conventional correlations by country

In order to examine the association between correlation estimates and math proficiency levels, averages of the first plausible value for math proficiency (PV1MATH) were used. As described in Chapter 3, in order to ensure the comparability, the average of the first plausible value for math proficiency for each super-school was calculated by employing a new set of senate weights that summed to 1000 within each super-school. Descriptive statistics for the weighted averages of math proficiency by country are presented in Table 4.9. Mean math proficiencies for super-schools in Jordan displayed the narrowest range whereas those for Chinese Taipei demonstrated the widest range in comparison to other countries.

Table 4.9. Descriptive statistics for the weighted averages of PV1MATH for super-schools by country

<u>Country</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
Kazakhstan	429	44	372	423	516	144	0.60	47
Jordan	386	41	325	382	467	142	0.47	38
Greece	460	46	365	468	527	162	-0.60	41
Netherlands	528	70	415	527	624	209	-0.15	102
Portugal	488	52	396	495	565	169	-0.29	59
Chinese Taipei	557	72	441	554	686	244	0.15	80

Figure 4.37 consists of six scatterplots displaying the relationships between conventional correlation estimates and weighted average of PV1MATH for super-schools by each country. Patterns observed at the super-school level were somewhat similar to those obtained from country-level analyses. Except for Chinese Taipei, a positive and approximately linear relationship was observed; correlations between math proficiency and math self-efficacy were larger for higher achieving super-schools. In Chinese Taipei, on the other hand, a negative relationship was

observed (the slope was -0.69); correlations were weaker as the math proficiency level of the super-school increased. Chinese Taipei was also the highest achieving country among selected countries. In particular, the weighted average of PV1MATH for super-school 9 in Chinese Taipei was the highest across countries and a substantial drop in correlation estimates was observed for super-schools 8 and 9.

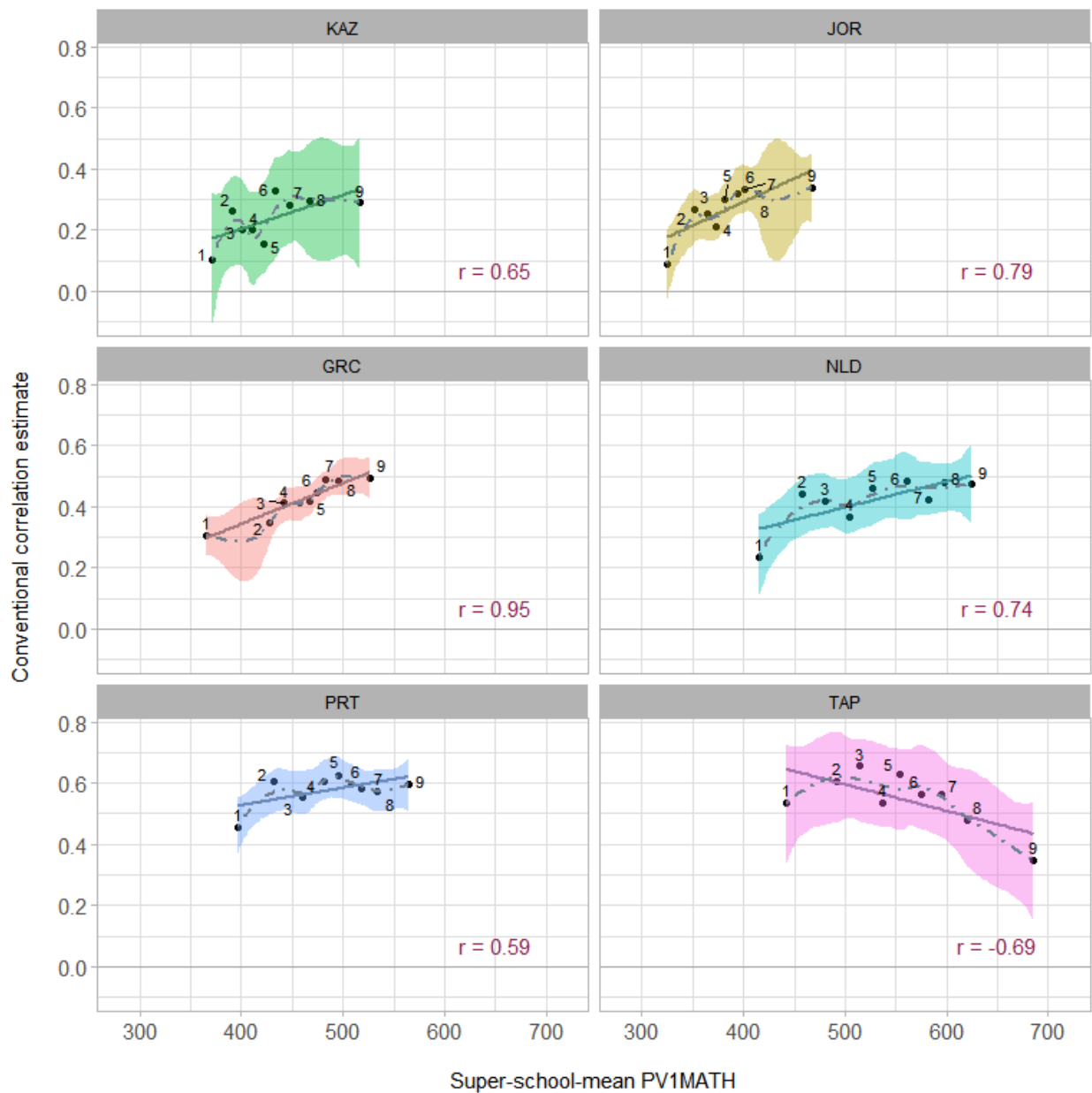


Figure 4.37. Relationships between conventional correlations and mean math proficiencies

Next, multidimensional IRT modeling was employed to account for the measurement error in the data. Similar to Phase 2 analyses, two-dimensional IRT models were fit to the item response data from super-schools using the same Q-matrix. In order to obtain within-super-school correlation estimates separately for each super-school within each country, multi-group 2PL MIRT models were fit to the data from each country. Multi-group MIRT models were fit to the data by keeping the parameters equal across the super-schools within each country.

Overall, correlation estimates from the MIRT models were larger than the conventional estimates, a pattern similar to that observed at the country level in Phase 2. For example, the minimum estimate obtained from the super-schools in Kazakhstan increased from 0.10 to 0.17 and the maximum estimate increased from 0.33 to 0.43. However, correlation estimates still displayed substantial variability across super-schools within countries. In fact, MIRT correlation estimates were substantially more dispersed than conventional correlation estimates. In contrast to those obtained for the conventional estimates (Table 4.8), interquartile ranges of the estimates were doubled in all countries except for the Netherlands. These findings were inconsistent with the findings from Phase 2 as the MIRT estimates demonstrated a slightly narrower interquartile range than the conventional estimates when the data were analyzed at the country-level.

Descriptive statistics of the MIRT correlation estimates are presented below in Table 4.10.

Table 4.10. Descriptive statistics for the within-super-school correlation estimates from MIRT models

<u>Country</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
Kazakhstan	0.29	0.09	0.17	0.32	0.43	0.26	-0.11	0.16
Jordan	0.34	0.14	0.05	0.37	0.48	0.43	-0.92	0.14
Greece	0.49	0.08	0.33	0.50	0.58	0.25	-0.46	0.13
Netherlands	0.47	0.09	0.29	0.49	0.60	0.31	-0.46	0.08
Portugal	0.65	0.05	0.55	0.65	0.71	0.16	-0.60	0.06
Chinese Taipei	0.58	0.08	0.45	0.55	0.69	0.24	0.01	0.14

Displayed in Figure 4.38 below, the distributions of the correlation estimates were shifted to the right demonstrating larger values in comparison to the conventional correlation estimates. In contrast to those obtained from country-level analyses in Phase 2, the shapes of the distributions indicated that MIRT estimates were more heterogeneous in comparison to conventional estimates.

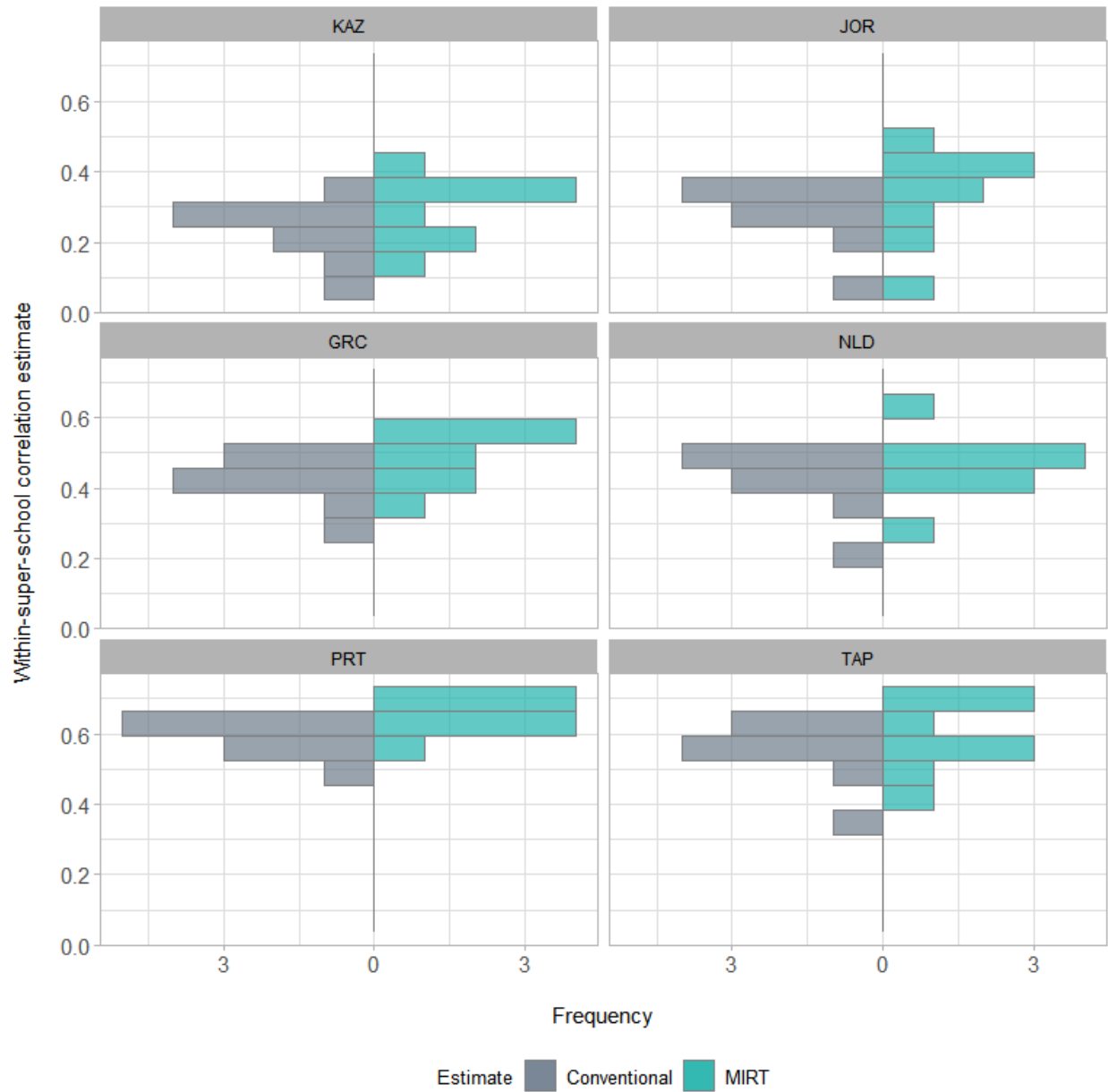


Figure 4.38. Distribution of within-super-school conventional correlations and MIRT correlations

As displayed in Figure 4.39 below, similar to those obtained for the conventional estimates, a positive and approximately linear relationship between MIRT correlation estimates and mean math achievement was observed in all countries except for Chinese Taipei. Moreover, consistent with the country-level results obtained in Phase 2, the relationship was slightly

weakened after employing MIRT modeling for countries that demonstrated positive relationships. For instance, the slope of the relationship decreased from 0.59 to 0.44 in Portugal. For Chinese Taipei, on the other hand, the negative relationship that was observed between conventional correlation estimates and mean math achievement was effectively equal in magnitude when MIRT modeling was employed (i.e., the slopes were -0.69 and -0.71, respectively).

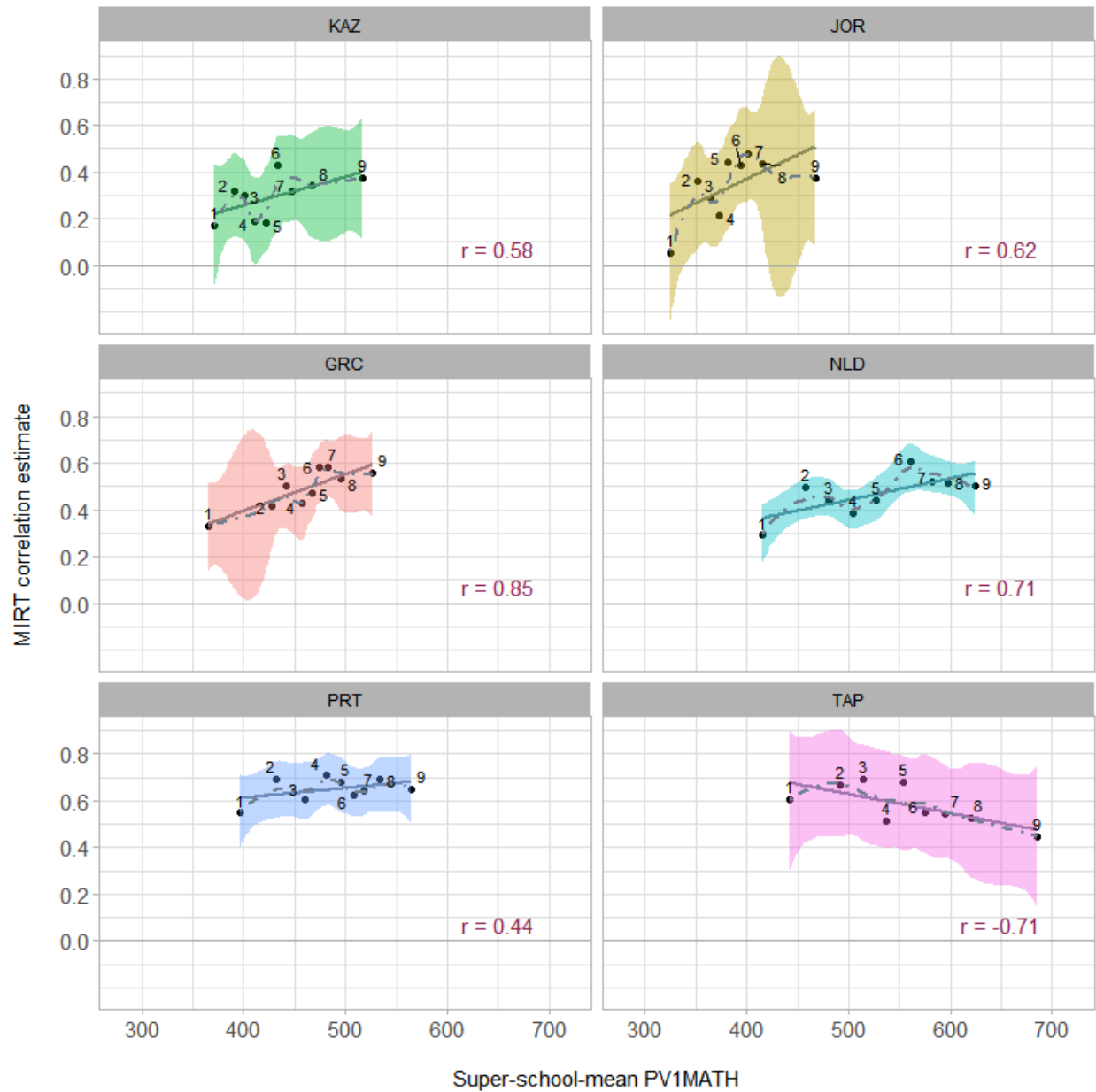


Figure 4.39. Relationships between MIRT correlations and mean math proficiencies

Similar to Phase 2 and Phase 4, the EAP reliability indices obtained from the MIRT models were examined. As demonstrated below, EAP reliabilities for the math scale varied substantially across super-schools within countries as well as across countries. In particular, reliabilities were lower in Kazakhstan and Jordan. For super-schools in the Netherlands, the reliabilities were more homogenous in comparison to those observed in other countries. Super-

school 1 (i.e., schools falling into bottom 15% of the distribution based on mean math achievement) in Greece as well as in Portugal demonstrated substantially lower reliabilities for the math scale whereas super-school 9 (i.e., schools falling into top 15% of the distribution based on mean math achievement) in Chinese Taipei exhibited a substantially lower reliability for the math scale.

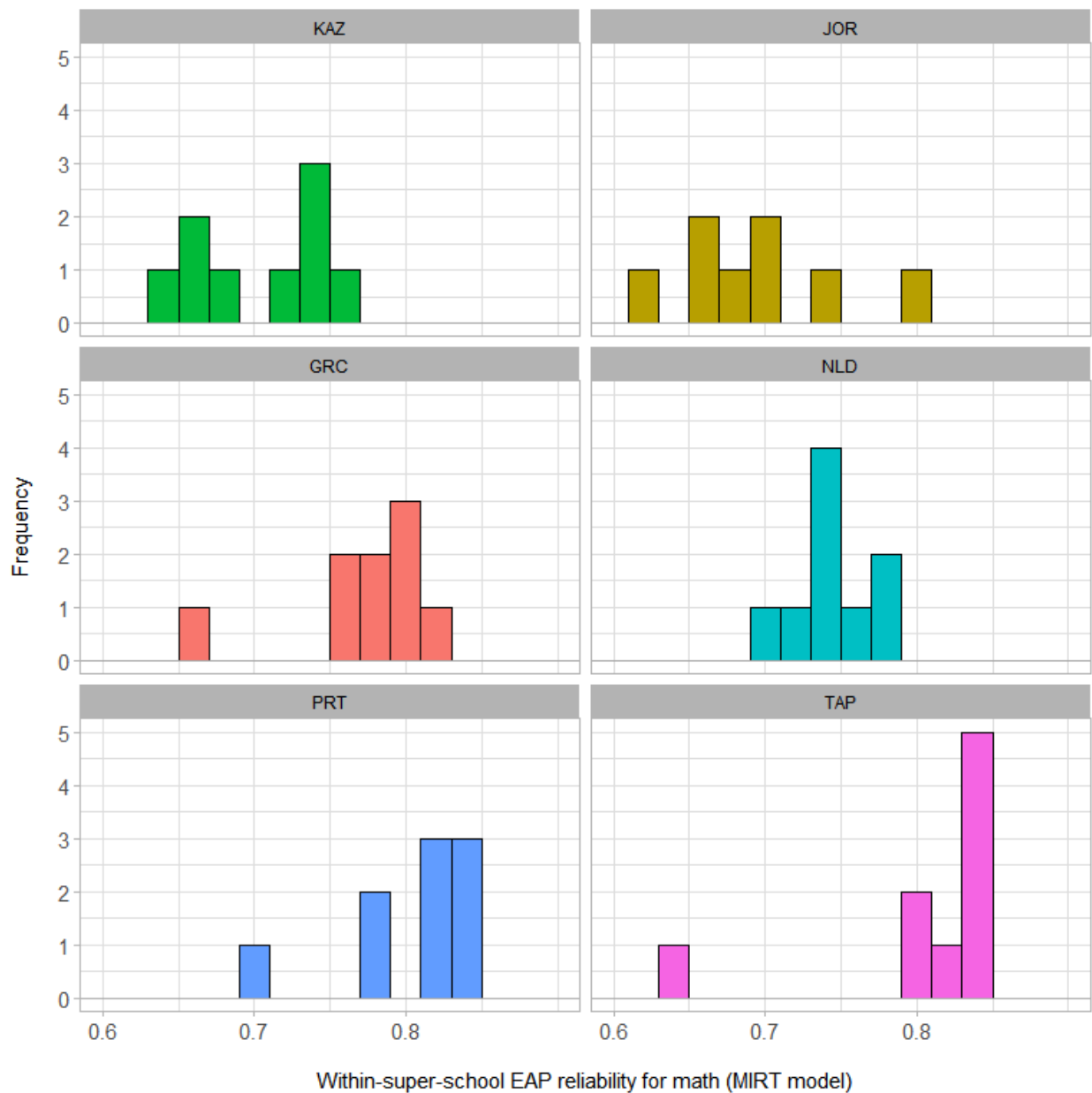


Figure 4.40. Distributions of EAP reliabilities gathered from the MIRT models for the math proficiency scale

When reliabilities for the math proficiency scale were examined in relation to mean math performance levels of the super-schools (Figure 4.41), a positive and approximately linear relationship was observed in most countries. The relationship was effectively zero in the Netherlands and Chinese Taipei (the slopes were 0.07 and 0.05, respectively). Note that super-school 1 in Greece and Portugal, and super-school 9 in Chinese Taipei were excluded in the plots due to unusually low reliabilities. For Kazakhstan and Jordan, the two countries in the sample with lowest conventional correlations at the country-level, the relationship between the reliabilities for the math proficiency scale and mean math performance was quite strong (the slopes were 0.91 and 0.97, respectively). That is, super-schools with lower mean math performances demonstrated considerably lower reliabilities for the math proficiency scale. These findings were consistent with those found in Phase 2 and Phase 4.

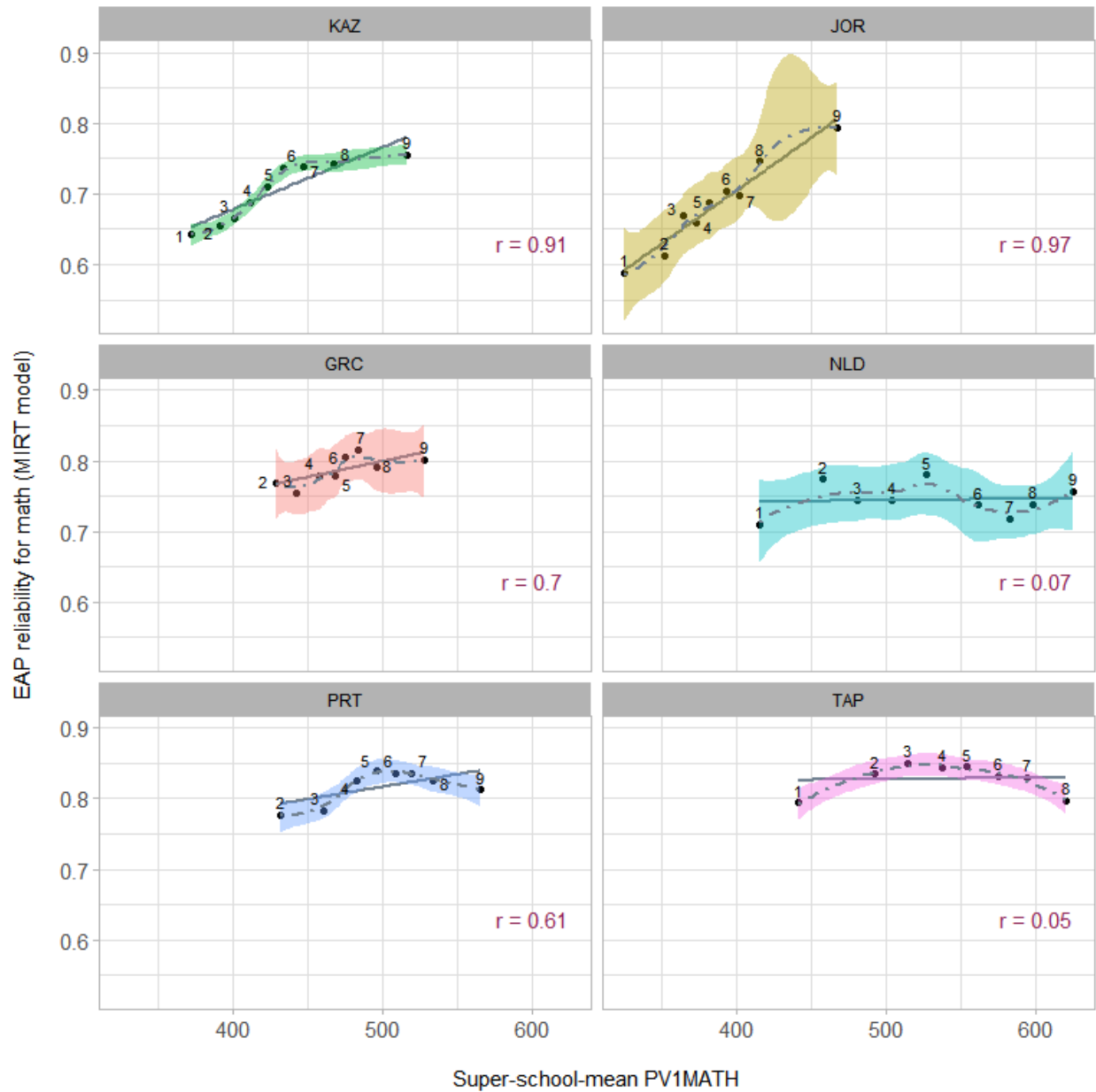


Figure 4.41. Relationship between empirical reliability indices for math proficiency and mean math performance

Reliabilities for the math proficiency scale were also examined in relation to the correlation estimates obtained from the MIRT models. The scatterplots in Figure 4.42 below display positive relationships in all countries, similar to those observed in Phase 2 and Phase 4. However, the relationship was considerably stronger for Kazakhstan, Jordan, and Greece (lower

achieving countries among all six countries). Note that super-school 1 in Greece and Portugal, and super-school 9 in Chinese Taipei were excluded in the plots due to unusually low reliabilities.

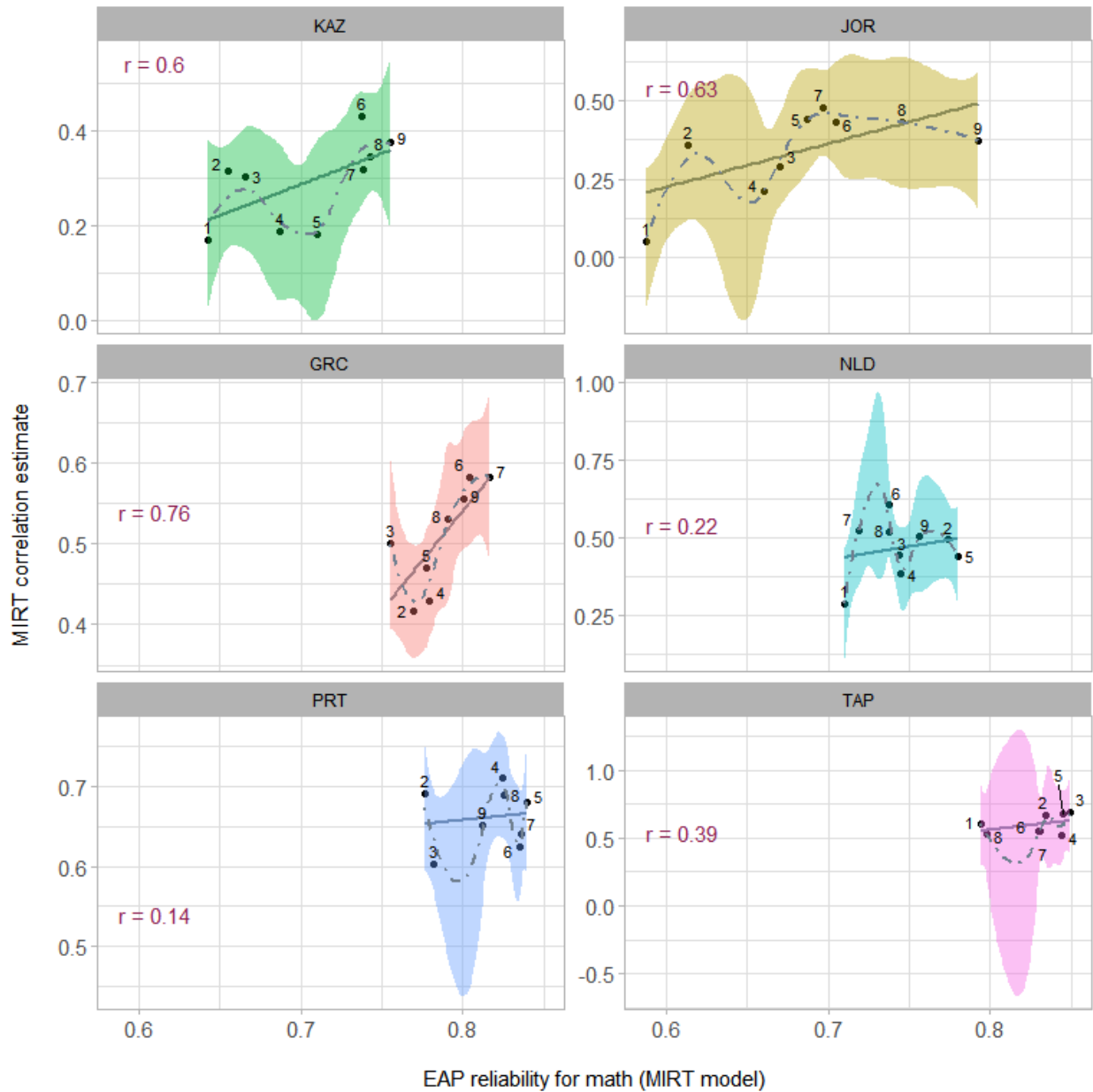


Figure 4.42. Relationship between empirical reliability indices for math proficiency and MIRT correlation estimates

Distributions of the reliabilities for the math self-efficacy scale (Figure 4.43) were more homogenous in comparison to those for the math proficiency scale (Figure 4.40). On the other hand, larger variation was observed in Portugal and Chinese Taipei (higher achieving countries among all six countries). Moreover, similar to the reliabilities for the math proficiency scale, the super-school 9 in Chinese Taipei again exhibited low reliability.

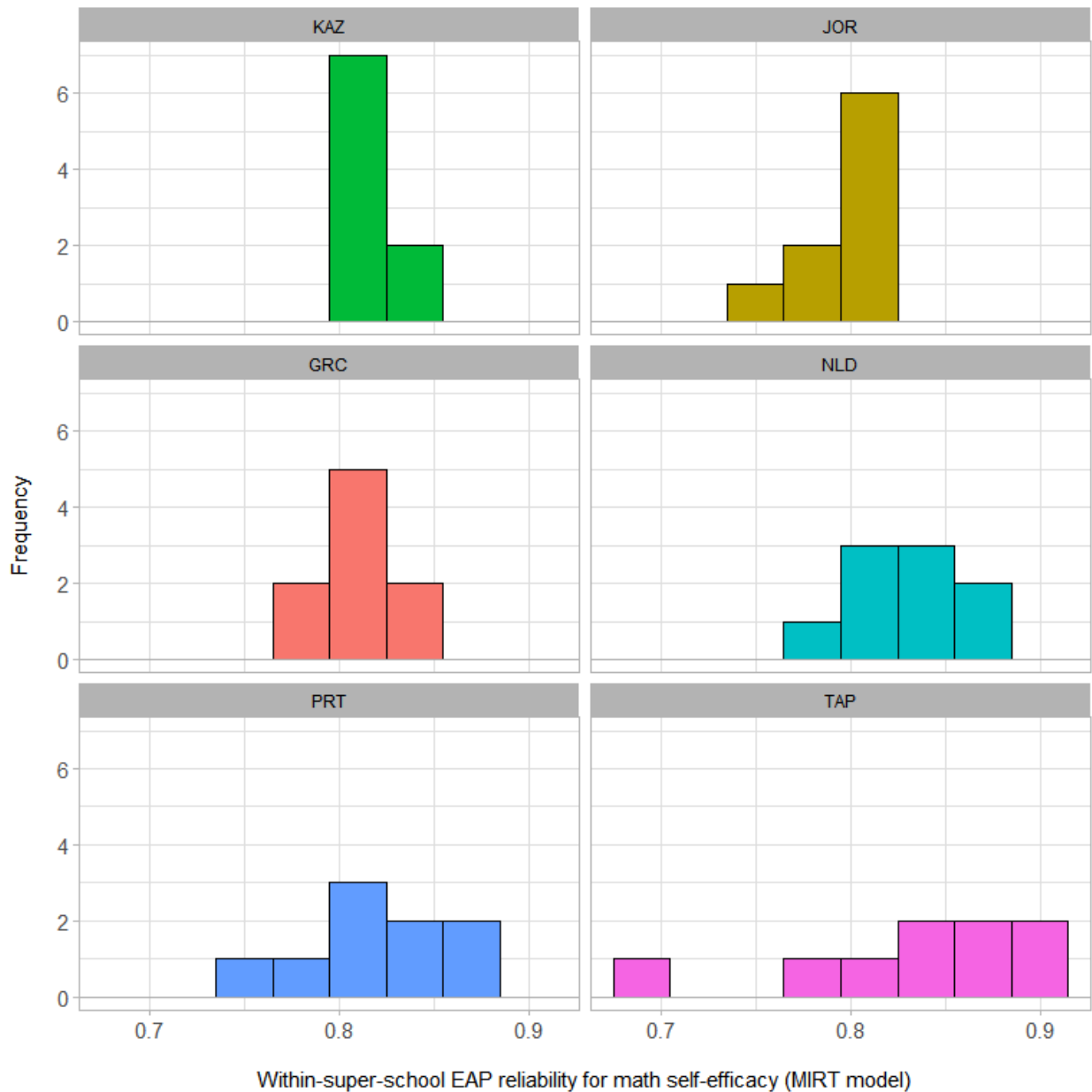


Figure 4.43. Distributions of EAP reliabilities gathered from the MIRT models for the math self-efficacy scale

When examined in relation to mean math performance, a negative and approximately linear relationship was observed in all countries. Figure 4.44 below shows that higher achieving super-schools possessed lower reliabilities for the math self-efficacy scale. Note that super-school 9 in Chinese Taipei was excluded in the plots below due to unusually low reliability. Lower reliabilities for the math self-efficacy scale observed in higher achieving groups may be due to ceiling effects resulting from low or no variance in students' responses. These findings were somewhat inconsistent with those found in Phase 2. A positive, albeit weak (the slope was 0.24), relationship was observed between country-mean math performance and reliabilities for the math self-efficacy scale in lower achieving countries whereas the relationship was effectively zero in higher achieving countries (Figure 4.13).

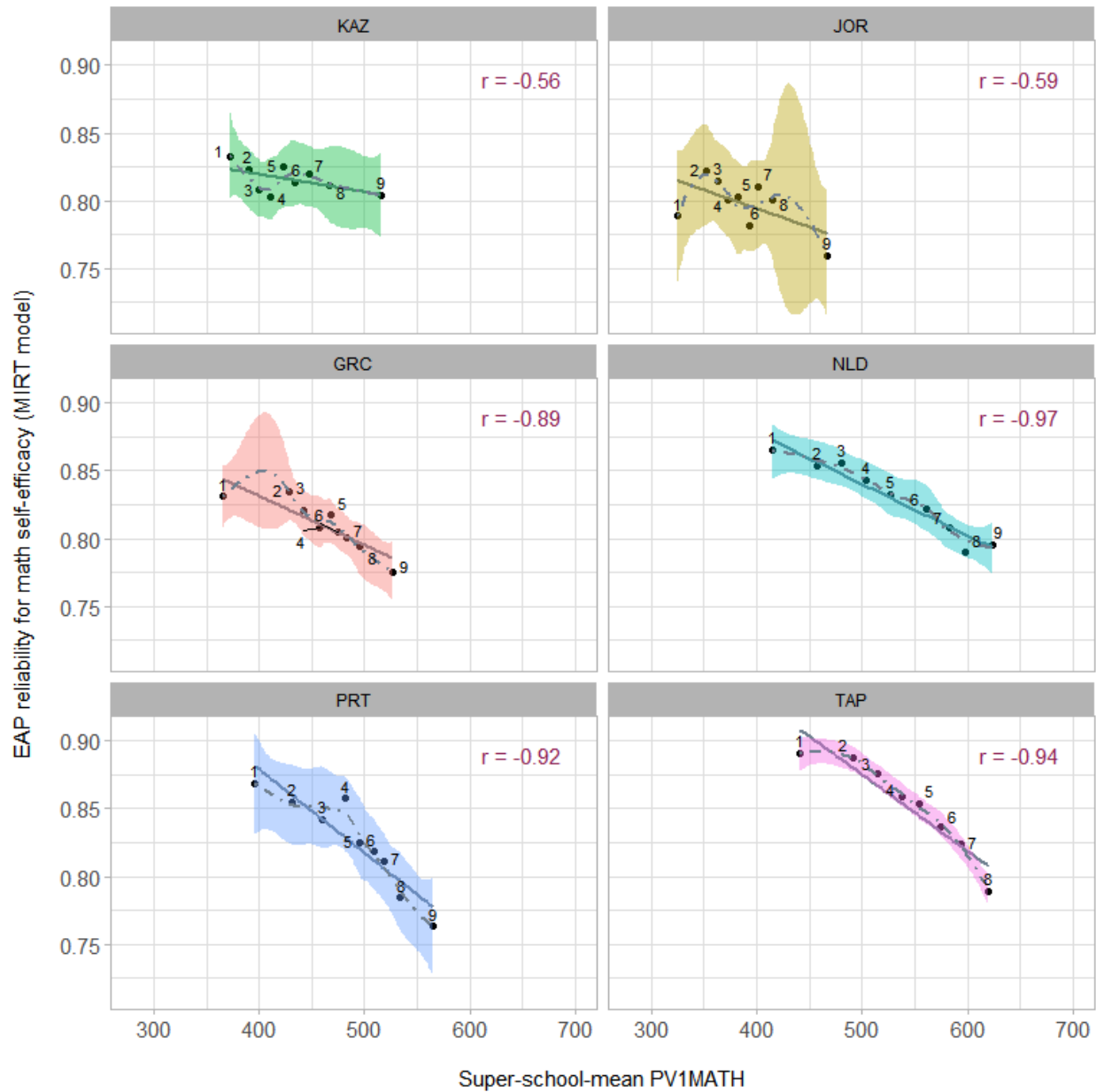


Figure 4.44. Relationship between empirical reliability indices for math self-efficacy and mean math performance

Reliabilities for the math self-efficacy scale were also examined in relation to average math self-efficacy of super-schools within each country. Note that, similar to those calculated for math proficiency, the average math self-efficacy for each super-school was calculated by employing the senate weights that summed to 1000 within each super-school. Descriptive

statistics for the distributions of average math-self efficacy by country are presented in Table 4.11. Akin to the distributions of super-schools' mean math performance, super-schools' mean math self-efficacy levels varied substantially in Chinese Taipei whereas a very narrow range was observed in Jordan.

Table 4.11. Descriptive statistics for the weighted averages of MATHEFF for super-schools by country

<u>Country</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
Kazakhstan	0.14	0.18	-0.15	0.13	0.41	0.56	0.16	0.18
Jordan	-0.03	0.14	-0.18	-0.06	0.29	0.47	1.10	0.04
Greece	-0.14	0.27	-0.72	-0.09	0.30	1.02	-0.56	0.15
Netherlands	-0.17	0.28	-0.61	-0.19	0.24	0.85	-0.12	0.42
Portugal	0.27	0.37	-0.34	0.39	0.81	1.15	-0.18	0.39
Chinese Taipei	0.16	0.53	-0.71	0.22	1.01	1.72	-0.10	0.50

Super-schools with higher mean math self-efficacy levels demonstrated lower reliabilities for the math self-efficacy scale. Figure 4.45 below demonstrates that a negative and approximately linear relationship was observed in all countries. The slopes were larger than 0.80 in absolute value except for Kazakhstan where a substantially weaker relationship was observed (the slope was -0.19). Note that super-school 9 in Chinese Taipei was again excluded in the plots due to unusually low reliability. These findings were somewhat similar to those found in Phase 2 (Figure 4.14). Although the relationship was effectively zero in lower achieving countries, a negative and approximately linear relationship was found between country-mean math self-efficacy index and reliabilities for the math self-efficacy scale in Phase 2.

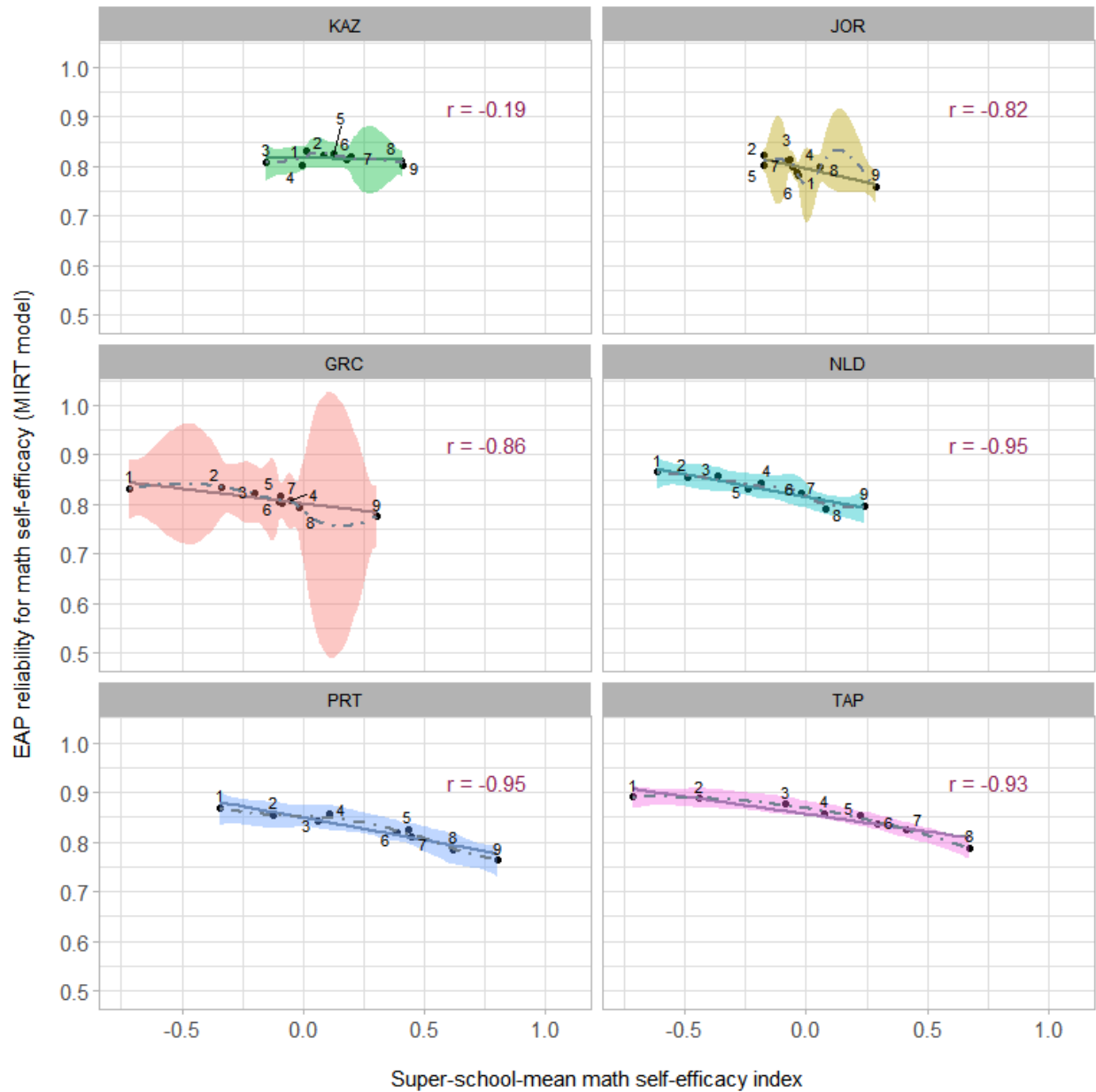


Figure 4.45. Relationship between empirical reliability indices for math self-efficacy and mean math self-efficacy

4.6 Phase 6: Refinement with a Composite Background Variable

The last phase of the analytical procedures examined whether using a composite background measure substantially reduced the impact of measurement error on correlation

estimates. To this end, a background measure that comprised math self-efficacy, math anxiety, and math self-concept measures was created by employing principal component analysis within each country. Although the literature suggests that these three domain-specific background measures are closely related constructs, the data were examined to ensure that they supported data reduction procedures. Thus, a total of sixty-one correlation matrices resulting from responses to 18 items from the three background measures was explored by using three main criteria⁴: (i) the determinant, (ii) Bartlett's test of sphericity, and (iii) the Kaiser-Meyer-Olkin measure of sampling adequacy (MSA).

The determinants of all sixty-one correlation matrices were found to be greater than .00001 indicating no linear dependencies among items. Bartlett's test of sphericity (Bartlett, 1951) compares the correlation matrix to an identity matrix with the same dimensions to test whether the data exhibits substantial correlations among observed variables. The results were found to be significant in all countries, indicating that all the correlation matrices were significantly different from the identity matrix. MSA is another measure to assess the correlation matrix in terms of the adequacy of the inter-correlations of the items for data reduction procedures. MSA values that are closer to 1 indicate a predominance of substantial correlation entries in the correlation matrix. Observed MSA values ranged from 0.83 (Romania) to 0.95 (Denmark), exceeding the 0.80 threshold that is commonly used in the literature (Hair et al., 2009; Rencher, 2002).

After ensuring the appropriateness of the correlation matrices, PCA was employed using the *FactoMineR* package (Le et al., 2008) in R. Eigenvalues of the first principal component

⁴ The procedures were executed in R; the Bartlett's test of sphericity and the Kaiser-Meyer-Olkin measure of sampling adequacy are provided in the *psych* package (Revelle, 2019) in R.

ranged from 3.93 (Romania) to 8.53 (Denmark) and proportion of the variance explained by the first component ranged from 22% (Romania) to 47% (Denmark) with a mean of 36%.

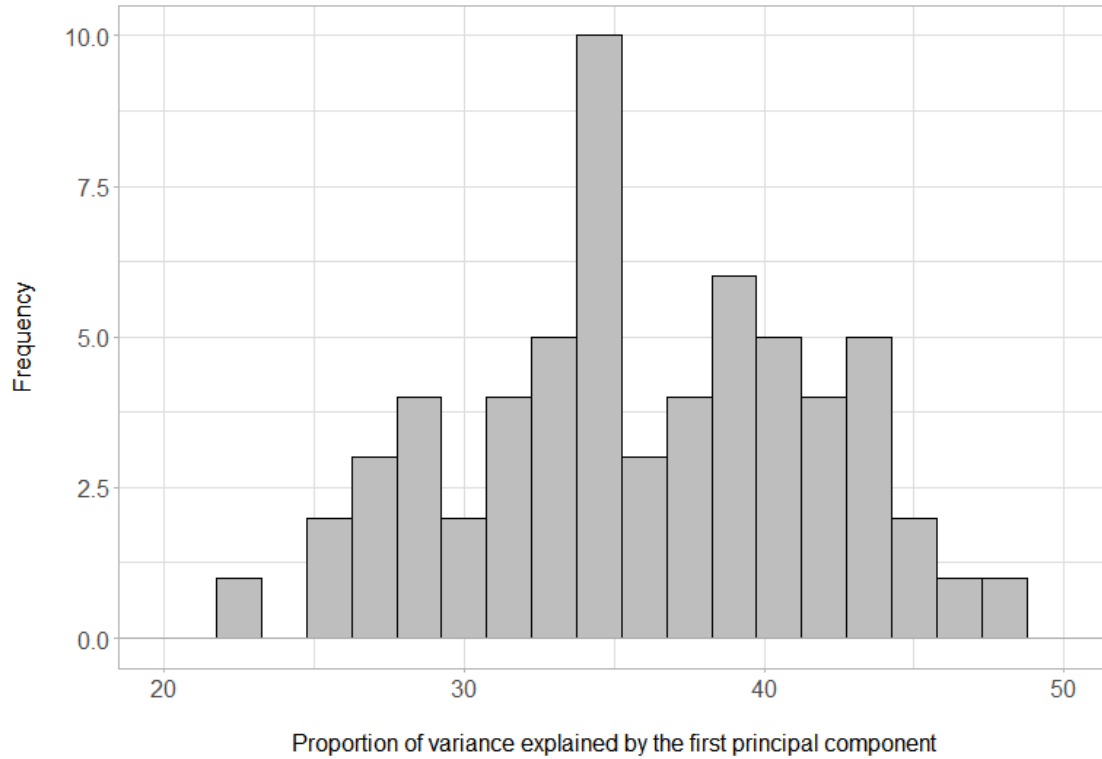


Figure 4.46. Distribution of proportion of variance explained by the first principal component

Ordinary regression models were fit to data from each country by employing the composite background variable (i.e., the first principal component) as the background measure. Resulting within-country correlation estimates between math proficiency and the background variable were compared to those when the math self-efficacy measure alone was employed as the background measure. Note that only one third of the sample was employed in both of these models as described in Chapter 3. Estimates from the models employing the composite measure were labelled as “OLS_composite” and those from the models employing the math self-efficacy measure alone were labelled as “OLS_matheff”.

Descriptive statistics of the conventional correlation estimates from the two separate sets of models are presented below. Although the minimum OLS_composite estimate (0.11) was smaller than the minimum OLS_matheff estimate (0.14), OLS_composite estimates were larger in general. The interquartile range for the OLS_composite estimates was narrower than the one for the OLS_matheff estimates (0.20 and 0.23, respectively).

Table 4.12. Descriptive statistics for the within-country conventional correlation estimates from OLS models with the composite measure and with math self-efficacy measure alone

<u>Conventional Estimates</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
OLS_matheff	0.44	0.14	0.14	0.49	0.65	0.51	-0.61	0.23
OLS_composite	0.47	0.13	0.11	0.51	0.69	0.58	-0.72	0.20

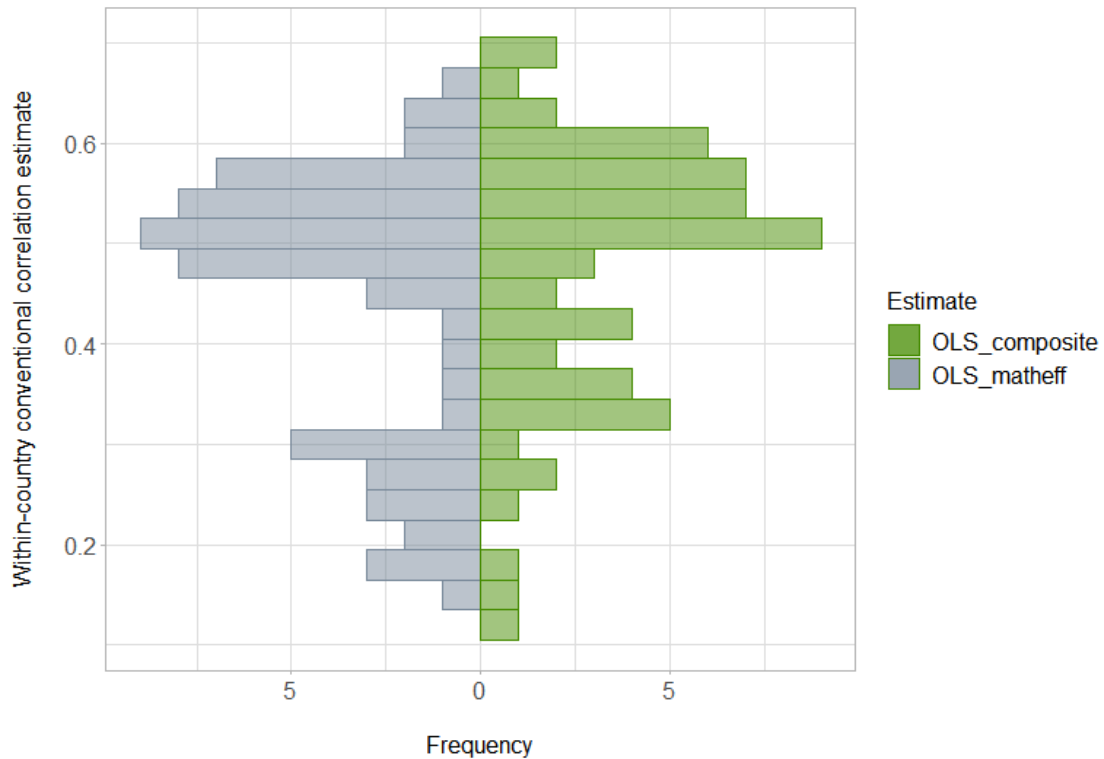


Figure 4.47. Distribution of the within-country conventional correlations employing the composite measure and the math self-efficacy measure alone (only one third of the sample was available for the models due to the rotated context questionnaire)

As is evident from the descriptive statistics and from the graph above, the distribution of the OLS_composite estimates shifted toward higher values relative to those employing math self-efficacy alone. Moreover, the shapes of the distributions were slightly different. The distribution of the OLS_composite estimates was less heterogeneous. Even though the distribution of the OLS_composite estimates still displayed bimodality, the two peaks were not as distinct as those evident in the distribution of the OLS_matheff estimates.

The relationships between correlation estimates and country-mean math performances were also examined to investigate if a similar pattern was seen when a composite background measure was employed in the models. The relationship between OLS_composite correlation estimates and country-mean math proficiencies was approximately linear up to 475 for country-mean math performance. In higher achieving countries (country-mean math performances above 475), the relationship was weakened and negative (the slope was -0.15). Figure 4.48 below displays a side-by-side comparison of OLS_matheff and OLS_composite correlation estimates and their relationship to country-mean math performances. The two scatterplots display similar patterns in lower achieving countries; the correlation between the background measure and math proficiency was stronger as the country-mean math proficiency was higher. In addition, the association was slightly weakened when the composite measure was used as the background measure instead of math self-efficacy measure alone.

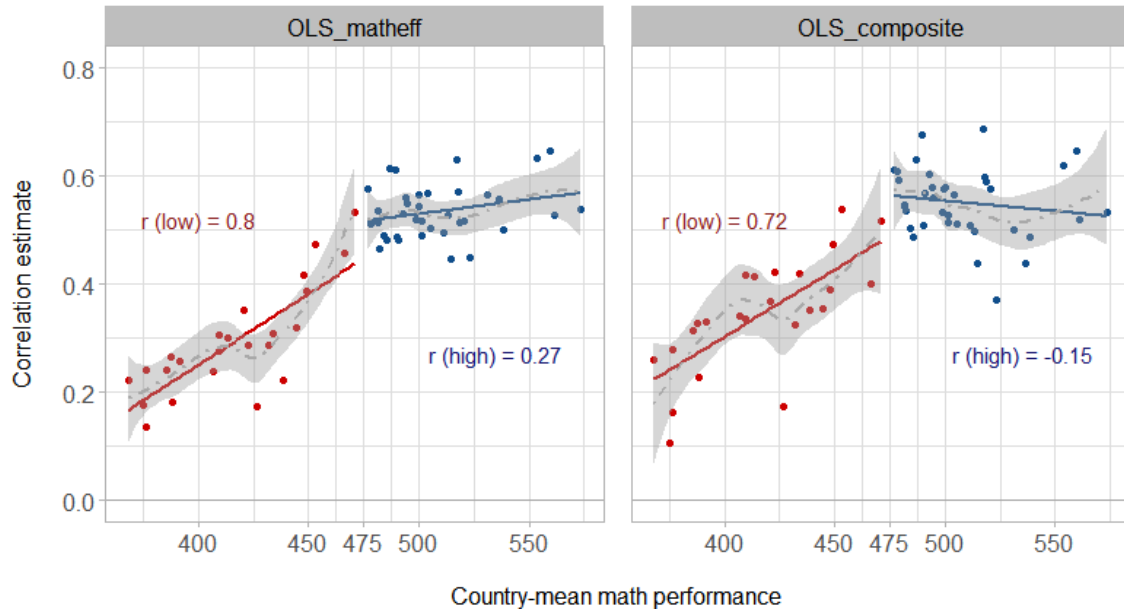


Figure 4.48. Comparison of the relationships between OLS correlation estimates and country-mean math performance (only one third of the sample was available for the models due to the rotated context questionnaire)

In higher achieving countries, on the other hand, the sign of the association between OLS estimates and country-mean math performance was reversed when the composite measure was used. Even though the association was weak, OLS_composite estimates were smaller for countries with higher mean math performances. Note that the fit of the linear regression line was worse when OLS_composite estimates were plotted against country-mean math performance and this was more apparent in higher achieving countries.

For the next step, multilevel modeling was employed to take clustering of the data into account. A set of two-level linear models in which the composite measure was used as a covariate at the student-level was fit to the data from each of the sixty-one participating countries. The distribution of standardized regression coefficients (i.e. the correlation estimates of the within-country relationships between the composite measure and math proficiency) were examined in comparison to those when math self-efficacy measure alone was employed as the

background measure. Estimates from the models employing the composite measure were labelled as “MLM_composite” and those from the models employing the math self-efficacy measure alone were labelled as “MLM_matheff”.

In order to better understand the impact of the reduction in the sample size and examine if patterns observed were similar to those obtained in Phase 3, MLM_matheff estimates were first compared to OLS_matheff estimates. Consistent with the results from Phase 3, the correlation estimates from the MLM models were typically smaller than those obtained from the OLS models. In addition, MLM_matheff estimates were more heterogeneously distributed than OLS_matheff estimates. The interquartile range for the MLM_matheff estimates was wider than the one for the OLS_matheff estimates (0.26 and 0.23, respectively). Nonetheless, the bimodality that was seen before almost disappeared, similar to the findings from Phase 3. Both sets of estimates demonstrated larger heterogeneity in comparison to those obtained in Phase 1 and Phase 3 due to the reduction in the sample size and proportionally larger uncertainty.

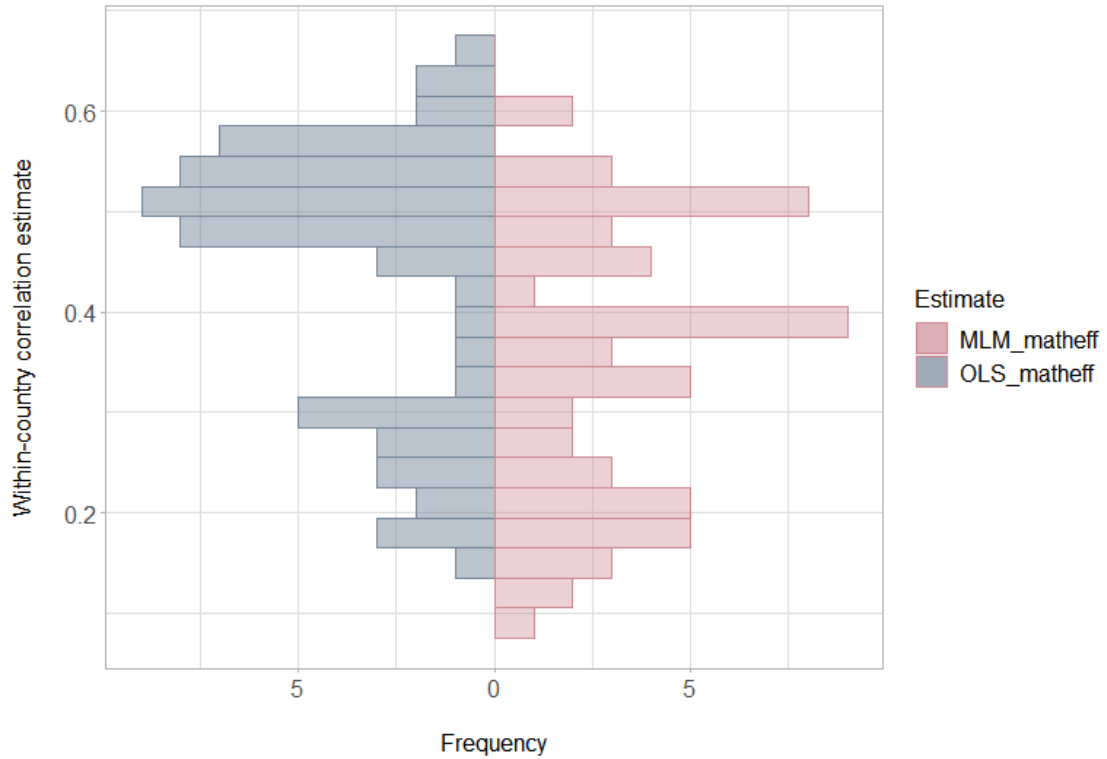


Figure 4.49. Distribution of the within-country OLS and MLM estimates employing the math self-efficacy measure alone (only one third of the sample was available for the models due to the rotated context questionnaire)

In relation to country-mean math performances, the correlation estimates obtained in this phase exhibited patterns that were very similar to those obtained in Phase 1 and Phase 3. Displayed in Figure 4.50, the association was again weakened after employing MLM models and the slopes were trivially smaller in magnitude in comparison to those obtained in Phase 1 and Phase 3 (Figure 4.4 and Figure 4.6). The increase in the heterogeneity in the distributions of the estimates was more apparent for higher achieving countries.

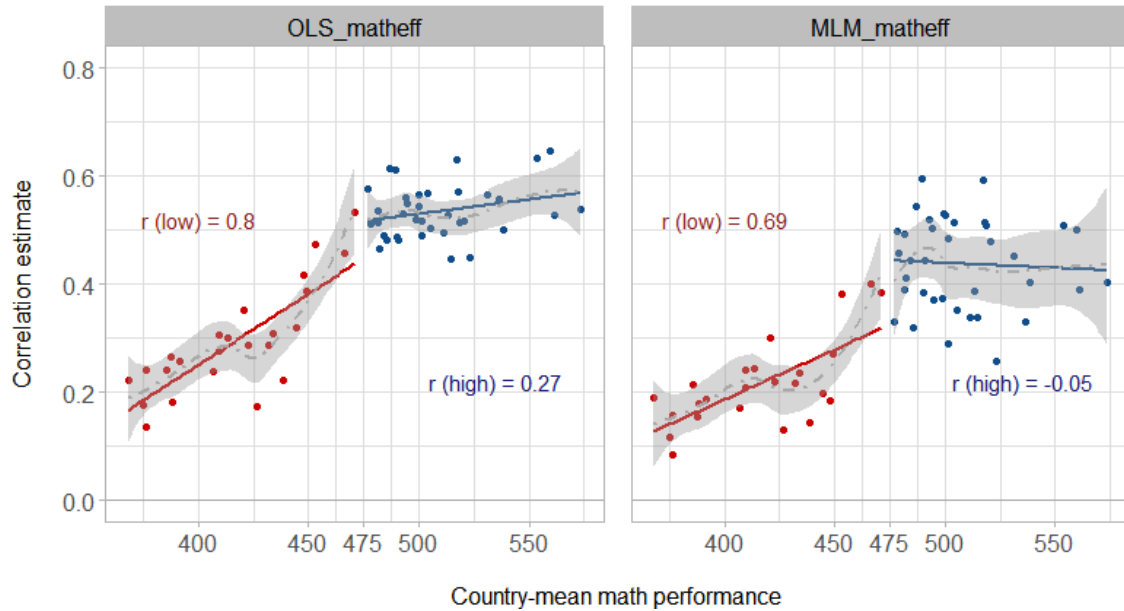


Figure 4.50. Comparison of the relationships between OLS_matheff and MLM_matheff estimates and country-mean math performance (only one third of the sample was available for the models due to the rotated context questionnaire)

Next, MLM_matheff correlation estimates were compared with MLM_composite estimates to examine the changes after employing a composite measure instead of math self-efficacy alone as a background measure. Descriptive statistics of the MLM estimates from the two separate sets of models are presented in Table 4.13. MLM_composite estimates were generally larger, ranging from 0.11 to 0.66. Moreover, the interquartile range for the MLM_composite estimates was narrower than the one for the MLM_matheff estimates (0.21 and 0.26, respectively). Note that the interquartile range for the MLM_composite estimates was also narrower than the one for the MLM estimates obtained in Phase 3 (0.21 and 0.22, respectively) despite the larger uncertainty caused by the reduction in the sample size.

Table 4.13. Descriptive statistics for the within-country correlation estimates from MLM models with the composite measure and with math self-efficacy measure alone

<u>MLM Estimates</u>	<u>mean</u>	<u>SD</u>	<u>min</u>	<u>median</u>	<u>max</u>	<u>range</u>	<u>skewness</u>	<u>IQR</u>
MLM_matheff	0.35	0.14	0.08	0.37	0.60	0.51	-0.12	0.26
MLM_composite	0.40	0.13	0.11	0.41	0.66	0.54	-0.12	0.21

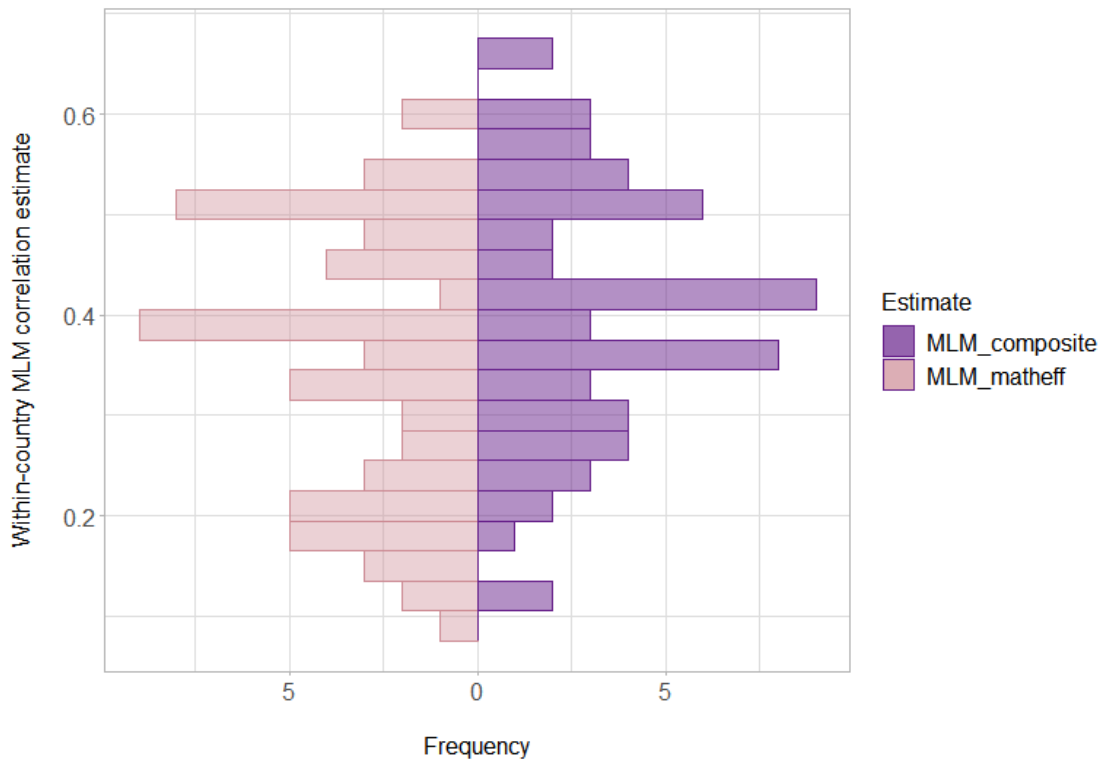


Figure 4.51. Distribution of the within-country MLM estimates employing the composite measure and the math self-efficacy measure alone (only one third of the sample was available for the models due to the rotated context questionnaire)

The histogram above illustrates that the distribution of the MLM_composite estimates was slightly shifted toward higher values in comparison to the distribution of MLM_matheff estimates. In addition, the distribution of the MLM_composite estimates was less dispersed than the distribution of the MLM_matheff estimates.

The two scatterplots below demonstrate the relationship between correlation estimates and country-mean math performance. In lower achieving countries, the relationship between MLM correlation estimates and country-mean math performance was still positive and approximately linear but weakened when the composite measure was employed (the slopes were 0.69 and 0.55, respectively). In higher achieving countries, on the other hand, the relationship was negative and larger in absolute value when the composite measure was employed (the slopes were -0.05 and -0.24, respectively).

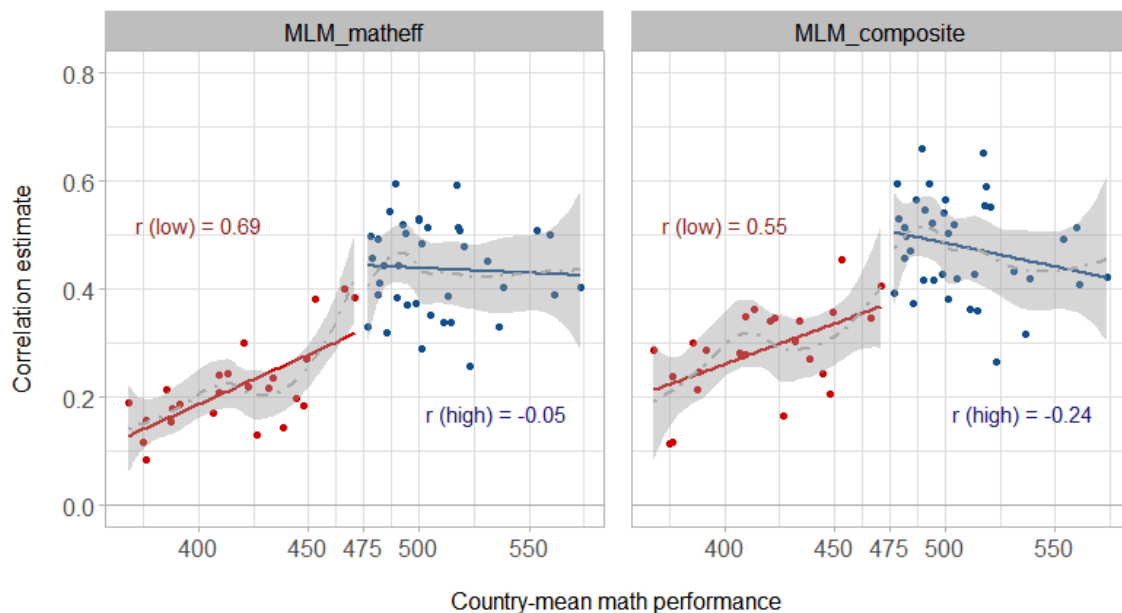


Figure 4.52. Comparison of the relationships between MLM correlation estimates and country-mean math performance (only one third of the sample was available for the models due to the rotated context questionnaire)

The distributions of the MLM_composite estimates and the OLS_composite estimates were also compared to examine the impact of clustering on the correlation estimates. Similar to the patterns seen previously, correlation estimates were generally smaller when clustering was taken into account by employing multilevel modeling. Moreover, the shapes of the distributions

were slightly different for OLS_composite and MLM_composite estimates. Similar to the previous findings, the bimodality that was seen before was not as apparent.

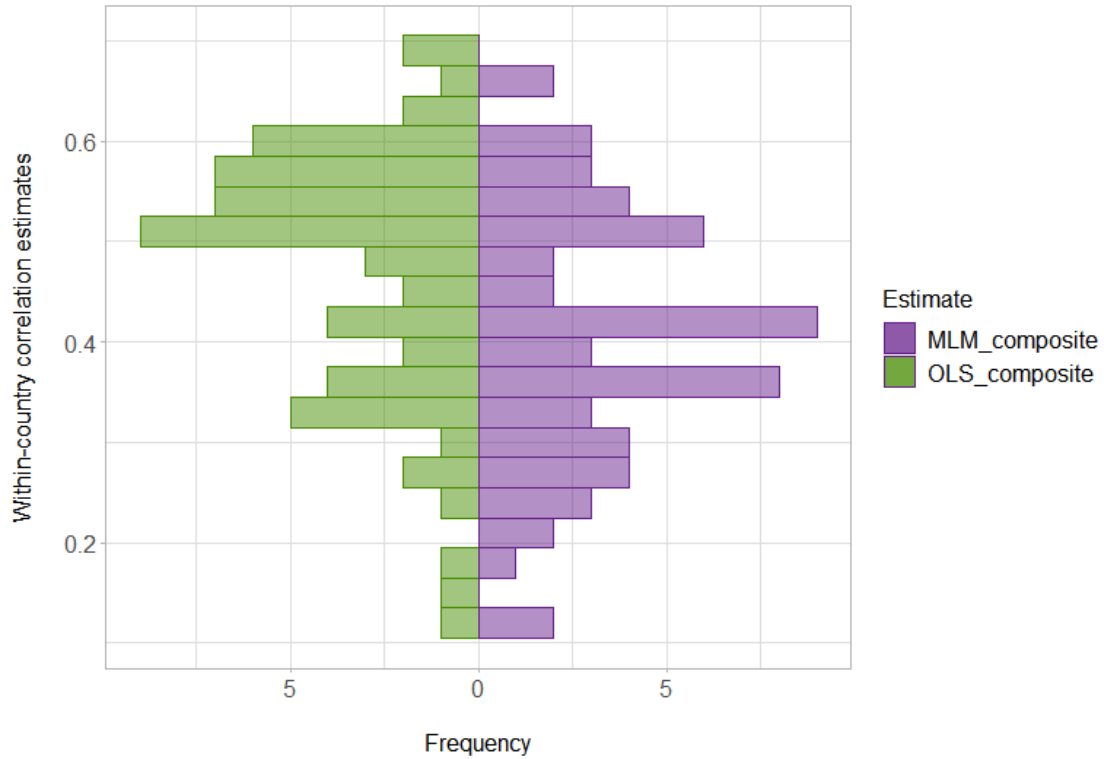


Figure 4.53. Distribution of the within-country OLS and MLM estimates employing the composite measure (only one third of the sample was available for the models due to the rotated context questionnaire)

Figure 4.54 below displays a side-by-side comparison of OLS_composite and MLM_composite correlation estimates plotted against country-mean math performances. In higher achieving countries, the association was negative and stronger when multilevel modeling was employed (the slopes were -0.15 and -0.24, respectively). The relationship in lower achieving countries was still positive but weakened after multilevel modeling was employed. It is important to note that the fit of the linear regression lines were worse than those that were obtained from OLS models and, again, the heterogeneity was more apparent in higher achieving countries.

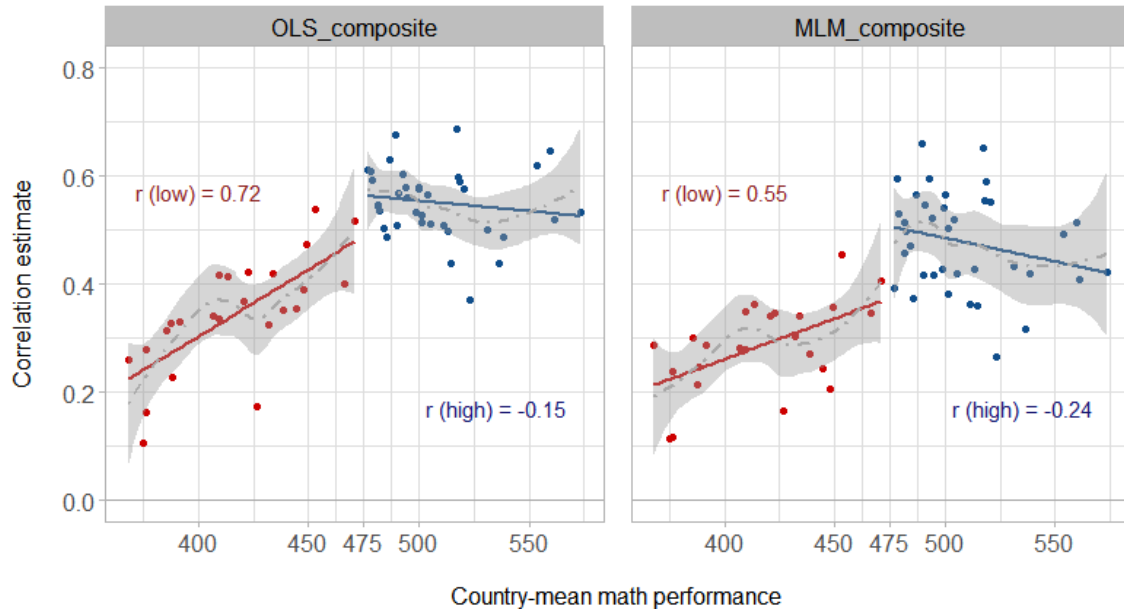


Figure 4.54. Comparison of the relationships between correlation estimates and country-mean math performance (only one third of the sample was available for the models due to the rotated context questionnaire)

The changes in the correlation estimates after employing multilevel modeling were examined in relation to within-country ICCs. As Figure 4.55 demonstrates, the changes were larger for countries with large ICCs. Furthermore, the relationship was approximately linear with a negative slope (-0.66).

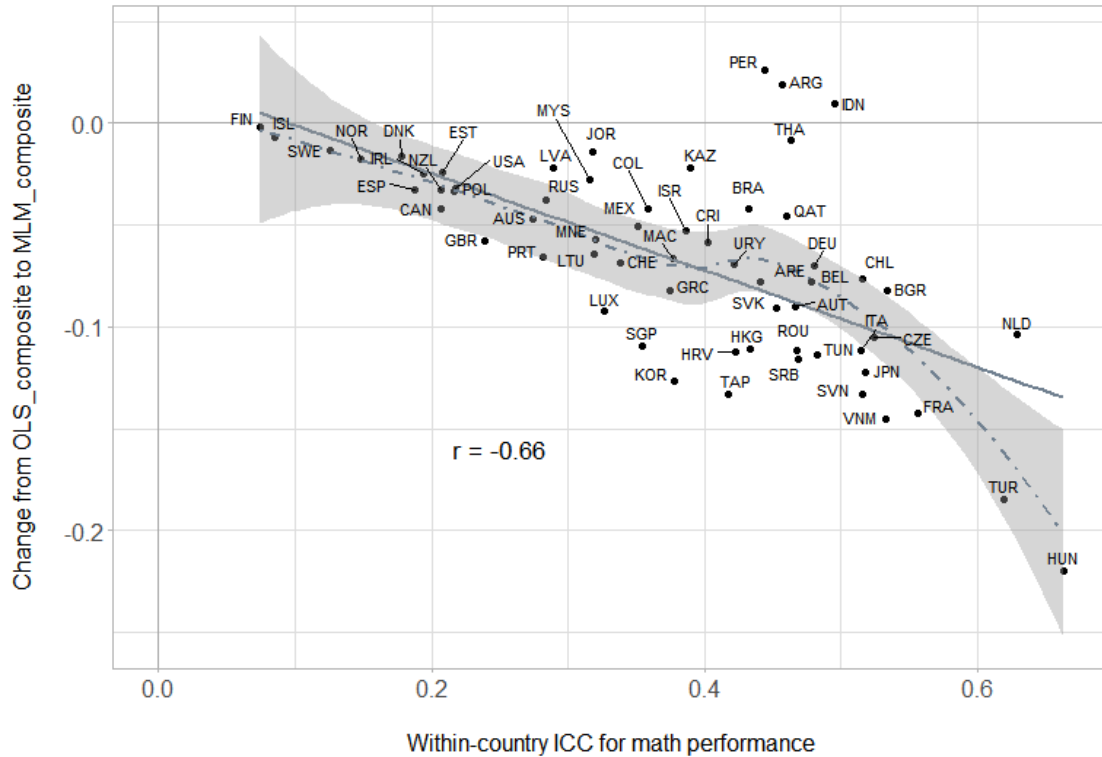


Figure 4.55. Relationship between changes in the correlation estimates from OLS to MLM and within-country intra-class correlations for math proficiency outcome (only one third of the sample was available for the models due to the rotated context questionnaire)

For example, the correlation between students' math performance and math self-efficacy almost stayed the same for Finland, as it had the smallest ICC across all countries in the dataset. By contrast, Hungary, with the largest ICC, had the largest change in the correlation estimate from conventional analyses to MLM (a decrease of 0.22). These results are very similar to those found in Phase 3.

Chapter 5: Discussion

5.1 Summary of Findings

The goal of this study was to investigate the impact of measurement error and clustering on correlations reported between background data and proficiency scales in ILSAs. As discussed in depth in Chapter 2, statistical artefacts such as heterogeneous measurement error and nested data are shown to impair the accuracy of correlation estimates. In this regard, a step-wise model refinement approach was employed to explore the operating characteristics of various modeling techniques through application to data from PISA 2012. The dataset contained 468,200 students from 61 participating countries. The estimates of correlations between math self-efficacy and math achievement across countries were the focus of this study.

Within-Country Relationships

Conventional Estimates (Phase 1)

Despite advances in methodology, ordinary regression models are commonly used when analyzing ILSA data. Therefore, as the baseline for comparisons across models in the model-refinement process, OLS models were fit to the data from each country and standardized regression coefficients between math proficiency and math self-efficacy were examined. Substantial variation was observed across countries. Although all were positive, the distribution of conventional within-country correlations was bimodal; rather weak relationships were observed in certain countries such as Colombia ($r_{OLS} = 0.15$) and Indonesia ($r_{OLS} = 0.17$) whereas rather strong relationships were observed in others such as Chinese Taipei ($r_{OLS} = 0.66$) and Portugal ($r_{OLS} = 0.63$).

At the between-country level, the relationship between country-mean math proficiency and country-mean math self-efficacy was positive ($r=0.55$) and no attitude-achievement paradox was observed, consistent with the findings from previous studies on math self-efficacy (Lee, 2009). However, there were outliers such as Japan and Korea, two countries exhibiting low levels of math self-efficacy at the country-level despite their high mean math performances. When examined at the within-country level, these two countries displayed strong, positive relationships between math proficiency and math self-efficacy (the correlation estimates were 0.58 and 0.62, respectively).

Although variation across countries is expected to some extent, it is concerning to observe such big differences in within-country relationships given that a strong and positive relationship between math self-efficacy and math performance is well established in the literature. It has been discussed in the literature that such a heterogeneity across countries may result from imperfect measurement, particularly in lower achieving and developing countries (Rutkowski & Rutkowski, 2010). In line with this argument, conventional correlation estimates varied substantially, especially in lower achieving countries, and a strong, positive relationship was observed between correlation estimates and lower achieving countries' mean math achievement ($r_{(low)} = 0.82$). In higher achieving countries (i.e., country-mean math performance greater than 475), correlation estimates were more homogeneous and their relation to mean math performance was substantially weaker ($r_{(high)} = 0.28$).

Accounting for Measurement Error Only (Phase 2)

Students' math performance and math self-efficacy levels are not directly observed quantities but, instead, they are latent constructs measured by means of students' responses to various questions. Therefore, if the uncertainty in these variables were ignored in the models, the

resulting within-country correlation estimates would be biased. In addition, substantial differences in the amount of measurement error in the data across countries can lead to greater heterogeneity in the correlation estimates. To this end, multidimensional IRT modeling was employed to determine if the resulting correlation estimates displayed greater homogeneity across countries than the conventional correlation estimates obtained from the OLS models in Phase 1.

Displayed in Figure 4.5, the distribution of MIRT correlation estimates was also bimodal and rather heterogeneous. The interquartile range was only slightly narrower than the one for the distribution of conventional estimates (0.20 and 0.22, respectively). However, MIRT estimates were typically larger than the conventional estimates. Furthermore, similar to the conventional correlation estimates, MIRT estimates were positively correlated with county-mean math performance. In comparison to patterns observed in Phase 1, the relationship was still strong and positive for lower achieving countries, but slightly weakened after employing MIRT modeling (the slopes were 0.82 and 0.78, respectively). In higher achieving countries, the relationship was weakened considerably and the slope was effectively zero (0.05).

The distribution of empirical reliabilities for the math proficiency scale were noticeably more heterogeneous and skewed towards smaller values in lower achieving countries in comparison to higher achieving countries. For lower achieving countries with average math proficiency scores smaller than 475, the slope of the relationship between country-mean math performance and empirical math reliability was 0.92. In higher achieving countries, this relationship was weaker but still strong and positive (the slope was 0.58). These findings indicate a disjuncture between the test information function and the distributions of skills in many countries, suggesting that the math proficiency scale was particularly less precise for lower levels

of math proficiency. Furthermore, MIRT correlation estimates between math proficiency and math self-efficacy were smaller for countries exhibiting lower reliabilities for the math proficiency scale (Figure 4.15) and this was more evident in lower achieving countries in comparison to higher achieving countries.

Overall, empirical reliabilities for the math self-efficacy scale did not display substantial variation across countries. However, in lower achieving countries they were more homogeneously distributed in comparison to those in higher achieving countries (Figure 4.12). In addition, there was a weak or null relationship between country-mean math performance and math self-efficacy scale reliability both in lower and in higher achieving countries (the slopes were 0.24 and -0.02, respectively). When examined in relation to countries' average math self-efficacy, a strong, negative relationship was observed ($r_{\text{(high)}} = -0.55$). Countries with lower levels of math self-efficacy such as Japan and Korea possessed higher empirical reliabilities for the math self-efficacy scale. By contrast, there was no evident association between country-mean math self-efficacy and reliabilities for the math self-efficacy scale in lower achieving countries. Nevertheless, similar to the findings for the math proficiency scale, MIRT correlation estimates between math proficiency and math self-efficacy were weaker for countries that possessed lower empirical reliabilities for the math self-efficacy scale and this relationship was more evident in lower achieving countries. These findings suggest that correlations were attenuated when the amount of noise in the data was larger.

Accounting for Clustering Only (Phase 3)

PISA employs a two-stage stratified sampling design, which involves first sampling schools that have eligible 15-year-old students and then selecting a random sample of students from these randomly sampled schools (OECD, 2014). The use of multi-stage sampling technique

results in a nested data structure and this has certain implications in statistical analysis as described in Chapter 2. Therefore, two-level linear modeling was employed to account for clustering of the data and determine if the estimates of within-country relationships between math self-efficacy and math achievement have greater homogeneity across countries than that displayed by the conventional correlation estimates obtained in Phase 1. The intra-class correlations (ICCs) obtained from the unconditional models were larger than 15% in most of the countries, indicating that multilevel modeling was an appropriate model choice.

Within-country correlation estimates obtained from the MLM models were typically smaller than both the conventional estimates and the MIRT estimates. Because the MLM induces a range restriction on both the predictor and the outcome if between-group variance is substantially larger than within-group variance (i.e., large ICC), these findings were expected. The correlation estimates still showed substantial variability across countries but they were more homogeneously distributed. The bimodality and skewness towards larger values that were observed in the distribution of the estimates from the OLS and MIRT models almost disappeared.

Similar to OLS and MIRT correlation estimates, MLM estimates were strongly correlated with countries' average math performances ($r_{(low)} = 0.71$) for lower achieving countries, but this relationship was slightly weaker in comparison to OLS and MIRT estimates. In higher achieving countries, no evident association was observed between correlation estimates and country-mean math performance and the slope of the relationship was effectively zero, similar to the findings from Phase 2. Furthermore, the decreases in the correlation estimates from the OLS models to the MLM models were larger for countries that possessed larger ICCs (the slope was -0.79). This

finding suggests that the impact on the correlation estimates are moderated by the degree of clustering in the data.

Accounting for Both Measurement Error and Clustering (Phase 4)

In the final step of the model refinement process, multilevel multidimensional mixture IRT modeling was employed to account for both measurement error and clustering in the data. The goal was to determine if the estimates of within-country relationships between math self-efficacy and math achievement had greater homogeneity across countries than that displayed by the conventional correlation estimates or the ones obtained from the models that account for either measurement error or clustering only. The Q-matrix that was developed in Phase 2 to fit the two-dimensional IRT models was also used to fit MLMixMIRT models. The models were constrained to have two latent classes based on previous research (von Davier, 2005a; von Davier, Yamamoto, et al., 2019). A smaller number of latent classes was preferred so that resulting latent classes could be interpreted similarly across countries. The models yielded two latent classes that consistently differed from each other by substantial differences in mean math ability estimates. Note that this is not a mathematical necessity of any MLMixMIRT model. The model was specified so that it should find two latent classes based on similarities and differences in response patterns. The results showed that the underlying latent class structure based on the patterns seen in students' responses were driven by overall math performance. Accordingly, the resulting latent classes were labelled "High Math Class" and "Low Math Class". The estimates of within-country relationships between math self-efficacy and math achievement were then compared between two latent classes as well as across models.

Correlation estimates for the High Math Class were more homogeneously distributed whereas those for the Low Math Class varied substantially. In fact, the estimates for the Low

Math Class exhibited the largest heterogeneity across all models employed in this study (Table 4.6). Moreover, the High Math Class demonstrated typically larger correlations between math proficiency and math self-efficacy in comparison to the Low Math Class. When compared to the correlation estimates obtained from the OLS, MIRT, and MLM models in previous phases, the estimates for the High Math Class were still substantially larger and more homogenous.

MLMixMIRT correlation estimates were examined in relation to both country-mean math proficiency and average math proficiency estimates by latent class. The results showed that there was a positive relationship between countries' average math proficiency and the correlation estimates between math proficiency and math self-efficacy in both High Math Class and Low Math Class (Figure 4.23). Furthermore, this positive relationship was noticeably stronger for Low Math Class (the slopes were 0.85 and 0.32, respectively). When examined in more detail, the slope for the High Math Class was effectively zero (0.02) in higher achieving countries but still positive (0.36) in lower achieving countries. These findings were consistent with those observed in previous phases. Even after employing MLMixMIRT modeling that accounts for both clustering and measurement error in the data, in lower achieving countries the correlations between math proficiency and math self-efficacy were still moderated by population-level performance. This implies that there were likely other factors, such as students' lower levels of meta-cognition (i.e., the Dunning-Kruger effect), that contributed to weaker correlations in lower performing countries.

EAP reliabilities derived from the MLMixMIRT model estimation revealed that the math proficiency scale was less precise at lower levels of math proficiency, in line with the results from the MIRT models in Phase 2. For the Low Math Class, estimated reliabilities for the math proficiency scale were typically smaller in comparison to those for the High Math Class. In

general, empirical reliabilities for the math proficiency scale were larger when the mean math proficiency was greater. This pattern was also more apparent for the Low Math Class (Figure 4.27 and Figure 4.28). These findings were consistent with those from the MIRT models employed in Phase 2.

Similar to the findings for the math proficiency scale, empirical reliabilities for the math self-efficacy scale were more heterogeneously distributed for the Low Math Class. However, reliabilities for the High Math Class were skewed towards smaller values in comparison to those for the Low Math Class. Furthermore, reliabilities for the math self-efficacy scale were negatively correlated with country-mean math self-efficacy levels; that is, countries with higher math self-efficacy levels typically displayed lower reliabilities for both latent classes (Figure 4.31). These findings were also consistent with those from the MIRT models employed in Phase 2.

For both the Low Math Class and the High Math Class, correlation estimates between math proficiency and math self-efficacy were larger for countries exhibiting greater reliabilities for the math proficiency scale. These findings were in line with the findings from the MIRT models employed in Phase 2, suggesting that correlations were attenuated when the amount of noise in the data was larger. Moreover, for the High Math Class, MLMixMIRT correlation estimates were again typically larger when the empirical reliabilities for the math self-efficacy scale were greater. For the Low Math Class, however, countries that possessed larger reliabilities for the math self-efficacy scale exhibited weaker correlations between math self-efficacy and math achievement.

Patterns of Relationships across Subgroups Within-Country (Phase 5)

One of the main hypotheses of this study was that levels of math proficiency might moderate the relationship between math proficiency and math self-efficacy. Employing advanced modeling techniques that account for both measurement error and clustering in the data was expected to mitigate this moderation. The findings from the previous phases revealed that math proficiency still played a role, though somewhat diminished, in the magnitude of correlation estimates, even after taking measurement error and clustering into account in the models. In this phase of the study, a supplementary analysis was conducted to explore if school-level correlations displayed similar patterns in relation to math proficiency. Unfortunately, preliminary analyses indicated that within-country school-level sample sizes were not sufficient to estimate complex models. Hence, as an alternative strategy, schools within each country were grouped into nine “super-schools” that were ordered based on their average math proficiency levels and Phase 1 and Phase 2 were repeated at the super-school level.

In order to limit the volume of the analyses, a set of countries exhibiting a wide range of country-level correlations were selected. Countries were ordered based on country-level conventional correlations and six countries were selected from the bottom, middle, and top quartiles: Kazakhstan ($r_{OLS} = 0.28$), Jordan ($r_{OLS} = 0.29$), Greece ($r_{OLS} = 0.48$), Netherlands ($r_{OLS} = 0.48$), Portugal ($r_{OLS} = 0.64$), and Chinese Taipei ($r_{OLS} = 0.66$). In each country, schools falling into the bottom 15% of the distribution were grouped into *super-school 1*, those in the top 15% were grouped into *super-school 9*, and the remaining schools were grouped into seven super-schools, each containing approximately 10% of the sample.

Conventional correlations between math proficiency and math self-efficacy at the super-school level within a country were larger for countries with larger country-level correlations.

Interestingly, for the top four countries with larger country-level correlation estimates, the maximum correlation estimate for the super-schools was close to the country-level correlation estimate. Furthermore, except for Chinese Taipei, a positive and approximately linear relationship at the super-school level was observed between math proficiency and the correlation estimates, consistent with the findings from previous phases. In Chinese Taipei, the highest achieving country among the selected countries, correlations were weaker when the math proficiency level of the super-school was higher (the slope was -0.69). In particular, a substantial drop in correlation estimates was observed for super-schools 8 and 9 in Chinese Taipei.

In agreement with the country-level findings from Phase 2, MIRT correlation estimates between math proficiency and math self-efficacy were larger than conventional estimates at the super-school level. Except for Chinese Taipei, higher achieving super-schools still exhibited larger correlations but the slopes were weakened in all five countries in comparison to those obtained from the OLS models. Chinese Taipei, again, demonstrated a negative relationship between math proficiency and correlation estimates (the slope was -0.71) even after measurement error was taken into account. This paradoxical relationship indicates that higher performing students' self-assessments of their math proficiency were lower in Chinese Taipei. Although this finding is at odds with the patterns evident in other countries, it is in line with findings from previous studies on response style bias in certain cultures (Han et al., 2015; He & van de Vijver, 2015) and modesty bias among higher performers (Ehrlinger et al., 2008; Min et al., 2016).

Overall, empirical reliabilities obtained from the MIRT models for the math proficiency scale demonstrated substantial variation within and across countries. Super-schools in Kazakhstan and Jordan, the two lowest achieving countries, particularly possessed lower reliabilities for the

math proficiency scale. In addition, super-school 1 (i.e., schools falling into the bottom 15% of the distribution based on mean math achievement) in both Greece and Portugal demonstrated unusually low reliabilities for the math proficiency scale. These findings are in agreement with those observed in previous phases, indicating that the math proficiency scale was less precise at lower levels of proficiency. The highest performing super-school in Chinese Taipei also possessed unusually low reliability for the math proficiency scale. Greater uncertainty is expected in the tails of the proficiency scale, especially when a fixed set of items is used to measure a wide range of ability levels. However, the discrepancy between actual ability levels and the levels that the math proficiency scale used in PISA 2012 targets is arguably more problematic for lower levels of math ability. It was also evident, especially in lower achieving countries, that correlation estimates were larger when reliabilities for the math proficiency scale were larger.

Consistent with the findings from Phase 2 and Phase 4, empirical reliabilities for the math self-efficacy scale demonstrated larger variation in higher achieving countries such as Portugal and Chinese Taipei. This relationship was also evident when reliabilities for the math self-efficacy scale were examined in relation to mean math performance levels by super-schools. A negative and approximately linear relationship was observed in all countries and the slopes were larger in magnitude in higher achieving countries. In addition, estimated reliabilities for the math self-efficacy scale were lower when mean math self-efficacy was higher. These findings were in agreement with those from Phase 2 and Phase 4, suggesting ceiling effects resulting from low or no variance in students' responses in groups with higher levels of math self-efficacy. Lower reliabilities for the math self-efficacy scale were also associated with larger correlation estimates between math proficiency and math self-efficacy. Note that lower reliabilities were typically

observed in higher achieving super-schools, which typically exhibited larger correlations between math proficiency and math self-efficacy.

Employing a Composite Background Variable (Phase 6)

Given that there have been numerous concerns raised in the literature regarding the background measures (Buckley, 2009; He & van de Vijver, 2015; Hopfenbeck & Maul, 2011; Kyllonen & Bertling, 2014; Rutkowski & Rutkowski, 2010), this study investigated if the impact of measurement error on correlation estimates could be mitigated by employing an attitudinal background measure comprising multiple indicators. To this end, a composite measure that comprised math-self-efficacy, math anxiety, and math self-concept was created by conducting principal component analyses within each country separately. The newly created composite measure was, then, employed in the models to compare the changes in the correlation estimates to those observed when math self-efficacy was used as a single background measure. Although employing a composite measure led to a change in the construct and has a different substantive interpretation, this procedure was proposed primarily as a methodological investigation. It should also be kept in mind that only one third of the sample could be employed in this phase due to the use of rotated design for the context questionnaire in PISA 2012 (OECD, 2013a). Hence, the uncertainty in the estimates was potentially larger due to the reduction in the sample size and this should be taken into consideration when interpreting the results from this phase.

First, within each country, two separate OLS models were fit to the data; one employing math self-efficacy alone (OLS_matheff) and the other employing the composite measure (OLS_composite) as the background variable. OLS_composite correlation estimates were larger in general and exhibited less heterogeneity. In addition, the distribution of OLS_matheff

estimates displayed bimodality, similar to the one obtained from the OLS models in Phase 1. However, the two peaks in the distribution of the estimates were not as distinct when the composite measure was employed (Figure 4.47).

Consistent with the findings from previous phases, in lower achieving countries, a positive and approximately linear relationship was observed between country-mean math performance and the correlation estimates but the relationship was weaker when the composite measure was employed in the models (the slopes were 0.80 and 0.72, respectively). In higher achieving countries, OLS_matheff estimates exhibited a weak but positive relationship (the slope was 0.27), a pattern similar to the one obtained from the OLS models in Phase 1. However, when the composite measure was employed in the models, a weak but negative relationship was observed (the slope was -0.15).

The impact of employing a composite measure on the estimates was investigated further by employing multilevel modeling to take clustering in the data into account. As with the OLS models, within each country, two separate MLM models were fit to the data; one employing math self-efficacy alone (MLM_matheff) and the other employing the composite measure (MLM_composite) as the background variable. MLM_composite correlation estimates were generally larger. This is somewhat consistent with the findings from previous phases. Taking clustering of the data into account typically led to smaller estimates whereas correcting for measurement error yielded larger estimates. Because the composite measure theoretically contains a smaller amount of measurement error in comparison to the math self-efficacy scale, observing larger estimates when the composite measure was employed was expected. Moreover, the interquartile range for the MLM_composite estimates was not only narrower than the one for the MLM_matheff estimates but also narrower than the one for the MLM estimates obtained in

Phase 3. This suggests that despite the larger uncertainty caused by the reduction in the sample size in this phase, the correlation estimates were more homogeneously distributed when the composite measure was employed in the models.

In lower achieving countries, the relationship between MLM correlation estimates and country-mean math performance was still positive but weaker when the composite measure was employed (the slopes were 0.69 and 0.55, respectively). In higher achieving countries, on the other hand, the relationship was negative and larger in absolute value (the slopes were -0.05 and -0.24, respectively). Note that OLS_composite estimates and MLM_composite estimates both demonstrated negative, albeit weak, relationships between country-mean math performances in higher achieving countries but the relationship was stronger when MLM was employed. These findings for higher achieving countries were somewhat inconsistent with those from previous phases in the study. OLS estimates obtained in Phase 1 were positively correlated with country-mean math achievement in higher achieving countries but the relationship was weaker (the slope was 0.28). When measurement error was taken into account in Phase 2, the slope of the relationship was effectively zero (0.05). When MLM models were fit to the data to take the clustering into account, the slope of the relationship was again effectively zero (-0.04). Finally, when both measurement error and clustering were taken into account with MLMixMIRT models, the slope of the relationship was once again effectively zero (0.02) for the High Math Class in higher achieving countries. Therefore, observing a negative relationship between the correlation estimates and mean math performances after employing the composite measure may be due to the change in the construct and greater sampling error resulting from employing one third of the sample.

Conclusions

The findings from this study offer both methodological and substantive takeaways. The step-wise refinement of the models conducted in the first four phases showed that conventional correlations might provide an incomplete picture and the resulting inferences could be puzzling or even misleading. The distribution of conventional correlation estimates obtained from the OLS models in Phase 1 demonstrated a large spread suggesting that the relationships between math proficiency and math self-efficacy were as weak as 0.15 in certain countries whereas as strong as 0.66 in others, as expected. When examined in relation to country-level math performances, weaker correlations were mostly observed in lower achieving countries. This finding is consistent with previous studies that lower performers tend to misjudge their own knowledge and skills due to lack of metacognitive competence (Kruger & Dunning, 1999), resulting in weaker correlations between their math self-efficacy and their math performances. Moreover, a moderate correlation (0.51) was found between conventional correlation estimates and within-country variances of the math self-efficacy variable, suggesting that restricted range of the predictor might be a factor behind the heterogeneity observed in conventional correlation estimates. If these were considered alone, one might conclude that certain cultural factors lead to such big differences across countries or could not be sure if these were due to certain statistical artefacts. Indeed, many argued that the Dunning-Kruger effect might result from artefacts such as measurement error or restriction of range (Ehrlinger et al., 2008; Krueger & Mueller, 2002).

The MIRT and the MLM models served as vehicles to understand the characteristics of the PISA data and how impactful the measurement error and clustering could be on the correlation estimates. The findings from the MIRT models demonstrated that the amount of measurement error varied across countries and larger amounts of noise in the data led to

attenuated correlation estimates. In addition, a positive relationship was found between the empirical reliabilities of the math proficiency scale and country-level math performances, indicating a mismatch between the skill distributions in many countries and the test information function. No evident relationship was observed between country-level mean math performances and the empirical reliabilities for the math self-efficacy scale. However, countries with higher levels of mean math self-efficacy exhibited lower reliabilities for the math self-efficacy scale. This may indicate that the math self-efficacy scale did not function as well for higher levels of math self-efficacy as for the lower levels of math self-efficacy, consistent with ceiling effects. The findings from the MLM models showed that the ICCs based on the cognitive outcome ranged from 15% to 65% across countries and taking clustering into account led to changes in the estimates as large as 0.3 in countries with large ICCs. These findings suggested that both measurement error and clustering were impactful on the estimates and should be accounted for to obtain unbiased estimates of correlations.

The MLMixMIRT models employed in Phase 4 not only properly accounted for measurement error and clustering in the data, as the characteristics of the PISA data require, but also took the possibility of distinct underlying subpopulations within countries into account. The correlation estimates between math self-efficacy and math proficiency were moderated by mean math performance even after measurement error and clustering were accounted for. Even in higher achieving countries, a weaker association between math self-efficacy and math proficiency in lower achieving groups was consistently seen across countries. Higher performing groups in both lower and higher achieving countries exhibited larger empirical reliabilities for the math proficiency scale and more homogeneously distributed correlations between math proficiency and math achievement. Hypothesizing a mixture of subpopulations within each

country by employing the MLMixMIRT modeling helped understand underlying factors behind the substantial variation observed in conventional correlations across countries in Phase 1 and shed light on how this pattern emerges from greater randomness in the responses of lower performing groups.

In alignment with the findings from Phase 4, the results from the analyses conducted in Phase 5 yielded similar patterns of relationships at the within-country level. Higher achieving groups of schools exhibited stronger correlations between math proficiency and math self-efficacy in five out of six countries chosen for this phase of the analyses. These findings support the hypothesis that the relationship between math proficiency and math self-efficacy was moderated by math performance. In Chinese-Taipei, however, weaker correlations were observed in higher achieving groups of schools. This finding requires further investigation to examine if response style bias or modesty bias explains observing such a different pattern in Chinese-Taipei. Finally, the findings from Phase 6 provided more insights on the impact of measurement error on the correlation estimates and indicated that employing an attitudinal background measure comprising multiple indicators could mitigate the impact of measurement error on the estimates.

5.2 Limitations and Future Research

This study carried out a comprehensive, methodological investigation of the impact of measurement error and clustering on correlation estimates by conducting a systematic sequence of statistical modeling techniques. Nonetheless, it has certain limitations that should be noted and further investigated. First, this study employed data from 61 PISA 2012 countries as an instructive example to explore the impact of heterogeneous measurement error and clustering on

the statistical inferences. These statistical artefacts are not unique to the PISA data and they are particularly common in data from ILSAs. Conducting similar investigations by means of applications to data from other ILSAs such as TIMSS can support some of the findings. For instance, one of the main findings of this study was the moderating impact of population-level math performance on the correlations between math self-efficacy and math achievement from lower performing countries, even after taking measurement error and clustering into account. This should be further investigated to better understand the factors contributing to such differences between lower performing and higher performing countries.

Moreover, findings from the MIRT and MLMixMIRT models revealed that the empirical reliabilities for the math proficiency scale varied across countries and data from lower performing countries possessed larger amounts of measurement error. This warrants further investigation of the precision of the scale in capturing the construct in lower performing countries. Careful examination of the test information functions can offer more insight into these findings. As mentioned before, when the same set of items is used to assess a wide range of ability levels, measurement of the construct is expected to be less than optimal, particularly in the tails of the skill distribution it targets. However, this can create various obstacles. For instance, Tijmstra and others (2020) demonstrated that the root-mean square deviation, which is commonly used to detect country-specific item misfit and DIF in ILSAs including PISA, could be less sensitive in lower performing countries due to the misalignment between the test information function and the skill distributions of those countries. Consequently, this is a threat to the ability to detect DIF and ensure the measurement invariance across countries that is essential to the ILSA claim of the comparability of measurement (Tijmstra et al., 2020).

In the Phase 4 analyses, the MLMixMIRT models were constrained to have two latent classes based on previous research (von Davier, 2005a; von Davier, Yamamoto, et al., 2019) and for keeping the model structure the same across countries ensuring the interpretability of the resulting latent classes. Testing models with larger numbers of latent classes was beyond the scope of this study but should be investigated in future research. In addition, the hierarchical mixture distribution used in the MLMixMIRT models does ensure that the latent classes are the same across schools within countries, but not across countries. Hence, although resulting latent classes were interpreted the same way across countries, more nuanced interpretations were possible for the latent class structure within each country and should be considered in future studies. It should be emphasized that the MLMixMIRT modeling was employed in this study not because it is a more complex and advanced approach but because it accounts for the structure and characteristics of the PISA data. Nevertheless, indirect support for the models can be further considered by examining proper goodness of fit statistics. Furthermore, exploratory analyses of within-country latent class sizes did not suggest any evident pattern but further investigation can offer more insight to better interpret the findings from the MLMixMIRT models.

Furthermore, the original analysis plan for Phase 5 was to replicate the step-wise model refinement approach (that was carried out at the country-level) at the school-level in order to determine if school-level correlations exhibited similar patterns in relation to math proficiency. However, the data did not support employing complex modeling techniques with so many parameters to be estimated, given the small sample sizes within schools. In order to avoid the convergence issues encountered when running complex models, schools within each country were grouped based on their average math performances to conduct further analyses. Although super-school level analyses provided some valuable insights, the grouping into super-schools did

not support employing MLMixMIRT modeling to account for clustering. These issues highlight some of the obstacles with employing advanced modeling techniques.

One of the findings from Phase 5 was the negative relationship that was observed in Chinese Taipei between correlation estimates and mean math performance. This result was inconsistent with what was observed in other countries, as well as those obtained from the county-level analyses. This may be signaling the modesty bias that is common among top performers (Ehrlinger et al., 2008; Min et al., 2016). However, considering the limitations faced in conducting the Phase 5 analyses, this finding should be further investigated.

Finally, employing a composite measure in Phase 6 led to a two-thirds reduction in sample size due to the use of rotated design for the context questionnaire in PISA 2012 (OECD, 2013a). Previous studies emphasized methodological concerns about possible biases resulting from the use of a rotated design for the context questionnaire (von Davier, 2014). Regardless, the comparisons made between the findings from Phase 6 and previous phases should be made with due consideration of this limitation.

5.3 Final Remarks

Results from PISA, as well as those from other international large-scale assessments, have helped to inform educational research, policy, and practices for over three decades. Over the years, many refinements and improvements have been made to enhance the accuracy and validity of the results, while striving to meet the needs of the current educational reforms and practices (Kirsch & Braun, 2020). For example, advanced statistical modeling techniques have been developed in conjunction with exponential increases in computing power. These advancements have enabled researchers and practitioners to improve the psychometric quality of

assessments and to strengthen the statistical inferences based on the data generated by these assessments.

Nonetheless, less advanced techniques that do not properly account for certain characteristics of ILSA data are still commonly used. Although they are relatively convenient and better known, various studies demonstrated that employing conventional modeling techniques can impair the statistical estimates and lead to misleading or puzzling findings. In particular, measurement error and clustering are commonly overlooked when analyzing data from ILSAs. For instance, Schmidt and Burroughs (2015) illustrated, by employing data from PISA 2012, that spending more time on applied math was negatively correlated with math performance at the between-country level, even though the correlations were positive at the within-country level. In calculating within-country correlations, they employed multilevel linear modeling, which accounts for clustering – as is appropriate for PISA data. On the other hand, multilevel linear modeling does not take measurement error into account, leading to misleading results. Indeed, their findings from within-country analyses suggested a complicated picture, indicating that exposure to applied math was associated with higher math performance only up to a certain point and this relationship was reversed for rather higher levels of exposure to applied math. As others pointed out (Hansen & Strietholt, 2018; Wihardini, 2016), PISA's opportunity to learn measure exhibits problems with its validity and can lead to biased estimates when measurement error associated is not accounted for.

In light of similar examples from the literature, the focus of this study was the unresolved issues regarding the heterogeneity in relationships between certain background variables and cognitive measures seen across countries participating in ILSAs. To this end, a comprehensive investigation was carried out to examine the operating characteristics of competing modeling

techniques. As an extended, illustrative example, this study explored the impact of the clustered nature of the data and the heterogeneous measurement error on the correlations between math self-efficacy and math proficiency by employing data from PISA 2012. The findings showed that advanced modeling techniques not only were more appropriate given the characteristics of the data, but also provided greater insight about the patterns of relationships across countries.

References

- Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2-3), 162-172. doi:10.1016/j.stueduc.2005.05.008
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response theories: An approach to errors-in-variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47-76. Retrieved from <https://www.jstor.org/stable/1165238>
- Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1-43. doi:10.2307/2981882
- Akyol, S. P., Krishna, K., & Wang, J. (2018). *Taking PISA seriously: How accurate are low stakes exams? NBER Working Paper No. 24930*. Cambridge, MA: National Bureau of Economic Research. Retrieved from <https://www.nber.org/papers/w24930.pdf>
- Allison, P. D. (2009). *Quantitative applications in the social sciences: Fixed effects regression models*. Thousand Oaks, CA: SAGE Publications, Inc. doi:10.4135/9781412993869
- Anderson, T. W. (1984). Estimating linear statistical relationships. *The Annals of Statistics, 12*(1), 1-45. Retrieved from <https://www.jstor.org/stable/2241032>
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi:10.1007/BF02293814
- Bandura, A. (1977). *Social learning theory*. New York, NY: General Learning Press.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Macmillan.

- Bartlett, M. S. (1951). The effect of standardization on a Chi-square approximation in factor analysis. *Biometrika*, 38(3/4), 337-344. doi:10.2307/2332580
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01
- Battauz, M., Bellio, R., & Gori, E. (2011). Covariate measurement error adjustment for multilevel models with application to educational data. *Journal of Educational and Behavioral Statistics*, 36(3), 283-306. doi:10.3102/1076998610366262
- Baumgartner, H., & Steenkamp, J. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38, 143–156.
doi:10.1509/jmkr.38.2.143.18840
- Benjamini, Y., Heller, R., & Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906), 4255–4271. doi:10.1098/rsta.2009.0127
- Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, 45(250), 164-180. Retrieved from <https://www.jstor.org/stable/2280676>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika*, 37(1), 29–51. doi:10.1007/BF02291411

- Bock, R. D., & Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.
doi:10.1007/BF02293801
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.
doi:10.1177/014662168200600405
- Bolt, D., & Newton, J. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814–833. doi:10.1177/0013164410388411
- Braun, H. I., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-scale Assessments in Education*, *5*(17), 1-16. doi:10.1186/s40536-017-0050-x
- Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514. Retrieved from <https://pdfs.semanticscholar.org/d137/7d77df410cc2cce07f968b8b08701f18f572.pdf>
- Buckley, J. (2009, June 2). *Cross-national response styles in international educational assessments: Evidence from PISA 2006*. Retrieved March 4, 2019, from RTI International: https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf
- Buonaccorsi, J. P. (1996). Measurement error in the response in the General Linear Model. *Journal of the American Statistical Association*, *91*(434), 633-642. Retrieved from <https://www.jstor.org/stable/2291659>

- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Boca Raton, FL: Taylor & Francis Group.
- Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement, 76*(1), 114-140. doi:10.1177/0013164415584576
- Cho, S. (2007). *A Multilevel Mixture IRT Model for DIF analysis*. Unpublished doctoral dissertation. Athens, GA: University of Georgia.
- Cho, S., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35*(3), 336–370. doi:10.3102/1076998609353111
- Cohen, J. (1986). An uncertainty principle in demography and the unisex issue. *The American Statistician, 40*(1), 32-39. doi:10.1080/00031305.1986.10475351
- Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude x treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology, 67*(6), 717-724. doi:10.1037/0022-0663.67.6.717
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.
- De Boeck, P., Cho, S., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement, 35*(8), 583–603. doi:10.1177/0146621611428446

- De Jong, M. G., & Steenkamp, J. E. (2010). Finite mixture multilevel multidimensional ordinal IRT models. *Psychometrika*, 75(1), 3-32. doi:10.1007/S11336-009-9134-Z
- De Jong, M. G., Steenkamp, J. E., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, 34(2), 260-278. doi:10.1086/518532
- DeShon, R. P. (2003). A generalizability theory perspective on measurement error corrections in validity generalization. In K. R. Murphy, *Validity generalization: A critical review* (pp. 365-402). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Diez-Roux, A. V. (1998). Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health*, 88(2), 216-222.
doi:10.2105/ajph.88.2.216
- Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology*, 44, 247–296. doi:10.1016/B978-0-12-385522-0.00005-6
- Edelen, M., & Reeve, B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(1st Suppl), 5-18.
doi:10.1007/s11136-007-9198-0
- Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, 105(1), 98–121.
doi:10.1016/j.obhdp.2007.05.002

- Emberson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Finch, W. H., & Finch, M. E. (2013). Investigation of specific learning disability and testing accommodations based Differential Item Functioning using a Multilevel Multidimensional Mixture Item Response Theory model. *Educational and Psychological Measurement, 73*(6), 973–993. doi:10.1177/0013164413494776
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character.*, 222, 309–368. doi:10.1098/rsta.1922.0009
- Fox, J. P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software, 20*(5), 1-16. doi:10.18637/jss.v020.i05
- Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288. doi:10.1007/BF02294839
- Fox, J. P., & Glas, C. A. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*(2), 169-191. doi:10.1007/BF02294796
- Fuller, W. A. (1987). *Measurement error models*. Ames, IA: John Wiley & Sons.
- Glas, C. A., & Pimentel, J. L. (2008). Modeling non-ignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*(6), 907-922. doi:10.1177/0013164408315262

- Gnaldi, M., Bacci, S., & Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification, 10*, 53–70. doi:10.1007/s11634-014-0196-0
- Goldstein, H. (1979). Some models for analyzing longitudinal data on educational attainment. *Journal of the Royal Statistical Society, 142*(4), 407-442. Retrieved from <https://www.jstor.org/stable/2982551>
- Goldstein, H., & Spiegelhalter, D. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, 159*(3), 385-443. Retrieved from <https://www.jstor.org/stable/2983325>
- Goltz, H. H., & Smith, M. L. (2010). Yule-Simpson's paradox in research. *Practical Assessment, Research, and Evaluation, 15*(15), 1-9. Retrieved from <https://pareonline.net/getvn.asp?v=15%26n=15>
- Gonzalez, E., & Rutkowski, L. (2009). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In M. von Davier, & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 2, pp. 125-156). Princeton, NJ: IEA-ETS Research Institute. Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02.pdf
- Gortazar, L., Kutner, D., & Inoue, K. (2014). *Education quality and opportunities for skills development in Albania : An analysis of PISA 2000-2012 results (English)*. Washington, DC: World Bank Group. Retrieved from <http://documents.worldbank.org/curated/en/729461468336309841/Education-quality->

and-opportunities-for-skills-development-in-Albania-an-analysis-of-PISA-2000-2012-
results

Goudie, R. J., Turner, R. M., De Angelis, D., & Thomas, A. (2017). MultiBUGS: A parallel implementation of the BUGS modeling framework for faster Bayesian inference. *arXiv Preprint No. 1704.03216*. Retrieved from <https://arxiv.org/abs/1704.03216>

Gustafsson, J. (2018). International large scale assessments: Current status and ways forward. *Scandinavian Journal of Educational Research*, 62(3), 328-332.
doi:10.1080/00313831.2018.1443573

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education*, 20, 261-273. doi:10.2307/749515

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (7th Ed. ed.). Upper Saddle River, NJ: Prentice Hall.

Han, S., Liou-Mark, J., Yu, K. T., & Zeng, S. (2015). Self-efficacy and attitudes towards mathematics of undergraduates: A U. S. and Taiwan comparison. *Journal of Mathematics Education*, 8(1), 1-15. Retrieved from http://educationforatoz.com/images/Han_Spring_2015_.pdf

Hansen, K. Y., & Strietholt, R. (2018). Does schooling actually perpetuate educational inequality in mathematics performance? A validity question on the measures of opportunity to learn in PISA. *ZDM Mathematics Education*, 50, 643-658. doi:10.1007/s11858-018-0935-3

- He, J., & van de Vijver, F. J. (2015). Effects of a general response style on cross-cultural comparisons: Evidence from the teaching and learning international survey. *Public Opinion Quarterly*, 79(Special Issue), 267–290. doi:10.1093/poq/nfv006
- Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What’s wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, 82(6), 903-918. doi:10.1037//0022-3514.82.6.903
- Heyneman, S., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. In L. Rutkowski, M. von Davier, & D. Rutkowski, *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 37-72). Boca Raton, FL: CRC Press Taylor & Francis Group.
- Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing*, 11, 95–121. doi:10.1080/15305058.2010.529977
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications, Inc.
- Jang, Y., Kim, S., & Cohen, A. S. (2018). The impact of multidimensionality on extraction of latent classes in mixture Rasch models. *Journal of Educational Measurement*, 55(3), 403–420. doi:10.1111/jedm.12185
- Jeon, M. (2019). A specialized confirmatory mixture IRT modeling approach for multidimensional tests. *Psychological Test and Assessment Modeling*, 61(1), 91-123.
Retrieved from

<https://proxy.bc.edu/login?url=https%3A%2F%2Fsearch.proquest.com%2Fdocview%2F2203049729%3Faccou>

- Jeon, M., & Rabe-Hesketh, S. (2012). Profile-likelihood approach for estimating generalized linear mixed models with factor structures. *Journal of Educational and Behavioral Statistics, 37*(4), 518–542. doi:10.3102/1076998611417628
- Jones, T. A. (1979). Fitting straight lines when both variables are subject to error. *Mathematical Geology, 11*(1), 1-25. doi:10.1007/BF01043243
- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger, & O. D. Duncan, *Structural equation models in the social sciences* (pp. 85-112). New York, NY: Seminar.
- Ju, U., & Falk, C. (2019). Modeling response styles in cross-country self-reports: An application of a multilevel multidimensional nominal response model. *Journal of Educational Measurement, 56*(1), 169-191. doi:10.1111/jedm.12205
- Jude, N., & Kuger, S. (2018). *Questionnaire development and design for international large-scale assessments: Current practice, challenges, and recommendations*. National Academy of education. Retrieved from http://naeducation.org/wp-content/uploads/2018/02/2018-Questionnaire-Design-for-ILSA_v02-1.pdf
- Junker, B. W., Patz, R. J., & VanHoudnos, N. M. (2016). Markov Chain Monte Carlo for Item Response Models. In W. J. van der Linden, *Handbook of Item Response Theory: Volume two: Statistical tools* (pp. 271-312). Boca Raton, FL: Taylor & Francis Group.

- Kamata, A. (1998). *One-parameter hierarchical generalized linear logistic model: An application of HGLM to IRT*. Unpublished doctoral dissertation. East Lansing, MI: Michigan State University.
- Kamata, A. (2001). Item analysis by hierarchical linear model. *Journal of Educational Measurement, 38*, 79-93. Retrieved from <http://www.jstor.org/stable/1435439>
- Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success (No. w20749)*. Retrieved July 31, 2019, from <http://www.nber.org/papers/w20749>
- Keesling, J. W. (1972). *Maximum likelihood approaches to causal flow analysis*. Unpublished doctoral dissertation. Chicago, IL: University of Chicago.
- Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research, 49*(2), 161-177. doi:10.1080/00273171.2013.866536
- King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191-207. doi:10.1017/S000305540400108X
- Kirsch, I., & Braun, H. (2020). Changing times, changing needs: Enhancing the utility of international large-scale assessments. *Large-scale Assessments in Education, 8*(10), 1-24. doi:10.1186/s40536-020-00088-9

Kish, L. (1965). Cluster sampling and subsampling. In L. Kish, *Survey Sampling* (pp. 148-181).
New York, NY: John Wiley and Sons, Inc.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2), 180-188. doi:10.1037//0022-3514.82.2.180

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134. doi:10.1037/0022-3514.77.6.1121

Kyllonen, P., & Bertling, J. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski, *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 277-285). Boca Raton, FL: Taylor & Francis Group, LLC.

Kyllonen, P., Burrus, J., Roberts, R., & Van de gaer, E. (2010, February). *Cross-cultural comparative questionnaire issues: Paper for the PISA 2012 questionnaire expert group meeting*. Retrieved March 20, 2019, from Australian Council for Educational Research (ACER):
https://www.acer.org/files/kyllonen_burrus_lietz_roberts_vandegaer_pisaqeg2010.pdf

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

- Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1-18. doi:10.18637/jss.v025.i01
- Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences*, 19(3), 355–365. doi:10.1016/j.lindif.2008.10.009
- Lee, J., & Stankov, L. (2013). Higher-order structure of noncognitive constructs and prediction of PISA 2003 mathematics achievement. *Learning and Individual Differences*, 26, 119–130. doi:10.1016/j.lindif.2013.05.004
- Levin, H. M. (2012). More than just test scores. *Prospects*, 42(3), 269–284. doi:10.1007/s11125-012-9240-z
- Lewis-Beck, M. S., Bryman, A., & Liao, T. F. (2004). *The SAGE Encyclopedia of Social Science Research Methods*. Thousand Oaks, CA: Sage Publications, Inc.
doi:10.4135/9781412950589
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons, Inc.
- Lockheed, M. E., & Wagemaker, H. (2013). International large-scale assessments: Thermometers, whips or useful policy tools? *Research in Comparative and International Education*, 8(3), 296-306. doi:10.2304/rcie.2013.8.3.296
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

- Lu, Y., & Bolt, D. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-scale Assessments in Education*, 3(2), 1-18. doi:10.1186/s40536-015-0012-0
- Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, 59(4), 615-630. doi:10.1177/0013164499594004
- Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS — a Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337. doi:10.1023/A:1008929526011
- Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education*, 30(5), 520-540. doi:10.2307/749772
- Mai, Y., & Zhang, Z. (2018). Software packages for bayesian multilevel modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 25, 650–658. doi:10.1080/10705511.2018.1431545
- Maris, E. (1999). Estimating multiple classification Latent Class Models. *Psychometrika*, 64(2), 187-212. doi:10.1007/BF02294535
- Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, 47(1), 213-231. doi:10.1037/0022-3514.47.1.213

- Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*(2), 331-353. doi:10.1037/edu0000281
- Marsh, W. H., Roche, L. A., Pajares, F., & Miller, D. (1997). Item-specific efficacy judgments in mathematical problem solving: The downside of standing too close to trees in a forest. *Contemporary Educational Psychology, 22*(3), 363-377. doi:10.1006/ceps.1997.0942
- Martin, M. O., Mullis, I., & Hooper, M. (2016). *TIMSS: Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and International Association for the Evaluation of Educational Achievement (IEA). Retrieved from <https://timssandpirls.bc.edu/publications/timss/2015-methods.html>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. doi:10.1007/BF02296272
- Matsueda, R. L. (2011). *Key advances in the history of structural equation modeling (Working Paper no. 114)*. Center for Statistics and the Social Sciences, University of Washington. Retrieved from <https://www.csss.washington.edu/files/working-papers/2012/wp114.pdf>
- Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of maths anxiety and its influence on young adolescents' course enrolment and performance in mathematics. *Journal of Educational Psychology, 82*(1), 60-70. doi:10.1037/0022-0663.82.1.60

- Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel laureates. *The New England Journal of Medicine*, 367(16), 1562-1564. doi:10.1056/NEJMon1211064
- Min, I., Cortina, K., & Miller, K. (2016). Modesty bias and the attitude-achievement paradox across nations: A reanalysis of TIMSS. *Journal of Psychology and Education*, 51, 359–366. doi:10.1016/j.lindif.2016.09.008
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49(3), 359-381. doi:10.1007/BF02306026
- Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, 11(1), 81-91. doi:10.1177/014662168701100106
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196. doi:10.1007/BF02294457
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55(2), 195-215. doi:10.1007/BF02295283
- Morony, S., Kleitman, S., Lee, Y. P., & Stankov, L. (2013). Predicting achievement: Confidence vs self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research*, 58, 79–96. doi:10.1016/j.ijer.2012.11.002
- Mun, E., Huo, Y., White, H. R., Suzuki, S., & de la Torre, J. (2019). Multivariate higher-order IRT model and MCMC algorithm for linking individual participant data from multiple studies. *Frontiers in Psychology*, 10(1328), 1-13. doi:10.3389/fpsyg.2019.01328

- Muraki, E. (1997). A Generalized Partial Credit Model. In W. J. van der Linden, & R. K. Hambleton, *Handbook of Modern Item Response Theory* (pp. 153-164). New York, NY: Springer. doi:10.1007/978-1-4757-2691-6_9
- Murphy, K. R. (2003). *Validity generalization: A critical review*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22(3), 376–398. doi:10.1177/0049124194022003006
- Muthén, B., & Asparouhov, T. (2016). Multidimensional, Multilevel, and Multi-timepoint Item Response Modeling. In W. J. van der Linden, *Handbook of Item Response Theory: Volume one: Models* (pp. 527-540). Boca Raton, FL: Taylor & Francis Group.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide* (Seventh Edition ed.). Los Angeles, CA: Muthén and Muthén.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14(4), 535–569. doi:10.1080/10705510701575396
- OECD. (2005). *PISA 2003 Technical Report*. Paris: OECD Publishing. Retrieved from <http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/35188570.pdf>
- OECD. (2009). *PISA Data Analysis Manual*. Paris: OECD Publishing. Retrieved from <https://www.oecd-ilibrary.org/docserver/9789264056275->

en.pdf?expires=1575602148&id=id&accname=guest&checksum=A43E934677D4ADEC
763DD47429FABDA7

OECD. (2013a). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing.
doi:10.1787/9789264190511-en

OECD. (2013b). *PISA 2012 results: Ready to learn : Students' engagement, drive and self-beliefs (Volume III)*. Paris: OECD Publishing. doi:10.1787/9789264201170-en

OECD. (2014). *PISA 2012 Technical Report*. Paris: OECD. Retrieved from
<https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>

OECD. (2017). *PISA 2015 Assessment and analytical framework: Science, reading, mathematics, financial literacy and collaborative problem solving*. Paris: OECD Publishing. doi:10.1787/9789264281820-en

OECD. (2019a). *PISA 2018 Assessment and analytical framework*. Paris: OECD Publishing.
doi:10.1787/b25efab8-en

OECD. (2019b). *PISA 2018 results (Volume I): What students know and can do*. Paris: OECD Publishing. doi:10.1787/5f07c754-en

OECD. (2019c). *PISA 2021 Context Questionnaire Framework (Field Trial Version)*. Retrieved from <https://www.oecd.org/pisa/publications/pisa-2021-assessment-and-analytical-framework.htm>

OECD. (2020). *PISA products - PISA*. Retrieved from PISA:
<https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm>

- Pajares, F., & Miller, D. M. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem-solving: A path analysis. *Journal of Educational Psychology*, 86(2), 193-203. doi:10.1037/0022-0663.86.2.193
- Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education*, 16(3), 223-243. doi:10.1207/S15324818AME1603_4
- Pepper, D. (2020). When assessment validation neglects any strand of validity evidence: An instructive example from PISA. *Educational Measurement: Issues and Practice*, 39(4), 8-20. doi:10.1111/emip.12380
- Porter, A., & Gamoran, A. (2002). Progress and challenges for large-scale studies. In A. Porter, & A. Gamoran, *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 3-26). Washington, DC: National Academy Press.
- R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized Multilevel Structural Equation Modeling. *Psychometrika*, 69(2), 167-190. doi:<https://doi.org/10.1007/BF02295939>
- Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2012). Multilevel structural equation modeling. In R. H. Hoyle, *Handbook of Structural Equation Modeling* (pp. 512-531). New York, NY: The Guilford Press.

- Rasch, G. (1960). On general laws and the meaning of measurement in psychology. *The Fourth Berkeley Symposium on Mathematical Statistics and Probability. June 20-July 30, 1960.* Statistical Laboratory, University of California.
- Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. K. Edwards, *Statistics: Textbooks and monographs* (Vol. 137, pp. 459-496). New York, NY: Marcel Dekker.
- Ravand, H. (2015). Item response theory using hierarchical generalized linear models. *Practical Assessment, Research, and Evaluation, 20*(7), 1-17. Retrieved from <https://pareonline.net/getvn.asp?v=20&n=7>
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36. doi:10.1177/0146621697211002
- Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.
- Reise, S., Widaman, K., & Pugh, R. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566. doi:10.1037/0033-2909.114.3.552
- Rencher, A. C. (2002). *Methods of Multivariate Analysis*. New York, NY: John Wiley & Sons, Inc.
- Revelle, W. (2019). psych: Procedures for personality and psychological research. *R package version 1.9.12*. Evanston, IL. Retrieved from <https://CRAN.R-project.org/package=psych>

- Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. Retrieved from <http://www.jstor.org/stable/2087176>
- Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, 82(3), 795–819. doi:10.1007/s11336-016-9544-7
- Rost, J. (1990). Rasch Models in Latent Classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271-282. doi:10.1177/014662169001400305
- Rost, J., & von Davier, M. (1995a). Polytomous Mixed Rasch Models. In G. H. Fischer, & I. W. Molenaar, *Rasch models: Foundations, recent developments, and applications* (pp. 371-379). New York, NY: Springer. doi:10.1007/978-1-4612-4230-7_20
- Rost, J., & von Davier, M. (1995b). Mixture distribution Rasch models. In G. H. Fischer, & I. W. Molenaar, *Rasch models: Foundations, recent developments, and applications* (pp. 257-268). Springer: New York, NY. doi:10.1007/978-1-4612-4230-7_20
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.
- Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, 4, 395–421. doi:10.1146/annurev-statistics-060116-054045
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford Press.

- Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski, *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 118-153). Boca Raton, FL: Taylor & Francis Group.
- Rutkowski, D., Rutkowski, L., & von Davier, M. (2014). A brief introduction to modern international large-scale assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski, *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 3-9). Boca Raton, FL: Taylor & Francis Group.
- Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': the importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, 42(3), 411-430. doi:10.1080/00220272.2010.487546
- Samejima, F. (1997). Graded Response Model. In W. J. van der Linden, & R. K. Hambleton, *Handbook of Modern Item Response Theory* (pp. 85-100). New York, NY: Springer. doi:10.1007/978-1-4757-2691-6_5
- Schmidt, W. H., & Burroughs, N. A. (2015). Puzzling out PISA: What can international comparisons tell us about American education? *American Educator*, 39(1), 24-31. Retrieved from <https://files.eric.ed.gov/fulltext/EJ1063875.pdf>
- Schofield, L. S. (2015). Correcting for measurement error in latent variables used as predictors. *The Annals of Applied Statistics*, 9(4), 2133-2152. doi:10.1214/15-aos877

- Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables with covariates. *Psychometrika*, *80*(3), 727–747. doi:10.1007/s11336-014-9415-z
- Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to Structural Equation Modeling*. New York, NY: Routledge.
- Schunk, D. H. (1989). Self-efficacy and academic behaviors. *Educational Psychology Review*, *1*(3), 173-208. Retrieved from <https://www.jstor.org/stable/23359217>
- Schwartz, S. (1994). The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. *American Journal of Public Health*, *84*(5), 819-824. doi:10.2105/ajph.84.5.819
- Sen, S., & Cohen, A. S. (2019). Applications of Mixture IRT Models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177-191. doi:10.1080/15366367.2019.1583506
- Sen, S., & Terzi, R. (2019). A comparison of software packages available for DINA model estimation. *Applied Psychological Measurement*, 1–15. doi:10.1177/0146621619843822
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, *13*(2), 238-241. Retrieved from www.jstor.org/stable/2984065
- Singer, J., & Braun, H. I. (2018, April 6). Testing international education assessments. *Science*, *360*(6384), pp. 38-40. doi:10.1126/science.aar4952

- Snijders, T. A. (2005). Power and sample size in multilevel modeling. (B. S. Everitt, & D. C. Howell, Eds.) *Encyclopedia of Statistics in Behavioral Science*, 3, 1570–1573. Retrieved from <https://pdfs.semanticscholar.org/a769/1eb67c5806e154da58a74f7b1a1bc9ccb58a.pdf>
- Stankov, L., Lee, J., & von Davier, M. (2018). A note on construct validity of the anchoring method in PISA 2012. *Journal of Psychoeducational Assessment*, 36(7), 709–724. doi:10.1177/0734282917702270
- StataCorp. (2019). *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.
- Sulis, I., & Toland, M. (2017). Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *Journal of Early Adolescence*, 37(1), 85–128. doi:10.1177/0272431616642328
- Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York, NY: Guilford.
- Tijmstra, J., Bolsinova, M., Liaw, Y., Rutkowski, L., & Rutkowski, D. (2020). Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *Journal of Educational Measurement*, 57(4), 566-583. doi:10.1111/jedm.12263
- Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, 1-30. doi:10.1111/bmsp.12188

- Van de gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology, 43*, 1205-1228. doi:10.1177/0022022111428083
- van de Vijver, F. J., & He, J. (2016). Bias assessment and prevention in non-cognitive outcome measures in context assessments. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan, *Assessing Context of Learning: An international perspective* (pp. 229-253). Cham: Springer.
- van der Linden, W. J. (2016). *Handbook of Item Response Theory: Volume one: Models*. Boca Raton, FL: Taylor & Francis Group.
- Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 36*(1), 213-239. doi:10.1111/j.0081-1750.2003.t01-1-00131.x
- Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica, 58*(2), 220–233. doi:10.1046/j.0039-0402.2003.00257.x
- Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research, 17*(1), 33–51. doi:10.1177/0962280207081238
- Vermunt, J. K., & Magidson, J. (2005). *Technical guide for latent GOLD 4.0: Basic and advanced [Computer software and manual]*. Belmont, CA: Statistical Innovations, Inc.
- Vigen, T. (2019, May 9). *tylervigen.com*. Retrieved from Spurious correlations: <https://tylervigen.com/spurious-correlations>

- von Davier, M. (2005a). *A General Diagnostic Model applied to language testing data*. Princeton, NJ: Educational Testing Service. Retrieved from <https://files.eric.ed.gov/fulltext/EJ1111422.pdf>
- von Davier, M. (2005b). mdlm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: ETS.
- von Davier, M. (2007). *Mixture Distribution Diagnostic Models*. Princeton, NJ: Educational Testing Service.
- von Davier, M. (2009). Mixture distribution item response theory, latent class analysis, and diagnostic mixture models. In S. Embertson, *Measuring psychological constructs: Advances in model-based approaches* (pp. 11-34). Washington, DC: American Psychological Association.
- von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling*, 52(1), 8-28. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf
- von Davier, M. (2014). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski, *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175-203). Boca Raton, FL: CRC Press.
- von Davier, M. (2018). Detecting and treating errors in tests and surveys. *Quality Assurance in Education*, 26(2), 243-262. doi:10.1108/QAE-07-2017-0036

- von Davier, M., & Rost, J. (2007). Mixture Distribution Item Response Models. In C. R. Rao, & S. Sinharay, *Handbook of Statistics: Psychometrics* (pp. 643-663). Amsterdam, The Netherlands: Elsevier.
- von Davier, M., & Rost, J. (2016). Logistic Mixture-Distribution Response Models. In W. J. van der Linden, *Handbook of Item Response Theory* (pp. 393-407). Boca Raton, FL: Taylor & Francis Group, LLC.
- von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski, *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 155-174). Boca Raton, FL: Taylor & Francis Group.
- von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the Generalized Partial-Credit Model. *Applied Psychological Measurement*, 28(6), 389–406. doi:10.1177/0146621604268734
- von Davier, M., DiBello, L., & Yamamoto, K. (2006). *Reporting test outcomes using models for cognitive diagnosis*. Princeton, NJ: Educational Testing Service. Retrieved from <https://files.eric.ed.gov/fulltext/EJ1111387.pdf>
- von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). *Plausible values: What are they and why do we need them?* Retrieved May 1, 2019, from Issues and Methodologies in Large-Scale Assessments: IERI Monograph Series: http://www.ierinstitute.org/IERI_Monograph_Volume_02.pdf

- von Davier, M., Shin, H.-J., Khorrarnadel, L., & Stankov, L. (2018). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement, 42*(4), 291-306. doi:10.1177/0146621617730389
- von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorrarnadel, K., Weeks, J., . . . Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice, 26*(4), 466-488. doi:10.1080/0969594X.2019.1586642
- Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health, 29*, 75-90. doi:10.1146/annurev.publhealth.29.020907.090821
- Wihardini, D. (2016). *An investigation of the relationship of student performance to their opportunity-to-learn in PISA 2012 mathematics: The case of Indonesia*. Berkeley, CA: University of California.
- Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger, & O. D. Duncan, *Structural equation models in the social sciences* (pp. 69-83). New York, NY: Seminar.
- Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society, 159*(2), 201-212. Retrieved from <http://links.jstor.org/sici?sici=0964-1998%281996%29159%3A2%3C201%3AAFMEIM%3E2.0.CO%3B2-V>
- Yap, S. C., Donnellan, M. B., Schwartz, S. J., Kim, S. Y., Castillo, L. G., Zamboanga, B. L., . . . Vazsonyi, A. T. (2014). Investigating the structure and measurement invariance of the

multigroup ethnic identity measure in a multiethnic sample of college students. *Journal of Counseling Psychology*, *61*(3), 437–446. doi:10.1037/a0036253

Zhang, J., Lu, J., Chen, F., & Tao, J. (2019). Exploring the correlation between multiple latent variables and covariates in hierarchical data based on the Multilevel Multidimensional IRT Model. *Frontiers in Psychology*, *10*, 1-17. doi:10.3389/fpsyg.2019.02387

Appendix

Table A.1. Country names and abbreviations

Country abbr.	Country name	Country abbr.	Country name
<i>ARE</i>	United Arab Emirates	<i>KAZ</i>	Kazakhstan
<i>ARG</i>	Argentina	<i>KOR</i>	Korea
<i>AUS</i>	Australia	<i>LTU</i>	Lithuania
<i>AUT</i>	Austria	<i>LUX</i>	Luxembourg
<i>BEL</i>	Belgium	<i>LVA</i>	Latvia
<i>BGR</i>	Bulgaria	<i>MAC</i>	Macao-China
<i>BRA</i>	Brazil	<i>MEX</i>	Mexico
<i>CAN</i>	Canada	<i>MNE</i>	Montenegro
<i>CHE</i>	Switzerland	<i>MYS</i>	Malaysia
<i>CHL</i>	Chile	<i>NLD</i>	Netherlands
<i>COL</i>	Colombia	<i>NOR</i>	Norway
<i>CRI</i>	Costa Rica	<i>NZL</i>	New Zealand
<i>CZE</i>	Czech Republic	<i>PER</i>	Peru
<i>DEU</i>	Germany	<i>POL</i>	Poland
<i>DNK</i>	Denmark	<i>PRT</i>	Portugal
<i>ESP</i>	Spain	<i>QAT</i>	Qatar
<i>EST</i>	Estonia	<i>ROU</i>	Romania
<i>FIN</i>	Finland	<i>RUS</i>	Russian Federation
<i>FRA</i>	France	<i>SGP</i>	Singapore
<i>GBR</i>	United Kingdom	<i>SRB</i>	Serbia
<i>GRC</i>	Greece	<i>SVK</i>	Slovak Republic
<i>HKG</i>	Hong Kong-China	<i>SVN</i>	Slovenia
<i>HRV</i>	Croatia	<i>SWE</i>	Sweden
<i>HUN</i>	Hungary	<i>TAP</i>	Chinese Taipei
<i>IDN</i>	Indonesia	<i>THA</i>	Thailand
<i>IRL</i>	Ireland	<i>TUN</i>	Tunisia
<i>ISL</i>	Iceland	<i>TUR</i>	Turkey
<i>ISR</i>	Israel	<i>URY</i>	Uruguay
<i>ITA</i>	Italy	<i>USA</i>	United States of America
<i>JOR</i>	Jordan	<i>VNM</i>	Viet Nam
<i>JPN</i>	Japan		

Table A.2. Total and final sample sizes by country

Country	Total Sample		Final Sample	
	Number of Schools	Number of Students	Number of Schools	Number of Students
ARE	458	11500	457	7503
ARG	226	5908	226	3726
AUS	775	14481	772	9450
AUT	191	4755	185	3115
BEL	287	8597	273	5475
BGR	188	5282	188	3401
BRA	839	19204	831	12403
CAN	885	21544	880	14090
CHE	411	11229	409	7395
CHL	221	6856	220	4523
COL	352	9073	349	5839
CRI	193	4602	192	2863
CZE	297	5327	281	3465
DEU	230	5001	217	2805
DNK	341	7481	337	4841
ESP	902	25313	896	16656
EST	206	4779	205	3149
FIN	311	8829	301	5758
FRA	226	4613	226	2988
GBR	507	12659	507	8349
GRC	188	5125	187	3359
HKG	148	4670	148	3067
HRV	163	5008	163	3314
HUN	204	4810	202	3175
IDN	209	5622	209	3700
IRL	183	5016	183	3304
ISL	134	3508	134	2228
ISR	172	5055	170	3204
ITA	1194	31073	1191	20568
JOR	233	7038	233	4555
JPN	191	6351	190	4166
KAZ	218	5808	217	3854
KOR	156	5033	156	3351
LTU	216	4618	216	3027
LUX	42	5258	42	3427
LVA	211	4306	209	2819
MAC	45	5335	45	3542
MEX	1471	33806	1465	22301

Country	Total Sample		Final Sample	
	Number of Schools	Number of Students	Number of Schools	Number of Students
MNE	51	4744	50	3015
MYS	164	5197	164	3413
NLD	179	4460	174	2834
NOR	197	4686	196	3053
NZL	177	4291	177	2790
PER	240	6035	239	3897
POL	184	4607	182	3057
PRT	195	5722	195	3716
QAT	157	10966	154	6764
ROU	178	5074	178	3365
RUS	227	5231	227	3454
SGP	172	5546	172	3687
SRB	153	4684	152	3007
SVK	231	4678	224	3021
SVN	338	5911	302	3755
SWE	209	4736	205	3095
TAP	163	6046	163	4004
THA	239	6606	239	4392
TUN	153	4407	153	2820
TUR	170	4848	170	3188
URY	180	5315	180	3442
USA	162	4978	162	3258
VNM	162	4959	162	3269
TOTAL	17705	468200	17532	305051

Figure A.1. PISA 2012 Student Questionnaire Rotated Design

Form A		Form B		Form C	
Q. n°.	Description	Q. n°.	Description	Q. n°.	Description
ST01-28	Common Part (see Table 6.3)	ST01-28	Common Part (see Table 6.3)	ST01-28	Common Part (see Table 6.3)
ST29	Intrinsic and Instrumental Motivation for Mathematics	ST42	Mathematics Self-Concept (Q2, 4, 6, 7, 9); Mathematics Anxiety (Q1, 3, 5, 8, 10)	ST53	Learning Strategies (Control vs. Elaboration vs. Memorisation)
ST35	Subjective Norms	ST77	Teacher Support In Mathematics Class	ST55	Attendance of Out-of-School-Time Lessons
ST37	Mathematics Self-Efficacy	ST79	Teacher Behaviour: - Teacher Directed Instruction - Formative Assessment - Student Orientation	ST57	Total Hours of Out-of-School Study Time
ST43	Perceived Control of Mathematics Performance	ST80	Cognitive Activation in Mathematics Lessons	ST61	Experience with Applied Mathematics Tasks (Q1, 4, 6, 8); Experience with Pure Mathematics tasks (Q5, 7, 9)
ST44	Attributions to Failure In Mathematics	ST81	Disciplinary Climate	ST62	Familiarity with Mathematics Concepts
ST46	Mathematics Work Ethic	ST82	+ Anchoring Vignettes	ST69	Min In «Class Period»
ST48	Mathematics Intentions	ST83	Mathematics Teacher Support	ST70	N° of «Class Period» per Week
ST49	Mathematics Behaviour	ST84	+ Anchoring Vignettes	ST71	N° of All «Class Period» a Week
ST93	Perseverance	ST85	Mathematics Teacher's Classroom Management	ST72	Class Size
ST94	Openness for Problem Solving	ST86	Student-Teacher Relations	ST73	Experience with Word Problems
ST96	Problem-Solving Strategies (SJT)	ST87	Sense of Belonging to School	ST74	Experience with Procedural Tasks
ST101	Problem-Solving Strategies (SJT)	ST88	Attitude towards School: Learning Outcomes	ST75	Experience with Pure Mathematics Reasoning
ST104	Problem-Solving Strategies (SJT)	ST89	Attitude towards School: Learning Activities	ST76	Experience with Applied Mathematics Reasoning
ST53	Learning Strategies (Control vs. Elaboration vs. Memorisation)	ST91	Perceived Control of Success In School	ST42	Mathematics Self-Concept (Q2, 4, 6, 7, 9); Mathematics Anxiety (Q1, 3, 5, 8, 10)
ST55	Attendance of Out-of-School-Time Lessons	ST29	Intrinsic and Instrumental Motivation for Mathematics	ST77	Teacher Support In Mathematics Class
ST57	Total Hours of Out-of-School Study Time	ST35	Subjective Norms	ST79	Teacher Behaviour: - Teacher Directed Instruction - Formative Assessment - Student Orientation
ST61	Experience with Applied Mathematics Tasks (Q1, 4, 6, 8); Experience with Pure Mathematics tasks (Q5, 7, 9)	ST37	Mathematics Self-Efficacy	ST80	Cognitive Activation in Mathematics Lessons
ST62	Familiarity with Mathematics Concepts	ST43	Perceived Control of Mathematics Performance	ST81	Disciplinary Climate
ST69	Min In «Class Period»	ST44	Attributions to Failure In Mathematics	ST82	+ Anchoring Vignettes
ST70	N° of «Class Period» per Week	ST46	Mathematics Work Ethic	ST83	Mathematics Teacher Support
ST71	N° of All «Class Period» a Week	ST48	Mathematics Intentions	ST84	+ Anchoring Vignettes
ST72	Class Size	ST49	Mathematics Behaviour	ST85	Mathematics Teachers' Classroom Management
ST73	Experience with Word Problems	ST93	Perseverance	ST86	Student-Teacher Relations
ST74	Experience with Procedural Tasks	ST94	Openness for Problem Solving	ST87	Sense of Belonging to School
ST75	Experience with Pure Mathematics Reasoning	ST96	Problem-Solving Strategies (SJT)	ST88	Attitude towards School: Learning Outcomes
ST76	Experience with Applied Mathematics Reasoning	ST101	Problem-Solving Strategies (SJT)	ST89	Attitude towards School: Learning Activities
		ST104	Problem-Solving Strategies (SJT)	ST91	Perceived Control of Success In School

(OECD, 2013a; pg. 194)

Table A.3. *Q*-Matrix used in MIRT and MLMixMIRT models

Item	MATHEFF Dimension	MATH Dimension	Item	MATHEFF Dimension	MATH Dimension
ST37Q01	1	0	PM571Q01	0	1
ST37Q02	1	0	PM603Q01T	0	1
ST37Q03	1	0	PM800Q01	0	1
ST37Q04	1	0	PM803Q01T	0	1
ST37Q05	1	0	PM828Q01	0	1
ST37Q06	1	0	PM828Q02	0	1
ST37Q07	1	0	PM828Q03	0	1
ST37Q08	1	0	PM903Q01	0	1
PM00FQ01	0	1	PM903Q03	0	1
PM00GQ01	0	1	PM905Q01T	0	1
PM00KQ02	0	1	PM905Q02	0	1
PM033Q01	0	1	PM906Q01	0	1
PM034Q01T	0	1	PM906Q02	0	1
PM155Q01	0	1	PM909Q01	0	1
PM155Q02D	0	1	PM909Q02	0	1
PM155Q03D	0	1	PM909Q03	0	1
PM155Q04T	0	1	PM915Q01	0	1
PM192Q01T	0	1	PM915Q02	0	1
PM273Q01T	0	1	PM918Q01	0	1
PM305Q01	0	1	PM918Q02	0	1
PM406Q01	0	1	PM918Q05	0	1
PM406Q02	0	1	PM919Q01	0	1
PM408Q01T	0	1	PM919Q02	0	1
PM411Q01	0	1	PM923Q01	0	1
PM411Q02	0	1	PM923Q03	0	1
PM420Q01T	0	1	PM923Q04	0	1
PM423Q01	0	1	PM924Q02	0	1
PM442Q02	0	1	PM934Q01	0	1
PM446Q01	0	1	PM934Q02	0	1
PM446Q02	0	1	PM936Q01	0	1
PM447Q01	0	1	PM936Q02	0	1
PM462Q01D	0	1	PM939Q01	0	1
PM464Q01T	0	1	PM939Q02	0	1
PM474Q01	0	1	PM942Q01	0	1
PM496Q01T	0	1	PM942Q02	0	1
PM496Q02	0	1	PM942Q03	0	1
PM559Q01	0	1	PM943Q01	0	1
PM564Q01	0	1	PM943Q02	0	1
PM564Q02	0	1	PM948Q01	0	1

Item	MATHEFF Dimension	MATH Dimension
PM948Q02	0	1
PM948Q03	0	1
PM949Q01T	0	1
PM949Q02T	0	1
PM949Q03	0	1
PM953Q02	0	1
PM953Q03	0	1
PM953Q04D	0	1
PM954Q01	0	1
PM954Q02	0	1
PM954Q04	0	1
PM955Q01	0	1
PM955Q02	0	1
PM955Q03	0	1
PM957Q01	0	1
PM957Q02	0	1
PM957Q03	0	1
PM961Q02	0	1
PM961Q03	0	1
PM961Q05	0	1

Item	MATHEFF Dimension	MATH Dimension
PM967Q01	0	1
PM967Q03T	0	1
PM982Q01	0	1
PM982Q02	0	1
PM982Q03T	0	1
PM982Q04	0	1
PM985Q01	0	1
PM985Q02	0	1
PM985Q03	0	1
PM991Q01	0	1
PM991Q02D	0	1
PM992Q01	0	1
PM992Q02	0	1
PM992Q03	0	1
PM995Q01	0	1
PM995Q02	0	1
PM995Q03	0	1
PM998Q02	0	1
PM998Q04T	0	1

Table A.4. Sample size and mean PVIMATH range by super-school and country (Phase 5)

Country	Super-school ID	PVIMATH		Number of Schools	Number of Students
		Weighted Mean	Range		
KAZ	1	[354,385]		24	532
KAZ	2	(385,396]		15	319
KAZ	3	(396,404]		16	357
KAZ	4	(404,416]		15	321
KAZ	5	(416,429]		16	359
KAZ	6	(429,439]		15	320
KAZ	7	(439,458]		16	357
KAZ	8	(458,482]		15	335
KAZ	9	(482,588]		24	514
JOR	1	[255,347]		32	616
JOR	2	(347,358]		21	410
JOR	3	(358,369]		21	433
JOR	4	(369,377]		21	432
JOR	5	(377,385]		22	441
JOR	6	(385,397]		21	441
JOR	7	(397,406]		21	417
JOR	8	(406,425]		21	449
JOR	9	(425,606]		32	698
GRC	1	[321,414]		22	431
GRC	2	(414,436]		15	311
GRC	3	(436,452]		14	307
GRC	4	(452,462]		15	320
GRC	5	(462,472]		14	320
GRC	6	(472,478]		15	327
GRC	7	(478,490]		14	307
GRC	8	(490,500]		15	319
GRC	9	(500,602]		22	478
NLD	1	[360,448]		21	359
NLD	2	(448,469]		14	241
NLD	3	(469,493]		13	229
NLD	4	(493,511]		14	248
NLD	5	(511,548]		13	232
NLD	6	(548,571]		14	241
NLD	7	(571,592]		14	255
NLD	8	(592,608]		13	237
NLD	9	(608,671]		21	364
PRT	1	[370,416]		23	455
PRT	2	(416,450]		15	324

PV1MATH**Country Super-school ID Weighted Mean Range Number of Schools Number of Students**

Country	Super-school ID	Weighted Mean Range	Number of Schools	Number of Students
PRT	3	(450,470]	15	323
PRT	4	(470,486]	15	324
PRT	5	(486,502]	14	306
PRT	6	(502,512]	15	336
PRT	7	(512,527]	15	341
PRT	8	(527,543]	15	339
PRT	9	(543,609]	23	523
TAP	1	[371,475]	25	565
TAP	2	(475,504]	16	386
TAP	3	(504,529]	16	399
TAP	4	(529,548]	16	408
TAP	5	(548,557]	16	395
TAP	6	(557,583]	16	397
TAP	7	(583,610]	16	398
TAP	8	(610,632]	16	410
TAP	9	(632,742]	24	622

Table A.5. Sample size by country (Phase 6)

Country	Number of Schools	Number of Students	Country	Number of Schools	Number of Students
ARE	448	3681	MYS	164	1683
ARG	224	1808	NLD	174	1422
AUS	768	4668	NOR	194	1505
AUT	181	1554	NZL	176	1381
BEL	271	2675	PER	238	1856
BGR	184	1671	POL	178	1512
BRA	808	6030	PRT	195	1848
CAN	871	6958	QAT	150	3230
CHE	404	3690	ROU	178	1687
CHL	216	2264	RUS	223	1724
COL	345	2790	SGP	172	1831
CRI	192	1446	SRB	152	1488
CZE	280	1700	SVK	224	1491
DEU	215	1406	SVN	294	1850
DNK	337	2395	SWE	201	1508
ESP	892	8276	TAP	163	1994
EST	201	1587	THA	239	2190
FIN	298	2817	TUN	151	1386
FRA	226	1507	TUR	166	1571
GBR	507	4165	URY	177	1661
GRC	183	1667	USA	160	1619
HKG	148	1512	VNM	162	1628
HRV	163	1648	TOTAL	17316	150855
HUN	193	1584			
IDN	209	1823			
IRL	183	1650			
ISL	127	1117			
ISR	169	1580			
ITA	1166	10267			
JOR	233	2247			
JPN	189	2085			
KAZ	215	1912			
KOR	156	1685			
LTU	214	1524			
LUX	42	1683			
LVA	205	1389			
MAC	45	1765			
MEX	1427	11082			
MNE	50	1482			

Table A.6. Correlation estimates from the OLS models employed in Phase 1

Country	Std. Regression Coef. (Corr. Est.)	SE
ARE	0.36	0.01
ARG	0.21	0.02
AUS	0.59	0.01
AUT	0.53	0.02
BEL	0.46	0.01
BGR	0.29	0.02
BRA	0.30	0.01
CAN	0.56	0.01
CHE	0.59	0.01
CHL	0.34	0.01
COL	0.15	0.01
CRI	0.24	0.02
CZE	0.53	0.02
DEU	0.55	0.02
DNK	0.55	0.01
ESP	0.50	0.01
EST	0.53	0.02
FIN	0.52	0.01
FRA	0.54	0.02
GBR	0.59	0.01
GRC	0.47	0.02
HKG	0.55	0.02
HRV	0.55	0.01
HUN	0.61	0.02
IDN	0.17	0.02
IRL	0.55	0.02
ISL	0.51	0.02
ISR	0.47	0.02
ITA	0.50	0.01
JOR	0.28	0.02
JPN	0.58	0.01
KAZ	0.26	0.02
KOR	0.62	0.01
LTU	0.53	0.02
LUX	0.51	0.02
LVA	0.50	0.02
MAC	0.51	0.01
MEX	0.32	0.01
MNE	0.31	0.02
MYS	0.39	0.02

Country	Std. Regression Coef. (Corr. Est.)	SE
NLD	0.47	0.02
NOR	0.59	0.02
NZL	0.58	0.02
PER	0.21	0.02
POL	0.63	0.01
PRT	0.63	0.01
QAT	0.28	0.01
ROU	0.36	0.02
RUS	0.49	0.02
SGP	0.56	0.01
SRB	0.41	0.02
SVK	0.53	0.02
SVN	0.50	0.02
SWE	0.52	0.02
TAP	0.66	0.01
THA	0.22	0.02
TUN	0.31	0.02
TUR	0.46	0.02
URY	0.33	0.02
USA	0.56	0.02
VNM	0.50	0.02

Table A.7. Descriptive statistics for the item response data employed in Phase 2 and Phase 4

Item	Number of categories	Number of examinees who tried/responded to the category					Item-total corr. (MATHEFF)	Item-total corr. (MATH)
		tried	0	1	2	3		
ST37Q01	4	61906.35	2361.13	12113.29	26968.46	20463.46	0.5160	
ST37Q02	4	61890.43	2364.44	10683.78	24828.47	24013.73	0.6156	
ST37Q03	4	61736.89	3432.70	16166.47	23340.86	18796.86	0.6478	
ST37Q04	4	61636.77	2463.95	11747.92	26597.17	20827.73	0.5586	
ST37Q05	4	61586.26	2265.55	7066.31	18462.58	33791.83	0.5714	
ST37Q06	4	61701.77	5470.60	20832.51	20781.90	14616.76	0.5987	
ST37Q07	4	61779.09	3909.81	12294.90	20997.64	24576.73	0.5519	
ST37Q08	4	61797.59	5425.57	21143.00	22583.94	12645.08	0.5328	
PM00FQ01	2	14358.89	8250.10	6108.79				0.5177
PM00GQ01	2	18671.66	17202.09	1469.58				0.3365
PM00KQ02	2	17916.79	15333.43	2583.36				0.3790
PM033Q01	2	18638.07	5595.44	13042.63				0.3623
M034Q01T	2	17976.62	11356.77	6619.84				0.4789
PM155Q01	2	18539.48	7268.28	11271.20				0.4907
M155Q02D	3	18601.77	7450.97	2462.98	8687.82			0.5564
M155Q03D	3	18529.71	14544.17	2197.90	1787.65			0.5230
M155Q04T	2	18508.23	9122.99	9385.24				0.4107
M192Q01T	2	18628.60	11961.74	6666.86				0.5051
M273Q01T	2	18930.05	9796.32	9133.73				0.3866
PM305Q01	2	18696.45	8300.33	10396.12				0.3490
PM406Q01	2	18768.59	14621.93	4146.66				0.5551
PM406Q02	2	18698.54	15992.28	2706.26				0.5349
M408Q01T	2	19081.36	12718.58	6362.78				0.4212
PM411Q01	2	18667.01	10608.22	8058.79				0.5520
PM411Q02	2	18414.58	10460.95	7953.64				0.3721
M420Q01T	2	19050.58	10685.45	8365.13				0.4101
PM423Q01	2	18647.42	4765.49	13881.94				0.2653

Item	Number of categories	Number of examinees who tried/responded to the category					Item-total corr. (MATHEFF)	Item-total corr. (MATH)
		tried	0	1	2	3		
PM442Q02	2	18107.43	11976.11	6131.31			0.5536	
PM446Q01	2	19017.16	7643.49	11373.67			0.5377	
PM446Q02	2	18955.63	17638.27	1317.36			0.3934	
PM447Q01	2	18928.76	7100.67	11828.09			0.4714	
M462Q01D	3	18283.04	15091.33	1673.20	1518.50		0.4238	
M464Q01T	2	18739.81	14770.05	3969.76			0.5411	
PM474Q01	2	18807.90	6166.15	12641.75			0.3689	
M496Q01T	2	18681.43	9641.75	9039.69			0.5076	
PM496Q02	2	18846.95	7438.43	11408.53			0.4450	
PM559Q01	2	18765.09	7432.70	11332.39			0.4038	
PM564Q01	2	18457.14	10474.59	7982.55			0.3541	
PM564Q02	2	18432.76	10414.80	8017.96			0.3775	
PM571Q01	2	18544.84	10520.31	8024.53			0.4585	
M603Q01T	2	18661.20	11044.85	7616.35			0.3687	
PM800Q01	2	18766.63	2404.90	16361.73			0.1968	
M803Q01T	2	18512.65	14249.36	4263.29			0.5434	
PM828Q01	2	18612.29	13745.85	4866.44			0.5042	
PM828Q02	2	18851.48	9085.79	9765.69			0.4346	
PM828Q03	2	18645.76	13681.82	4963.95			0.4232	
PM903Q01	3	14445.84	10055.52	1864.14	2526.18		0.5773	
PM903Q03	2	14426.00	10349.14	4076.86			0.5316	
M905Q01T	2	14589.81	3668.92	10920.89			0.3699	
PM905Q02	2	14383.31	7364.97	7018.34			0.6363	
PM906Q01	2	18423.70	8087.45	10336.25			0.4632	
PM906Q02	3	18339.72	10331.02	2039.08	5969.62		0.6251	
PM909Q01	2	18640.06	2728.43	15911.63			0.3546	
PM909Q02	2	18895.86	8133.20	10762.66			0.4898	
PM909Q03	2	18874.39	13063.54	5810.85			0.6024	
PM915Q01	2	18500.68	11772.23	6728.45			0.4163	

Item	Number of categories	Number of examinees who tried/responded to the category					Item-total corr. (MATHEFF)	Item-total corr. (MATH)
		tried	0	1	2	3		
PM915Q02	2	18476.85	6508.74	11968.11			0.4585	
PM918Q01	2	14359.19	1907.47	12451.72			0.2154	
PM918Q02	2	14352.41	3203.31	11149.10			0.4191	
PM918Q05	2	14347.64	3512.14	10835.50			0.3942	
PM919Q01	2	14366.08	2554.96	11811.12			0.3961	
PM919Q02	2	14344.48	8151.01	6193.47			0.3856	
PM923Q01	2	14357.78	6194.92	8162.86			0.4943	
PM923Q03	2	14333.74	7198.39	7135.35			0.4155	
PM923Q04	2	14324.48	11958.44	2366.04			0.4824	
PM924Q02	2	14035.37	5368.98	8666.39			0.5188	
PM934Q01	2	4679.48	3681.26	998.23			0.5263	
PM934Q02	2	4646.09	2538.81	2107.28			0.3531	
PM936Q01	2	4792.45	2878.91	1913.53			0.5245	
PM936Q02	2	4747.53	3032.38	1715.14			0.5438	
PM939Q01	2	4665.13	2524.55	2140.58			0.2623	
PM939Q02	2	4815.26	3160.38	1654.88			0.4596	
PM942Q01	2	4994.79	2625.32	2369.47			0.4187	
PM942Q02	2	4802.57	4139.24	663.34			0.4878	
PM942Q03	3	4766.36	4008.71	193.76	563.89		0.5518	
PM943Q01	2	14214.31	7046.99	7167.31			0.3967	
PM943Q02	2	14184.02	13160.84	1023.17			0.3832	
PM948Q01	2	4850.25	999.04	3851.21			0.2733	
PM948Q02	2	4830.52	1912.20	2918.32			0.3904	
PM948Q03	2	4808.54	4385.52	423.02			0.3711	
M949Q01T	2	19007.19	7136.42	11870.77			0.5132	
M949Q02T	2	18976.99	12921.08	6055.91			0.5116	
PM949Q03	3	18909.74	12763.17	531.63	5614.94		0.5332	
PM953Q02	2	13954.43	7216.29	6738.14			0.5204	
PM953Q03	2	13836.12	6826.50	7009.62			0.5916	

Item	Number of categories	Number of examinees who tried/responded to the category					Item-total corr. (MATHEFF)	Item-total corr. (MATH)
		tried	0	1	2	3		
M953Q04D	3	13656.79	10572.42	1157.29	1927.08		0.5648	
PM954Q01	2	14326.77	4902.68	9424.09			0.5253	
PM954Q02	2	14271.86	9769.10	4502.76			0.4919	
PM954Q04	2	14241.19	10491.90	3749.29			0.5318	
PM955Q01	2	18628.39	6265.94	12362.45			0.3693	
PM955Q02	2	18560.79	12673.21	5887.59			0.5044	
PM955Q03	3	18519.70	15951.90	1050.12	1517.69		0.4821	
PM957Q01	2	4751.80	2151.43	2600.36			0.4208	
PM957Q02	2	4732.73	2961.96	1770.77			0.3286	
PM957Q03	2	4718.43	4320.75	397.68			0.3741	
PM961Q02	2	4726.71	4424.50	302.21			0.3177	
PM961Q03	2	4708.72	2671.93	2036.79			0.4226	
PM961Q05	3	4684.25	2237.78	1137.90	1308.56		0.4259	
PM967Q01	2	4810.97	3529.06	1281.91			0.5258	
M967Q03T	2	4562.36	4151.12	411.24			0.2316	
PM982Q01	2	18716.17	3054.21	15661.96			0.2561	
PM982Q02	2	18679.24	13389.76	5289.48			0.3483	
M982Q03T	2	18667.55	7071.51	11596.05			0.3544	
PM982Q04	2	18580.37	9961.47	8618.89			0.4995	
PM985Q01	2	4884.12	965.65	3918.47			0.2440	
PM985Q02	2	4872.35	2884.33	1988.02			0.4965	
PM985Q03	2	4567.66	3299.93	1267.73			0.5152	
PM991Q01	2	4851.88	1929.45	2922.44			0.3990	
M991Q02D	3	4841.11	4525.90	147.20	168.01		0.3865	
PM992Q01	2	18634.49	4887.59	13746.90			0.3861	
PM992Q02	2	18585.28	15148.68	3436.60			0.4560	
PM992Q03	2	18553.56	16922.02	1631.53			0.4400	
PM995Q01	2	14235.36	6116.52	8118.84			0.5719	
PM995Q02	2	14215.28	13617.27	598.02			0.2650	

Item	Number of categories	Number of examinees who tried/responded to the category					Item-total corr. (MATHEFF)	Item-total corr. (MATH)
		tried	0	1	2	3		
PM995Q03	2	14111.79	7814.16	6297.63			0.4906	
PM998Q02	2	18461.98	6621.74	11840.24			0.4331	
M998Q04T	2	18393.56	11284.72	7108.84			0.1761	

Table A.8. Estimated item parameters from the multi-group MIRT model employed in Phase 2

Item	a-param (MATHEFF)	a-param (MATH)	intercept	intercept	intercept	b-param	d-step	d-step	d-step
ST37Q01	0.7152		2.7866	1.3859	-0.5248	-1.0001	1.2919	0.1398	-1.4317
ST37Q02	1.1244		3.6191	1.9900	-0.2616	-0.9325	0.9608	0.1086	-1.0694
ST37Q03	1.1357		3.4509	1.2746	-0.8002	-0.6777	1.1097	-0.0175	-1.0922
ST37Q04	0.7960		2.8901	1.4993	-0.5145	-0.9545	1.1813	0.1535	-1.3348
ST37Q05	1.0307		3.1741	2.1950	0.7484	-1.1638	0.6477	0.0890	-0.7367
ST37Q06	0.7834		2.3872	0.4000	-0.9736	-0.4539	1.3385	-0.1536	-1.1849
ST37Q07	0.7364		2.2900	1.1471	-0.0274	-0.9079	0.9214	0.0084	-0.9298
ST37Q08	0.6064		2.0978	0.3263	-1.0818	-0.4340	1.6010	-0.1176	-1.4834
PM00FQ01		0.7985	-1.3628			1.0040			
PM00GQ01		0.8016	-4.0043			2.9384			
PM00KQ02		0.7326	-3.0059			2.4134			
PM033Q01		0.5383	0.6002			-0.6558			
PM034Q01T		0.7117	-1.3492			1.1152			
PM155Q01		0.7730	-0.0020			0.0016			
PM155Q02D		0.5356	-1.1447	0.5284		0.3384	-0.9188	0.9188	
PM155Q03D		0.7615	-2.8720	-2.3842		2.0302	-0.1884	0.1884	
PM155Q04T		0.5507	-0.4160			0.4444			
PM192Q01T		0.7777	-1.3917			1.0526			
PM273Q01T		0.4942	-0.4550			0.5416			
PM305Q01		0.4594	-0.0632			0.0810			
PM406Q01		1.1559	-3.0502			1.5523			
PM406Q02		1.3857	-4.4789			1.9013			
PM408Q01T		0.5823	-1.3032			1.3165			
PM411Q01		0.9098	-1.1568			0.7480			
PM411Q02		0.4706	-0.7099			0.8873			
PM420Q01T		0.5308	-0.6898			0.7645			
PM423Q01		0.3966	0.9156			-1.3578			

Item	a-param (MATHEFF)	a-param (MATH)	intercept	intercept	intercept	b-param	d-step	d-step	d-step
PM442Q02		0.9385	-1.8553			1.1629			
PM446Q01		0.9014	-0.0286			0.0187			
PM446Q02		1.2097	-5.2997			2.5770			
PM447Q01		0.7214	0.1729			-0.1410			
PM462Q01D		0.5693	-2.9333	-1.6517		2.3687	-0.6621	0.6621	
PM464Q01T		1.0914	-3.0637			1.6512			
PM474Q01		0.5518	0.4536			-0.4835			
PM496Q01T		0.7484	-0.6170			0.4850			
PM496Q02		0.6663	0.1153			-0.1018			
PM559Q01		0.5405	0.1121			-0.1220			
PM564Q01		0.4381	-0.6295			0.8453			
PM564Q02		0.4788	-0.6538			0.8032			
PM571Q01		0.6254	-0.7938			0.7466			
PM603Q01T		0.4690	-0.7710			0.9671			
PM800Q01		0.3696	1.8222			-2.9000			
PM803Q01T		1.0924	-2.9471			1.5869			
PM828Q01		0.8141	-2.1103			1.5248			
PM828Q02		0.5932	-0.3462			0.3433			
PM828Q03		0.6230	-1.7852			1.6855			
PM903Q01		0.7521	-2.6724	-1.7634		1.7347	-0.3555	0.3555	
PM903Q03		0.9704	-2.6001			1.5761			
PM905Q01T		0.5958	0.7168			-0.7077			
PM905Q02		1.3278	-1.6252			0.7200			
PM906Q01		0.6788	-0.2049			0.1776			
PM906Q02		0.6850	-1.9344	-0.2154		0.9230	-0.7381	0.7381	
PM909Q01		0.8211	1.7868			-1.2800			
PM909Q02		0.7473	-0.2116			0.1666			
PM909Q03		1.1394	-2.3740			1.2256			
PM915Q01		0.5653	-1.1475			1.1941			

Item	a-param (MATHEFF)	a-param (MATH)	intercept	intercept	intercept	b-param	d-step	d-step	d-step
PM915Q02		0.7513	0.2673			-0.2093			
PM918Q01		0.3933	1.6060			-2.4021			
PM918Q02		0.7501	0.8177			-0.6412			
PM918Q05		0.6370	0.6890			-0.6362			
PM919Q01		0.8359	1.2468			-0.8774			
PM919Q02		0.5209	-0.9276			1.0475			
PM923Q01		0.7384	-0.4964			0.3955			
PM923Q03		0.5601	-0.6770			0.7110			
PM923Q04		1.0924	-3.9393			2.1213			
PM924Q02		0.8509	-0.3224			0.2229			
PM934Q01		1.0537	-1.6640			0.9290			
PM934Q02		0.5168	-0.0675			0.0768			
PM936Q01		1.0929	-0.2052			0.1105			
PM936Q02		1.1337	-0.4688			0.2433			
PM939Q01		0.3795	-0.0549			0.0851			
PM939Q02		0.7738	-0.5443			0.4137			
PM942Q01		0.6814	0.1180			-0.1019			
PM942Q02		1.0582	-2.4323			1.3520			
PM942Q03		0.7901	-3.0969	-0.1493		1.2083	-1.0972	1.0972	
PM943Q01		0.5150	-0.5717			0.6530			
PM943Q02		1.2341	-5.6653			2.7005			
PM948Q01		0.6013	1.8074			-1.7683			
PM948Q02		0.7377	0.8222			-0.6556			
PM948Q03		0.9136	-2.8911			1.8615			
PM949Q01T		0.8512	0.0924			-0.0639			
PM949Q02T		0.7850	-1.7162			1.2860			
PM949Q03		0.4627	-3.4479	1.4324		1.2812	-3.1022	3.1022	
PM953Q02		0.8040	-1.0464			0.7656			
PM953Q03		1.0916	-1.2567			0.6773			

Item	a-param (MATHEFF)	a-param (MATH)	intercept	intercept	intercept	b-param	d-step	d-step	d-step
PM953Q04D		0.8697	-3.6023	-2.1428		1.9429	-0.4936	0.4936	
PM954Q01		0.9555	-0.0604			0.0372			
PM954Q02		0.8132	-2.0521			1.4844			
PM954Q04		1.0369	-2.8787			1.6331			
PM955Q01		0.5458	0.3833			-0.4132			
PM955Q02		0.7721	-1.7329			1.3203			
PM955Q03		0.7392	-3.8264	-1.9119		2.2833	-0.7618	0.7618	
PM957Q01		0.7360	0.4814			-0.3847			
PM957Q02		0.4737	-0.4454			0.5531			
PM957Q03		0.8808	-2.9619			1.9781			
PM961Q02		0.8485	-3.2376			2.2445			
PM961Q03		0.7184	-0.0886			0.0726			
PM961Q05		0.4245	-0.4075	0.0114		0.2745	-0.2903	0.2903	
PM967Q01		1.0778	-1.1193			0.6109			
PM967Q03T		0.5231	-2.4597			2.7659			
PM982Q01		0.4514	1.5155			-1.9750			
PM982Q02		0.4644	-1.4536			1.8411			
PM982Q03T		0.4700	0.2154			-0.2696			
PM982Q04		0.7332	-0.7711			0.6187			
PM985Q01		0.5066	1.7401			-2.0207			
PM985Q02		0.9194	-0.2159			0.1382			
PM985Q03		0.9732	-1.1164			0.6747			
PM991Q01		0.7213	0.7685			-0.6267			
PM991Q02D		0.7433	-3.7656	-1.4445		2.0616	-0.9184	0.9184	
PM992Q01		0.6398	0.8394			-0.7718			
PM992Q02		0.8374	-2.8132			1.9763			
PM992Q03		1.2703	-5.1658			2.3922			
PM995Q01		1.0086	-0.7454			0.4347			
PM995Q02		0.9292	-5.3938			3.4146			

Item	a-param (MATHEFF)	a-param (MATH)	intercept	intercept	intercept	b-param	d-step	d-step	d-step
PM995Q03		0.7311	-1.1721			0.9431			
PM998Q02		0.6573	0.2198			-0.1967			
PM998Q04T		0.1943	-0.6440			1.9498			

Table A.9. Correlation estimates and empirical reliabilities from the MIRT models employed in Phase 2

Country	Correlation Estimate	MATH			MATHEFF		
		EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
ARE	0.45	1.26	0.32	0.80	1.16	0.24	0.83
ARG	0.28	1.04	0.39	0.73	0.91	0.20	0.82
AUS	0.66	1.26	0.22	0.85	1.28	0.27	0.83
AUT	0.60	1.15	0.22	0.84	1.16	0.24	0.83
BEL	0.53	1.33	0.24	0.85	1.17	0.22	0.84
BGR	0.33	1.38	0.31	0.82	1.20	0.25	0.83
BRA	0.38	1.10	0.41	0.73	0.88	0.18	0.83
CAN	0.62	1.09	0.21	0.84	1.27	0.27	0.82
CHE	0.65	1.16	0.21	0.84	1.12	0.27	0.80
CHL	0.43	1.07	0.30	0.78	0.88	0.19	0.82
COL	0.22	0.98	0.42	0.70	0.73	0.17	0.81
CRI	0.33	0.84	0.34	0.71	0.81	0.17	0.82
CZE	0.62	1.20	0.22	0.85	0.97	0.23	0.81
DEU	0.62	1.18	0.21	0.85	1.13	0.29	0.79
DNK	0.65	0.99	0.21	0.82	1.00	0.20	0.83
ESP	0.58	1.14	0.24	0.83	1.02	0.24	0.81
EST	0.62	0.92	0.19	0.83	0.90	0.20	0.81
FIN	0.63	1.00	0.20	0.83	1.02	0.19	0.84
FRA	0.61	1.28	0.24	0.84	1.18	0.24	0.83
GBR	0.65	1.26	0.24	0.84	1.19	0.25	0.83
GRC	0.55	1.29	0.30	0.81	1.24	0.22	0.85
HKG	0.58	1.20	0.23	0.84	1.35	0.30	0.82
HRV	0.62	1.16	0.24	0.83	1.14	0.26	0.82
HUN	0.67	1.31	0.26	0.84	1.33	0.29	0.82
IDN	0.25	1.04	0.46	0.69	0.61	0.17	0.78
IRL	0.63	1.01	0.21	0.83	1.11	0.23	0.83

Country	Correlation Estimate	MATH			MATHEFF		
		EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
ISL	0.57	1.17	0.24	0.83	1.46	0.29	0.84
ISR	0.52	1.51	0.31	0.83	1.40	0.29	0.83
ITA	0.59	1.24	0.24	0.84	0.86	0.20	0.81
JOR	0.34	1.11	0.42	0.73	1.27	0.26	0.83
JPN	0.64	1.23	0.22	0.85	1.19	0.20	0.85
KAZ	0.30	0.99	0.30	0.77	0.95	0.25	0.79
KOR	0.67	1.28	0.22	0.85	1.33	0.22	0.86
LTU	0.62	1.17	0.24	0.83	1.13	0.24	0.82
LUX	0.58	1.35	0.25	0.85	1.47	0.30	0.83
LVA	0.57	1.01	0.21	0.82	0.83	0.19	0.81
MAC	0.57	1.13	0.21	0.85	1.08	0.26	0.81
MEX	0.40	0.96	0.34	0.74	0.83	0.18	0.82
MNE	0.38	1.16	0.36	0.77	1.15	0.23	0.84
MYS	0.49	1.15	0.34	0.77	0.73	0.17	0.81
NLD	0.54	1.11	0.21	0.84	1.02	0.20	0.83
NOR	0.66	1.20	0.24	0.83	1.47	0.27	0.84
NZL	0.63	1.34	0.24	0.85	1.19	0.23	0.84
PER	0.29	1.25	0.51	0.71	0.64	0.18	0.78
POL	0.70	1.17	0.21	0.85	1.27	0.27	0.82
PRT	0.71	1.28	0.25	0.84	1.23	0.29	0.81
QAT	0.36	1.59	0.50	0.76	1.59	0.29	0.85
ROU	0.43	1.12	0.27	0.80	0.91	0.20	0.82
RUS	0.54	1.17	0.24	0.83	0.94	0.21	0.82
SGP	0.58	1.36	0.23	0.86	1.27	0.36	0.78
SRB	0.49	1.26	0.29	0.81	1.07	0.21	0.84
SVK	0.65	1.44	0.26	0.85	1.02	0.24	0.81
SVN	0.54	1.13	0.22	0.84	1.24	0.31	0.80
SWE	0.60	1.21	0.25	0.83	1.16	0.24	0.83

Country	Correlation Estimate	MATH			MATHEFF		
		EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
TAP	0.69	1.62	0.24	0.87	1.67	0.33	0.83
THA	0.32	1.13	0.32	0.78	0.57	0.15	0.79
TUN	0.41	1.13	0.41	0.73	0.96	0.19	0.83
TUR	0.54	1.27	0.28	0.82	1.02	0.22	0.82
URY	0.43	1.25	0.37	0.77	0.86	0.20	0.81
USA	0.62	1.18	0.23	0.84	1.20	0.27	0.82
VNM	0.61	1.05	0.22	0.83	0.48	0.14	0.77

Table A.10. Parameter estimates from the MLM models employed in Phase 3

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
ARE	-0.02	0.47	0.55	0.46	-0.02	0.26	0.01	0.40	0.49
ARG	-0.12	0.48	0.52	0.48	-0.12	0.18	0.01	0.47	0.49
AUS	0.05	0.26	0.69	0.28	0.04	0.50	0.01	0.13	0.49
AUT	-0.13	0.52	0.55	0.48	-0.11	0.35	0.01	0.34	0.46
BEL	-0.10	0.53	0.53	0.50	-0.08	0.33	0.01	0.42	0.44
BGR	-0.15	0.56	0.49	0.54	-0.15	0.15	0.01	0.52	0.47
BRA	-0.12	0.45	0.56	0.45	-0.12	0.19	0.01	0.41	0.53
CAN	0.03	0.19	0.82	0.19	0.03	0.51	0.01	0.11	0.59
CHE	-0.03	0.33	0.65	0.33	-0.01	0.46	0.01	0.19	0.49
CHL	-0.16	0.56	0.47	0.54	-0.14	0.23	0.01	0.47	0.43
COL	-0.09	0.36	0.63	0.36	-0.10	0.11	0.01	0.35	0.62
CRI	-0.04	0.42	0.62	0.40	-0.04	0.15	0.02	0.39	0.60
CZE	-0.15	0.47	0.48	0.49	-0.11	0.37	0.01	0.31	0.39
DEU	-0.06	0.51	0.52	0.50	-0.05	0.39	0.01	0.35	0.39
DNK	0.08	0.15	0.76	0.17	0.09	0.52	0.01	0.09	0.52
ESP	-0.07	0.16	0.79	0.17	-0.06	0.46	0.01	0.12	0.60
EST	-0.04	0.19	0.81	0.19	-0.02	0.49	0.02	0.12	0.61
FIN	0.13	0.07	0.84	0.07	0.13	0.51	0.01	0.04	0.59
FRA	-0.12	0.61	0.44	0.58	-0.10	0.34	0.01	0.44	0.35
GBR	0.01	0.24	0.77	0.24	0.00	0.52	0.01	0.14	0.53
GRC	-0.12	0.40	0.68	0.37	-0.10	0.38	0.01	0.30	0.55
HKG	-0.02	0.43	0.58	0.43	-0.02	0.40	0.01	0.28	0.44
HRV	-0.02	0.43	0.57	0.43	-0.02	0.38	0.01	0.27	0.46
HUN	-0.25	0.73	0.39	0.65	-0.21	0.33	0.01	0.50	0.31
IDN	-0.03	0.48	0.50	0.49	-0.03	0.11	0.02	0.46	0.49
IRL	-0.02	0.18	0.82	0.18	-0.01	0.51	0.02	0.11	0.59

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
ISL	-0.05	0.12	0.89	0.12	-0.03	0.48	0.02	0.08	0.68
ISR	-0.05	0.43	0.63	0.41	-0.04	0.39	0.01	0.33	0.49
ITA	-0.15	0.55	0.50	0.52	-0.13	0.31	0.01	0.41	0.42
JOR	-0.03	0.34	0.67	0.33	-0.03	0.22	0.01	0.31	0.63
JPN	-0.03	0.56	0.46	0.55	-0.02	0.32	0.01	0.35	0.39
KAZ	0.04	0.39	0.61	0.39	0.04	0.19	0.02	0.36	0.58
KOR	-0.03	0.39	0.61	0.39	-0.02	0.46	0.01	0.18	0.46
LTU	-0.10	0.32	0.70	0.31	-0.07	0.45	0.02	0.21	0.53
LUX	-0.01	0.33	0.69	0.32	0.00	0.39	0.02	0.21	0.55
LVA	-0.12	0.30	0.71	0.30	-0.11	0.43	0.02	0.22	0.56
MAC	-0.17	0.44	0.70	0.38	-0.14	0.41	0.01	0.28	0.55
MEX	-0.09	0.36	0.64	0.36	-0.09	0.25	0.01	0.32	0.59
MNE	-0.09	0.35	0.66	0.35	-0.08	0.24	0.02	0.32	0.61
MYS	-0.01	0.31	0.69	0.31	-0.01	0.31	0.02	0.26	0.60
NLD	-0.02	0.63	0.37	0.63	-0.02	0.27	0.02	0.51	0.31
NOR	0.00	0.13	0.88	0.13	0.00	0.58	0.02	0.08	0.58
NZL	-0.05	0.25	0.76	0.25	-0.04	0.51	0.02	0.14	0.55
PER	-0.13	0.48	0.55	0.47	-0.13	0.17	0.01	0.48	0.52
POL	0.01	0.22	0.79	0.22	0.00	0.58	0.01	0.11	0.50
PRT	-0.05	0.32	0.69	0.32	-0.03	0.54	0.01	0.15	0.46
QAT	-0.04	0.48	0.53	0.47	-0.05	0.17	0.01	0.44	0.50
ROU	-0.01	0.50	0.56	0.47	-0.01	0.21	0.02	0.43	0.52
RUS	-0.07	0.31	0.71	0.30	-0.05	0.41	0.02	0.22	0.57
SGP	-0.01	0.36	0.64	0.36	-0.01	0.42	0.01	0.21	0.50
SRB	-0.03	0.48	0.54	0.47	-0.04	0.27	0.01	0.40	0.48
SVK	-0.13	0.44	0.52	0.46	-0.10	0.37	0.01	0.30	0.41
SVN	-0.14	0.56	0.44	0.56	-0.12	0.27	0.01	0.44	0.39
SWE	0.01	0.14	0.88	0.13	0.00	0.50	0.02	0.08	0.66

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
TAP	-0.03	0.45	0.58	0.44	-0.02	0.49	0.01	0.20	0.41
THA	-0.05	0.42	0.46	0.48	-0.05	0.16	0.01	0.38	0.44
TUN	-0.04	0.50	0.51	0.49	-0.04	0.18	0.01	0.44	0.48
TUR	-0.13	0.60	0.37	0.62	-0.12	0.21	0.01	0.50	0.34
URY	-0.11	0.45	0.58	0.44	-0.10	0.24	0.01	0.40	0.53
USA	-0.02	0.24	0.78	0.23	-0.01	0.49	0.02	0.14	0.57
VNM	-0.09	0.56	0.50	0.53	-0.09	0.33	0.01	0.43	0.42

Table A.11. Correlation estimates, class sizes, and class means from the MLMixMIRT models employed in Phase 4

Country	High Math Class				Low Math Class			
	Correlation Estimate	Class Size	MATH Mean	MATHEFF Mean	Correlation Estimate	Class Size	MATH Mean	MATHEFF Mean
ARE	0.59	0.55	-0.27	0.87	0.14	0.45	-1.38	0.51
ARG	0.51	0.58	-0.76	0.27	-0.04	0.42	-1.54	0.58
AUS	0.62	0.63	0.06	0.97	0.42	0.37	-0.62	0.50
AUT	0.58	0.57	0.28	1.06	0.28	0.43	-0.82	0.26
BEL	0.41	0.53	0.52	1.36	0.23	0.47	-0.74	0.16
BGR	0.57	0.45	-0.08	0.57	-0.11	0.55	-0.89	0.60
BRA	0.59	0.57	-0.85	0.28	-0.02	0.43	-1.56	0.37
CAN	0.50	0.65	0.18	1.01	0.44	0.35	-0.48	0.04
CHE	0.60	0.47	0.63	2.03	0.44	0.53	-0.46	0.49
CHL	0.61	0.57	-0.48	0.45	-0.01	0.43	-1.62	0.93
COL	0.53	0.45	-0.79	0.52	-0.07	0.55	-2.06	0.26
CRI	0.51	0.55	-0.83	0.29	-0.03	0.45	-1.34	0.65
CZE	0.63	0.53	0.30	0.90	0.30	0.47	-1.02	0.42
DEU	0.61	0.55	0.49	1.33	0.45	0.45	-0.77	0.55
DNK	0.72	0.37	-0.19	2.23	0.57	0.63	-0.29	0.45
ESP	0.49	0.67	-0.12	1.13	0.40	0.33	-0.74	0.09
EST	0.72	0.53	0.12	1.85	0.40	0.47	-0.23	0.45
FIN	0.59	0.61	0.20	0.90	0.34	0.39	-0.41	-0.09
FRA	0.61	0.58	0.28	1.14	0.25	0.42	-1.33	0.23
GBR	0.54	0.45	0.31	2.19	0.43	0.55	-1.09	0.37
GRC	0.51	0.59	-0.28	1.11	0.16	0.41	-1.21	0.04
HKG	0.42	0.49	0.80	3.82	0.52	0.51	-0.21	0.59
HRV	0.59	0.66	-0.33	0.89	0.52	0.34	-1.26	0.46
HUN	0.62	0.61	0.02	1.82	0.18	0.39	-2.47	0.34
IDN	0.32	0.50	-1.09	0.55	-0.08	0.50	-1.68	0.96
IRL	0.64	0.56	-0.04	0.78	0.34	0.44	-0.58	0.76

Country	High Math Class				Low Math Class			
	Correlation Estimate	Class Size	MATH Mean	MATHEFF Mean	Correlation Estimate	Class Size	MATH Mean	MATHEFF Mean
ISL	0.64	0.74	-0.15	0.75	0.39	0.26	-0.76	2.02
ISR	0.64	0.60	-0.06	0.97	0.27	0.40	-1.30	0.59
ITA	0.50	0.62	0.05	1.10	0.25	0.38	-0.91	0.17
JOR	0.56	0.74	-0.95	0.46	0.12	0.26	-1.69	2.66
JPN	0.56	0.28	0.24	2.79	0.66	0.72	0.05	0.09
KAZ	0.37	0.52	-0.38	0.98	-0.03	0.48	-0.76	1.30
KOR	0.64	0.51	0.53	1.16	0.56	0.49	-0.12	-0.21
LTU	0.69	0.55	0.01	0.80	0.26	0.45	-0.96	0.55
LUX	0.50	0.50	0.29	1.54	0.25	0.50	-0.92	0.30
LVA	0.59	0.55	0.06	0.82	0.21	0.45	-0.80	0.43
MAC	0.63	0.77	0.15	0.84	0.47	0.23	0.04	4.19
MEX	0.52	0.61	-0.68	0.54	0.07	0.39	-1.04	0.71
MNE	0.52	0.61	-0.79	0.42	0.05	0.39	-1.37	0.79
MYS	0.47	0.43	-0.38	0.86	0.16	0.57	-1.62	0.37
NLD	0.50	0.54	0.60	0.95	0.37	0.46	-0.80	0.29
NOR	0.68	0.85	-0.28	0.60	0.67	0.15	-0.44	4.72
NZL	0.63	0.61	0.02	0.76	0.39	0.39	-0.67	0.44
PER	0.46	0.39	-0.72	0.65	-0.10	0.61	-1.93	0.64
POL	0.72	0.62	0.14	0.91	0.61	0.38	-0.49	0.65
PRT	0.60	0.51	0.24	2.07	0.43	0.49	-1.25	0.56
QAT	0.57	0.69	-0.89	0.48	-0.04	0.31	-1.83	2.12
ROU	0.54	0.61	-0.26	0.55	-0.01	0.39	-0.78	0.50
RUS	0.55	0.63	-0.11	0.72	0.30	0.37	-0.62	0.38
SGP	0.52	0.36	0.94	4.12	0.48	0.64	0.02	0.94
SRB	0.61	0.59	-0.30	0.46	0.03	0.41	-0.99	0.90
SVK	0.61	0.56	0.18	1.54	0.26	0.44	-1.38	0.36
SVN	0.41	0.45	0.46	2.33	0.20	0.55	-0.88	0.65
SWE	0.36	0.45	0.24	2.24	0.19	0.55	-1.05	0.30

Country	High Math Class				Low Math Class			
	Correlation Estimate	Class Size	MATH Mean	MATHEFF Mean	Correlation Estimate	Class Size	MATH Mean	MATHEFF Mean
TAP	0.73	0.61	0.27	0.81	0.70	0.39	0.25	4.06
THA	0.46	0.45	-0.49	0.51	0.03	0.55	-1.21	0.83
TUN	0.57	0.42	-0.66	0.35	-0.07	0.58	-1.45	0.27
TUR	0.56	0.40	0.02	1.07	0.16	0.60	-1.42	0.42
URY	0.53	0.56	-0.45	0.46	-0.04	0.44	-1.82	0.42
USA	0.61	0.72	-0.24	0.90	0.31	0.28	-0.91	0.90
VNM	0.50	0.60	0.36	1.07	0.06	0.40	-0.66	0.29

Table A.12. Empirical reliabilities for the High Math Class from the MLMixMIRT models employed in Phase 4

Country	High Math Class					
	MATH			MATHEFF		
	EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
ARE	0.58	0.14	0.80	0.61	0.18	0.78
ARG	0.57	0.17	0.77	0.24	0.09	0.74
AUS	0.73	0.12	0.85	0.96	0.21	0.82
AUT	0.45	0.11	0.81	0.76	0.21	0.79
BEL	0.36	0.11	0.77	1.27	0.27	0.83
BGR	0.60	0.12	0.83	0.68	0.15	0.82
BRA	0.68	0.19	0.78	0.32	0.10	0.77
CAN	0.56	0.11	0.83	0.58	0.20	0.74
CHE	0.31	0.11	0.74	0.98	0.32	0.75
CHL	0.61	0.15	0.80	0.49	0.12	0.81
COL	0.61	7.29	0.08	0.48	0.12	0.80
CRI	0.69	0.22	0.76	0.34	0.09	0.78
CZE	0.46	0.10	0.82	0.62	0.16	0.79
DEU	0.35	0.10	0.78	0.65	0.22	0.74
DNK	0.99	0.17	0.86	2.64	0.32	0.89
ESP	0.58	0.12	0.83	0.54	0.18	0.75
EST	0.67	0.11	0.85	2.19	0.34	0.87
FIN	0.57	0.12	0.83	0.89	0.17	0.84
FRA	0.42	0.11	0.80	0.91	0.23	0.80
GBR	0.43	0.10	0.80	1.62	0.37	0.81
GRC	0.65	0.17	0.79	0.74	0.17	0.81
HKG	0.39	0.13	0.74	3.18	0.75	0.81
HRV	0.77	0.14	0.85	0.59	0.18	0.76
HUN	0.49	0.11	0.82	1.12	0.32	0.78
IDN	0.66	0.23	0.74	0.42	0.13	0.77

Country	High Math Class					
	MATH			MATHEFF		
	EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
IRL	0.66	0.13	0.84	0.67	0.17	0.80
ISL	0.70	0.14	0.83	0.70	0.17	0.81
ISR	0.63	0.13	0.83	0.96	0.21	0.82
ITA	0.49	0.11	0.82	0.89	0.21	0.81
JOR	0.54	0.19	0.74	0.38	0.14	0.73
JPN	0.94	0.13	0.88	13.36	0.73	0.95
KAZ	0.44	0.13	0.78	0.70	0.21	0.77
KOR	0.55	0.12	0.82	1.25	0.23	0.84
LTU	0.56	0.11	0.84	0.73	0.19	0.80
LUX	0.41	0.10	0.80	0.85	0.28	0.75
LVA	0.49	0.11	0.82	0.75	0.18	0.81
MAC	0.55	0.11	0.84	0.75	0.17	0.82
MEX	0.58	0.17	0.77	0.39	0.12	0.77
MNE	0.70	0.18	0.79	0.42	0.13	0.77
MYS	0.45	0.13	0.78	0.74	0.19	0.80
NLD	0.33	0.11	0.75	0.87	0.20	0.82
NOR	0.74	0.15	0.83	0.61	0.14	0.82
NZL	0.84	0.13	0.86	0.99	0.19	0.84
PER	0.72	0.25	0.75	0.62	0.15	0.81
POL	0.63	0.11	0.85	0.75	0.23	0.76
PRT	0.39	0.10	0.79	1.33	0.39	0.77
QAT	0.83	0.22	0.79	0.49	0.14	0.78
ROU	0.53	0.11	0.82	0.39	0.14	0.74
RUS	0.62	0.13	0.83	0.61	0.16	0.79
SGP	0.62	0.17	0.78	3.95	0.78	0.83
SRB	0.70	0.14	0.83	0.61	0.14	0.81
SVK	0.46	0.10	0.82	1.05	0.28	0.79

Country	High Math Class					
	MATH			MATHEFF		
	EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
SVN	0.32	0.10	0.76	1.68	0.59	0.74
SWE	0.44	1.14	0.28	1.00	0.32	0.76
TAP	0.82	0.13	0.86	1.02	0.21	0.83
THA	0.77	0.15	0.84	0.39	0.11	0.78
TUN	0.62	0.20	0.76	0.49	0.13	0.79
TUR	0.40	0.10	0.81	0.73	0.25	0.75
URY	0.67	0.16	0.80	0.37	0.11	0.77
USA	0.76	0.14	0.85	0.79	0.19	0.80
VNM	0.36	0.10	0.79	0.95	0.21	0.82

Table A.13. Empirical reliabilities for the Low Math Class from the MLMixMIRT models employed in Phase 4

Country	Low Math Class					
	MATH			MATHEFF		
	EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
ARE	0.53	0.26	0.67	0.93	0.16	0.85
ARG	0.55	0.26	0.68	3.06	0.32	0.91
AUS	0.67	0.17	0.80	0.93	0.19	0.83
AUT	0.55	0.18	0.76	0.49	0.12	0.80
BEL	0.55	0.18	0.76	0.42	0.10	0.81
BGR	0.47	0.17	0.74	0.98	0.20	0.83
BRA	0.64	0.32	0.67	3.18	0.33	0.91
CAN	0.57	0.14	0.80	0.71	0.15	0.83
CHE	0.53	0.15	0.78	0.53	0.13	0.81
CHL	0.50	0.23	0.68	1.35	0.22	0.86
COL	0.50	0.35	0.59	0.93	0.18	0.84
CRI	0.44	0.20	0.69	1.32	0.20	0.87
CZE	0.57	0.17	0.77	0.55	0.13	0.81
DEU	0.48	0.15	0.76	0.54	0.13	0.80
DNK	0.58	0.13	0.82	0.55	0.12	0.82
ESP	0.65	0.17	0.80	0.46	0.13	0.78
EST	0.47	0.12	0.80	0.38	0.10	0.79
FIN	0.58	0.14	0.81	0.76	0.14	0.84
FRA	0.63	0.24	0.72	0.45	0.10	0.81
GBR	0.81	0.26	0.76	0.63	0.12	0.84
GRC	0.80	0.32	0.71	0.57	0.11	0.84
HKG	0.68	0.15	0.82	0.85	0.16	0.85
HRV	0.85	0.24	0.78	0.47	0.16	0.74
HUN	1.21	0.51	0.71	0.79	0.14	0.85
IDN	0.69	0.31	0.69	5.80	0.50	0.92

Country	Low Math Class					
	MATH			MATHEFF		
	EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
IRL	0.63	0.15	0.80	0.55	0.16	0.77
ISL	0.86	0.18	0.83	17.11	0.75	0.96
ISR	0.81	0.33	0.71	0.87	0.19	0.82
ITA	0.52	0.16	0.76	0.46	0.11	0.81
JOR	0.66	0.29	0.70	12.01	0.51	0.96
JPN	0.69	0.12	0.85	0.58	0.10	0.85
KAZ	0.33	0.15	0.69	1.75	0.27	0.87
KOR	0.85	0.14	0.86	1.12	0.18	0.86
LTU	0.48	0.15	0.76	0.59	0.13	0.82
LUX	0.47	0.15	0.76	0.63	0.13	0.83
LVA	0.55	0.18	0.75	0.39	0.12	0.77
MAC	0.82	0.13	0.86	11.52	0.70	0.94
MEX	0.45	0.21	0.68	4.34	0.33	0.93
MNE	0.59	0.24	0.71	5.40	0.37	0.94
MYS	0.45	0.21	0.68	0.77	0.14	0.85
NLD	0.53	0.18	0.75	0.94	0.15	0.86
NOR	1.41	0.23	0.86	26.77	0.87	0.97
NZL	0.74	0.17	0.81	0.71	0.19	0.79
PER	0.84	0.47	0.64	0.90	0.19	0.82
POL	0.72	0.15	0.82	0.58	0.15	0.80
PRT	0.67	0.27	0.71	0.79	0.15	0.84
QAT	0.60	0.43	0.58	19.35	0.59	0.97
ROU	0.35	0.16	0.69	2.09	0.25	0.89
RUS	0.61	0.16	0.79	0.89	0.21	0.81
SGP	0.60	0.11	0.84	0.79	0.23	0.78
SRB	0.70	0.21	0.77	1.37	0.22	0.86
SVK	0.63	0.23	0.73	0.58	0.14	0.81

Country	Low Math Class					
	MATH			MATHEFF		
	EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
SVN	0.45	0.18	0.72	0.73	0.16	0.82
SWE	0.56	0.20	0.74	0.60	0.12	0.83
TAP	1.05	0.14	0.88	20.26	0.87	0.96
THA	0.54	0.25	0.68	3.69	0.36	0.91
TUN	0.51	0.27	0.65	0.47	0.12	0.79
TUR	0.55	0.25	0.68	0.67	0.13	0.84
URY	0.75	0.37	0.67	1.94	0.28	0.88
USA	0.64	0.18	0.79	1.37	0.27	0.84
VNM	0.34	0.13	0.72	0.42	0.13	0.77

Table A.14. Correlation estimates by super-school and country from the OLS models employed in Phase 5

Country	Super-school ID	Std. Regression Coef. (Corr. Est.)	SE	N	PV1MATH Weighted Mean	MATHEFF Weighted Mean
KAZ	1	0.10	0.05	532	371.78	0.02
KAZ	2	0.26	0.07	319	390.49	0.08
KAZ	3	0.20	0.06	357	400.20	-0.15
KAZ	4	0.20	0.07	321	411.16	0.00
KAZ	5	0.16	0.06	359	422.79	0.13
KAZ	6	0.33	0.07	320	433.61	0.18
KAZ	7	0.28	0.06	357	446.86	0.20
KAZ	8	0.30	0.06	335	467.03	0.40
KAZ	9	0.29	0.06	514	516.11	0.41
JOR	1	0.09	0.05	616	325.00	-0.04
JOR	2	0.27	0.05	410	352.08	-0.18
JOR	3	0.25	0.05	433	363.74	-0.07
JOR	4	0.21	0.06	432	373.31	-0.06
JOR	5	0.30	0.05	441	381.76	-0.18
JOR	6	0.32	0.06	441	393.45	-0.03
JOR	7	0.34	0.06	417	401.46	-0.08
JOR	8	0.32	0.05	449	415.35	0.06
JOR	9	0.34	0.04	698	467.19	0.29
GRC	1	0.31	0.05	431	365.48	-0.72
GRC	2	0.35	0.06	311	428.17	-0.34
GRC	3	0.41	0.05	307	442.13	-0.20
GRC	4	0.41	0.06	320	457.13	-0.05
GRC	5	0.42	0.06	320	467.87	-0.09
GRC	6	0.45	0.06	327	474.52	-0.10
GRC	7	0.49	0.06	307	483.24	-0.09
GRC	8	0.48	0.05	319	495.43	-0.01
GRC	9	0.49	0.04	478	527.21	0.30

Country	Super-school ID	Std. Regression Coef. (Corr. Est.)	SE	N	PV1MATH Weighted Mean	MATHEFF Weighted Mean
NLD	1	0.24	0.06	359	414.96	-0.61
NLD	2	0.44	0.07	241	457.56	-0.49
NLD	3	0.42	0.07	229	480.45	-0.37
NLD	4	0.37	0.07	248	503.68	-0.19
NLD	5	0.46	0.09	232	526.99	-0.24
NLD	6	0.49	0.06	241	560.91	-0.02
NLD	7	0.42	0.07	255	582.57	0.05
NLD	8	0.48	0.07	237	597.73	0.08
NLD	9	0.48	0.05	364	624.45	0.24
PRT	1	0.46	0.07	455	396.23	-0.34
PRT	2	0.61	0.05	324	431.38	-0.12
PRT	3	0.55	0.05	323	459.80	0.06
PRT	4	0.61	0.05	324	481.85	0.11
PRT	5	0.62	0.05	306	495.46	0.43
PRT	6	0.60	0.05	336	508.51	0.39
PRT	7	0.58	0.05	341	518.81	0.45
PRT	8	0.57	0.04	339	533.64	0.62
PRT	9	0.60	0.04	523	565.27	0.81
TAP	1	0.54	0.04	565	441.49	-0.71
TAP	2	0.60	0.04	386	491.86	-0.44
TAP	3	0.66	0.04	399	514.58	-0.09
TAP	4	0.53	0.04	408	537.45	0.07
TAP	5	0.63	0.04	395	553.58	0.22
TAP	6	0.56	0.04	397	574.46	0.29
TAP	7	0.56	0.04	398	594.43	0.41
TAP	8	0.48	0.05	410	620.14	0.68
TAP	9	0.35	0.04	622	685.68	1.01

Table A.15. Correlation estimates and empirical reliabilities by super-school from the MIRT models employed in Phase 5

Country	Super-school ID	Correlation Estimate	MATH			MATHEFF		
			EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
KAZ	1	0.17	1.69	0.94	0.64	1.07	0.22	0.83
KAZ	2	0.32	1.56	0.82	0.65	0.94	0.20	0.82
KAZ	3	0.30	1.59	0.80	0.67	0.66	0.16	0.81
KAZ	4	0.19	1.81	0.83	0.69	0.69	0.17	0.80
KAZ	5	0.18	1.77	0.72	0.71	0.98	0.21	0.83
KAZ	6	0.43	1.75	0.62	0.74	0.97	0.22	0.81
KAZ	7	0.32	1.79	0.63	0.74	1.03	0.23	0.82
KAZ	8	0.35	1.71	0.59	0.74	1.19	0.28	0.81
KAZ	9	0.38	1.63	0.53	0.76	1.10	0.27	0.80
JOR	1	0.05	1.84	1.29	0.59	0.81	0.22	0.79
JOR	2	0.36	1.47	0.93	0.61	0.65	0.14	0.82
JOR	3	0.29	1.85	0.91	0.67	0.73	0.17	0.81
JOR	4	0.21	1.57	0.81	0.66	0.70	0.17	0.80
JOR	5	0.44	1.86	0.85	0.69	0.53	0.13	0.80
JOR	6	0.43	1.70	0.71	0.70	0.45	0.13	0.78
JOR	7	0.48	1.54	0.67	0.70	0.57	0.13	0.81
JOR	8	0.43	1.92	0.66	0.75	0.59	0.15	0.80
JOR	9	0.37	2.05	0.54	0.79	0.63	0.20	0.76
GRC	1	0.33	1.78	0.92	0.66	1.65	0.34	0.83
GRC	2	0.42	1.86	0.56	0.77	2.40	0.48	0.83
GRC	3	0.50	1.51	0.49	0.76	2.00	0.43	0.82
GRC	4	0.43	1.58	0.45	0.78	2.07	0.49	0.81
GRC	5	0.47	1.68	0.48	0.78	2.24	0.50	0.82
GRC	6	0.58	1.62	0.39	0.80	1.71	0.42	0.80
GRC	7	0.58	1.47	0.33	0.82	1.66	0.41	0.80
GRC	8	0.53	1.50	0.40	0.79	1.77	0.46	0.79

Country	Super-school ID	Correlation Estimate	MATH			MATHEFF		
			EAP Variance	Error Variance	EAP Reliability	EAP Variance	Error Variance	EAP Reliability
GRC	9	0.56	1.72	0.43	0.80	2.28	0.66	0.78
NLD	1	0.29	1.54	0.63	0.71	1.60	0.25	0.87
NLD	2	0.49	1.40	0.41	0.77	1.08	0.19	0.85
NLD	3	0.44	1.08	0.37	0.74	1.45	0.24	0.86
NLD	4	0.39	0.95	0.33	0.74	1.22	0.23	0.84
NLD	5	0.44	1.04	0.29	0.78	1.10	0.22	0.83
NLD	6	0.60	0.79	0.28	0.74	1.31	0.28	0.82
NLD	7	0.52	0.71	0.28	0.72	1.13	0.27	0.81
NLD	8	0.52	0.77	0.27	0.74	0.94	0.25	0.79
NLD	9	0.50	1.02	0.33	0.76	1.24	0.32	0.80
PRT	1	0.55	1.53	0.68	0.69	1.37	0.21	0.87
PRT	2	0.69	1.77	0.51	0.78	1.25	0.21	0.85
PRT	3	0.60	1.57	0.44	0.78	1.38	0.26	0.84
PRT	4	0.71	1.70	0.36	0.82	1.56	0.26	0.86
PRT	5	0.68	1.73	0.33	0.84	1.70	0.36	0.82
PRT	6	0.62	1.55	0.31	0.84	1.45	0.32	0.82
PRT	7	0.64	1.48	0.29	0.84	1.40	0.33	0.81
PRT	8	0.69	1.28	0.27	0.83	1.42	0.39	0.78
PRT	9	0.65	1.26	0.29	0.81	1.58	0.49	0.76
TAP	1	0.61	1.25	0.32	0.79	1.23	0.15	0.89
TAP	2	0.67	1.06	0.21	0.83	1.28	0.16	0.89
TAP	3	0.69	1.24	0.22	0.85	1.63	0.23	0.88
TAP	4	0.52	1.09	0.20	0.84	1.49	0.25	0.86
TAP	5	0.68	1.06	0.19	0.85	1.59	0.27	0.85
TAP	6	0.55	0.90	0.18	0.83	1.30	0.25	0.84
TAP	7	0.55	1.06	0.22	0.83	1.38	0.29	0.82
TAP	8	0.53	0.83	0.21	0.80	1.42	0.38	0.79
TAP	9	0.45	0.54	0.29	0.65	1.11	0.46	0.70

Table A.16. Results from the PCA analysis conducted in Phase 6

Country	Determinant	MSA	First Principal Component	
			Eigenvalue	Variance Explained
ARE	0.00117	0.90	5.61	31.16
ARG	0.00226	0.87	4.67	25.92
AUS	0.00007	0.94	7.78	43.22
AUT	0.00012	0.92	6.96	38.65
BEL	0.00037	0.91	5.92	32.88
BGR	0.00040	0.88	5.21	28.96
BRA	0.00207	0.88	5.19	28.81
CAN	0.00006	0.94	7.73	42.93
CHE	0.00024	0.92	6.76	37.56
CHL	0.00036	0.91	6.24	34.65
COL	0.00143	0.90	5.47	30.38
CRI	0.00133	0.89	5.89	32.70
CZE	0.00016	0.92	6.95	38.63
DEU	0.00012	0.93	7.19	39.93
DNK	0.00008	0.95	8.53	47.38
ESP	0.00057	0.90	6.12	34.02
EST	0.00014	0.93	7.34	40.77
FIN	0.00007	0.94	8.13	45.15
FRA	0.00041	0.91	6.31	35.04
GBR	0.00016	0.94	7.60	42.23
GRC	0.00037	0.91	6.48	36.02
HKG	0.00004	0.93	7.60	42.22
HRV	0.00026	0.92	6.69	37.14
HUN	0.00031	0.91	6.38	35.46
IDN	0.00289	0.87	4.67	25.94
IRL	0.00011	0.94	7.72	42.91
ISL	0.00002	0.93	8.12	45.10
ISR	0.00026	0.90	6.35	35.31
ITA	0.00081	0.89	5.68	31.57
JOR	0.00097	0.89	5.56	30.91
JPN	0.00009	0.91	6.83	37.93
KAZ	0.00063	0.91	5.94	32.98
KOR	0.00010	0.92	7.12	39.55
LTU	0.00057	0.90	6.15	34.16
LUX	0.00012	0.91	6.62	36.80
LVA	0.00093	0.91	6.24	34.68
MAC	0.00007	0.93	7.24	40.24
MEX	0.00106	0.90	5.67	31.50
MNE	0.00091	0.89	5.42	30.11

Country	Determinant	MSA	First Principal Component	
			Eigenvalue	Variance Explained
MYS	0.00355	0.87	4.88	27.12
NLD	0.00018	0.92	6.91	38.37
NOR	0.00003	0.95	8.37	46.51
NZL	0.00020	0.93	7.29	40.48
PER	0.00285	0.87	4.98	27.69
POL	0.00006	0.94	7.90	43.90
PRT	0.00011	0.91	6.98	38.75
QAT	0.00011	0.92	6.17	34.26
ROU	0.01280	0.83	3.93	21.83
RUS	0.00078	0.91	6.32	35.13
SGP	0.00018	0.92	6.95	38.62
SRB	0.00041	0.90	6.14	34.10
SVK	0.00066	0.90	6.13	34.06
SVN	0.00030	0.90	6.32	35.10
SWE	0.00014	0.94	7.58	42.11
TAP	0.00002	0.93	7.41	41.15
THA	0.00204	0.87	5.00	27.77
TUN	0.00272	0.88	4.98	27.66
TUR	0.00056	0.89	5.95	33.05
URY	0.00040	0.90	6.05	33.59
USA	0.00005	0.94	7.93	44.03
VNM	0.00546	0.87	5.09	28.27

Table A.17. Correlation estimates from the OLS models employed in Phase 6

Country	OLS_matheff		OLS_composite	
	Std. Regression Coef. (Corr. Est.)	SE	Std. Regression Coef. (Corr. Est.)	SE
ARE	0.31	0.02	0.42	0.02
ARG	0.18	0.03	0.23	0.02
AUS	0.57	0.01	0.57	0.01
AUT	0.50	0.02	0.51	0.02
BEL	0.45	0.02	0.44	0.02
BGR	0.22	0.02	0.35	0.02
BRA	0.26	0.02	0.33	0.01
CAN	0.57	0.01	0.60	0.01
CHE	0.56	0.02	0.50	0.01
CHL	0.29	0.02	0.42	0.02
COL	0.14	0.02	0.28	0.02
CRI	0.24	0.03	0.34	0.03
CZE	0.52	0.02	0.53	0.02
DEU	0.53	0.02	0.50	0.02
DNK	0.54	0.02	0.58	0.02
ESP	0.49	0.01	0.50	0.01
EST	0.52	0.02	0.58	0.02
FIN	0.51	0.02	0.59	0.02
FRA	0.55	0.02	0.56	0.02
GBR	0.56	0.02	0.58	0.01
GRC	0.47	0.02	0.54	0.02
HKG	0.53	0.02	0.52	0.02
HRV	0.53	0.02	0.52	0.02
HUN	0.58	0.02	0.61	0.02
IDN	0.18	0.03	0.11	0.03
IRL	0.52	0.02	0.53	0.02
ISL	0.53	0.03	0.60	0.02
ISR	0.46	0.02	0.40	0.02
ITA	0.48	0.01	0.49	0.01
JOR	0.24	0.02	0.31	0.02
JPN	0.56	0.02	0.44	0.02
KAZ	0.29	0.02	0.33	0.02
KOR	0.63	0.02	0.62	0.02
LTU	0.51	0.02	0.59	0.02
LUX	0.49	0.02	0.51	0.02
LVA	0.48	0.03	0.57	0.03
MAC	0.50	0.02	0.49	0.02
MEX	0.30	0.01	0.41	0.01

Country	OLS_matheff		OLS_composite	
	Std. Regression Coef. (Corr. Est.)	SE	Std. Regression Coef. (Corr. Est.)	SE
MNE	0.28	0.03	0.34	0.03
MYS	0.35	0.02	0.37	0.03
NLD	0.45	0.03	0.37	0.03
NOR	0.61	0.02	0.68	0.02
NZL	0.57	0.02	0.58	0.02
PER	0.22	0.02	0.26	0.02
POL	0.63	0.02	0.69	0.02
PRT	0.61	0.02	0.63	0.02
QAT	0.24	0.02	0.16	0.02
ROU	0.32	0.03	0.35	0.03
RUS	0.47	0.02	0.54	0.02
SGP	0.54	0.02	0.53	0.02
SRB	0.39	0.03	0.47	0.03
SVK	0.51	0.02	0.55	0.02
SVN	0.49	0.02	0.52	0.02
SWE	0.51	0.02	0.61	0.02
TAP	0.65	0.02	0.65	0.02
THA	0.17	0.02	0.17	0.02
TUN	0.27	0.03	0.33	0.03
TUR	0.42	0.02	0.39	0.02
URY	0.31	0.03	0.42	0.02
USA	0.54	0.02	0.54	0.02
VNM	0.49	0.02	0.51	0.02

Table A.18. Parameter estimates from the MLM models employing MATHEFF only in Phase 6

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
ARE	-0.01	0.45	0.56	0.44	-0.01	0.24	0.01	0.40	0.52
ARG	-0.13	0.44	0.53	0.46	-0.13	0.18	0.02	0.45	0.50
AUS	0.07	0.26	0.70	0.27	0.05	0.51	0.01	0.14	0.51
AUT	-0.12	0.50	0.57	0.47	-0.09	0.35	0.02	0.34	0.48
BEL	-0.09	0.50	0.55	0.48	-0.08	0.34	0.01	0.40	0.46
BGR	-0.15	0.57	0.49	0.53	-0.14	0.14	0.02	0.54	0.48
BRA	-0.09	0.43	0.57	0.43	-0.09	0.19	0.01	0.40	0.54
CAN	0.05	0.21	0.80	0.21	0.04	0.52	0.01	0.13	0.59
CHE	0.01	0.32	0.63	0.34	0.02	0.45	0.02	0.19	0.48
CHL	-0.14	0.50	0.47	0.52	-0.13	0.22	0.02	0.44	0.43
COL	-0.10	0.35	0.62	0.36	-0.10	0.08	0.02	0.34	0.62
CRI	-0.03	0.42	0.62	0.40	-0.03	0.17	0.02	0.39	0.60
CZE	-0.15	0.50	0.45	0.52	-0.11	0.37	0.02	0.34	0.36
DEU	-0.05	0.49	0.52	0.48	-0.05	0.39	0.02	0.34	0.42
DNK	0.11	0.16	0.73	0.18	0.11	0.53	0.02	0.10	0.50
ESP	-0.07	0.18	0.77	0.19	-0.06	0.44	0.01	0.13	0.60
EST	-0.03	0.21	0.80	0.21	-0.01	0.48	0.02	0.14	0.61
FIN	0.13	0.07	0.82	0.07	0.13	0.51	0.02	0.04	0.58
FRA	-0.11	0.58	0.46	0.56	-0.09	0.37	0.02	0.40	0.36
GBR	0.01	0.24	0.78	0.24	0.00	0.50	0.01	0.16	0.56
GRC	-0.09	0.40	0.66	0.37	-0.07	0.38	0.02	0.28	0.54
HKG	-0.02	0.44	0.57	0.43	-0.02	0.39	0.02	0.29	0.46
HRV	-0.02	0.42	0.58	0.42	-0.02	0.38	0.02	0.27	0.47
HUN	-0.22	0.73	0.37	0.66	-0.18	0.33	0.02	0.51	0.30
IDN	-0.04	0.48	0.49	0.50	-0.04	0.12	0.02	0.47	0.48
IRL	-0.01	0.19	0.80	0.19	-0.01	0.48	0.02	0.12	0.60

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
ISL	-0.04	0.09	0.92	0.08	-0.02	0.52	0.03	0.05	0.68
ISR	-0.06	0.41	0.65	0.39	-0.05	0.40	0.02	0.33	0.50
ITA	-0.13	0.52	0.49	0.51	-0.11	0.32	0.01	0.40	0.42
JOR	-0.02	0.32	0.68	0.32	-0.02	0.21	0.02	0.31	0.64
JPN	-0.01	0.52	0.49	0.52	-0.01	0.33	0.02	0.32	0.42
KAZ	0.04	0.39	0.62	0.39	0.04	0.22	0.02	0.36	0.58
KOR	-0.03	0.38	0.63	0.38	-0.02	0.51	0.02	0.16	0.46
LTU	-0.08	0.32	0.69	0.32	-0.06	0.46	0.02	0.23	0.51
LUX	-0.02	0.33	0.69	0.33	-0.02	0.38	0.02	0.22	0.56
LVA	-0.12	0.29	0.72	0.29	-0.11	0.44	0.03	0.24	0.55
MAC	-0.17	0.42	0.70	0.38	-0.14	0.40	0.02	0.27	0.55
MEX	-0.08	0.35	0.64	0.35	-0.08	0.24	0.01	0.31	0.59
MNE	-0.06	0.31	0.67	0.32	-0.07	0.21	0.02	0.31	0.63
MYS	-0.01	0.32	0.69	0.32	-0.01	0.30	0.02	0.27	0.61
NLD	-0.01	0.62	0.36	0.63	-0.01	0.26	0.02	0.51	0.31
NOR	0.00	0.15	0.87	0.15	0.00	0.60	0.02	0.08	0.56
NZL	-0.04	0.21	0.80	0.21	-0.03	0.53	0.02	0.13	0.57
PER	-0.12	0.46	0.58	0.44	-0.12	0.19	0.02	0.45	0.54
POL	-0.01	0.22	0.79	0.22	-0.01	0.59	0.02	0.10	0.51
PRT	-0.03	0.28	0.72	0.28	-0.02	0.54	0.02	0.13	0.49
QAT	-0.05	0.45	0.53	0.46	-0.06	0.16	0.02	0.43	0.51
ROU	-0.01	0.50	0.56	0.47	-0.02	0.20	0.02	0.43	0.54
RUS	-0.05	0.29	0.73	0.28	-0.04	0.41	0.02	0.21	0.60
SGP	-0.01	0.36	0.65	0.35	-0.01	0.40	0.02	0.21	0.53
SRB	-0.03	0.47	0.54	0.47	-0.03	0.27	0.02	0.40	0.48
SVK	-0.12	0.44	0.54	0.45	-0.09	0.39	0.02	0.30	0.43
SVN	-0.10	0.51	0.48	0.52	-0.09	0.29	0.02	0.40	0.42
SWE	0.00	0.13	0.89	0.13	0.00	0.50	0.02	0.08	0.67

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
TAP	-0.03	0.43	0.60	0.42	-0.01	0.50	0.02	0.20	0.42
THA	-0.07	0.40	0.46	0.46	-0.06	0.13	0.02	0.38	0.45
TUN	-0.03	0.49	0.52	0.48	-0.04	0.16	0.02	0.45	0.51
TUR	-0.11	0.59	0.36	0.62	-0.11	0.18	0.02	0.51	0.34
URY	-0.10	0.44	0.60	0.42	-0.09	0.24	0.02	0.40	0.55
USA	0.00	0.22	0.78	0.22	0.00	0.49	0.02	0.14	0.57
VNM	-0.08	0.57	0.50	0.53	-0.08	0.34	0.02	0.44	0.41

Table A.19. Parameter estimates from the MLM models employing the composite measure in Phase 6

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
ARE	-0.01	0.45	0.56	0.44	-0.01	0.34	0.01	0.38	0.46
ARG	-0.13	0.44	0.53	0.46	-0.13	0.25	0.02	0.46	0.47
AUS	0.07	0.26	0.70	0.27	0.05	0.52	0.01	0.17	0.48
AUT	-0.12	0.50	0.57	0.47	-0.09	0.42	0.02	0.38	0.42
BEL	-0.09	0.50	0.55	0.48	-0.08	0.36	0.01	0.43	0.43
BGR	-0.15	0.57	0.49	0.53	-0.14	0.27	0.02	0.51	0.43
BRA	-0.09	0.43	0.57	0.43	-0.10	0.29	0.01	0.41	0.49
CAN	0.05	0.21	0.80	0.21	0.05	0.56	0.01	0.14	0.53
CHE	0.01	0.32	0.63	0.34	0.01	0.43	0.01	0.25	0.47
CHL	-0.14	0.50	0.47	0.52	-0.12	0.35	0.02	0.41	0.37
COL	-0.10	0.35	0.62	0.36	-0.11	0.24	0.02	0.32	0.57
CRI	-0.03	0.42	0.62	0.40	-0.03	0.28	0.02	0.37	0.55
CZE	-0.15	0.50	0.45	0.52	-0.12	0.43	0.02	0.36	0.31
DEU	-0.05	0.49	0.52	0.48	-0.05	0.43	0.02	0.41	0.36
DNK	0.11	0.16	0.73	0.18	0.10	0.56	0.02	0.12	0.44
ESP	-0.07	0.18	0.77	0.19	-0.06	0.47	0.01	0.13	0.56
EST	-0.03	0.21	0.80	0.21	-0.02	0.55	0.02	0.15	0.54
FIN	0.13	0.07	0.82	0.07	0.12	0.59	0.02	0.05	0.49
FRA	-0.11	0.58	0.46	0.56	-0.10	0.42	0.02	0.43	0.31
GBR	0.01	0.24	0.78	0.24	0.00	0.52	0.01	0.17	0.54
GRC	-0.09	0.40	0.66	0.37	-0.06	0.46	0.02	0.26	0.48
HKG	-0.02	0.44	0.57	0.43	-0.02	0.41	0.02	0.32	0.43
HRV	-0.02	0.42	0.58	0.42	-0.02	0.40	0.02	0.31	0.44
HUN	-0.22	0.73	0.37	0.66	-0.18	0.39	0.02	0.49	0.26
IDN	-0.04	0.48	0.49	0.50	-0.04	0.11	0.02	0.48	0.48
IRL	-0.01	0.19	0.80	0.19	-0.02	0.50	0.02	0.14	0.57

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
ISL	-0.04	0.09	0.92	0.08	-0.02	0.60	0.02	0.07	0.58
ISR	-0.06	0.41	0.65	0.39	-0.05	0.35	0.02	0.36	0.53
ITA	-0.13	0.52	0.49	0.51	-0.11	0.37	0.01	0.42	0.38
JOR	-0.02	0.32	0.68	0.32	-0.02	0.30	0.02	0.31	0.60
JPN	-0.01	0.52	0.49	0.52	-0.01	0.32	0.02	0.43	0.40
KAZ	0.04	0.39	0.62	0.39	0.04	0.30	0.02	0.38	0.54
KOR	-0.03	0.38	0.63	0.38	-0.02	0.49	0.02	0.19	0.45
LTU	-0.08	0.32	0.69	0.32	-0.06	0.53	0.02	0.21	0.45
LUX	-0.02	0.33	0.69	0.33	-0.03	0.42	0.02	0.23	0.53
LVA	-0.12	0.29	0.72	0.29	-0.12	0.55	0.02	0.25	0.46
MAC	-0.17	0.42	0.70	0.38	-0.16	0.42	0.02	0.32	0.53
MEX	-0.08	0.35	0.64	0.35	-0.08	0.36	0.01	0.30	0.53
MNE	-0.06	0.31	0.67	0.32	-0.07	0.28	0.02	0.31	0.59
MYS	-0.01	0.32	0.69	0.32	-0.01	0.34	0.02	0.29	0.59
NLD	-0.01	0.62	0.36	0.63	-0.01	0.27	0.02	0.56	0.30
NOR	0.00	0.15	0.87	0.15	0.00	0.66	0.02	0.09	0.47
NZL	-0.04	0.21	0.80	0.21	-0.03	0.54	0.02	0.14	0.54
PER	-0.12	0.46	0.58	0.44	-0.13	0.29	0.02	0.49	0.49
POL	-0.01	0.22	0.79	0.22	-0.01	0.65	0.02	0.12	0.41
PRT	-0.03	0.28	0.72	0.28	-0.02	0.57	0.02	0.15	0.45
QAT	-0.05	0.45	0.53	0.46	-0.06	0.12	0.02	0.45	0.52
ROU	-0.01	0.50	0.56	0.47	-0.02	0.24	0.02	0.42	0.52
RUS	-0.05	0.29	0.73	0.28	-0.05	0.50	0.02	0.22	0.52
SGP	-0.01	0.36	0.65	0.35	-0.01	0.42	0.02	0.23	0.51
SRB	-0.03	0.47	0.54	0.47	-0.03	0.36	0.02	0.38	0.43
SVK	-0.12	0.44	0.54	0.45	-0.10	0.46	0.02	0.33	0.36
SVN	-0.10	0.51	0.48	0.52	-0.09	0.38	0.02	0.41	0.35
SWE	0.00	0.13	0.89	0.13	0.00	0.59	0.02	0.09	0.55

Country	Unconditional Model				Conditional Model				
	Intercept	τ_{00}	σ^2	ICC	Intercept	Std. Regression Coef. (Corr. Est.)	SE	τ_{00}	σ^2
TAP	-0.03	0.43	0.60	0.42	-0.02	0.51	0.02	0.22	0.39
THA	-0.07	0.40	0.46	0.46	-0.07	0.17	0.02	0.39	0.44
TUN	-0.03	0.49	0.52	0.48	-0.04	0.21	0.02	0.43	0.49
TUR	-0.11	0.59	0.36	0.62	-0.11	0.21	0.02	0.52	0.33
URY	-0.10	0.44	0.60	0.42	-0.09	0.35	0.02	0.38	0.49
USA	0.00	0.22	0.78	0.22	0.00	0.51	0.02	0.15	0.54
VNM	-0.08	0.57	0.50	0.53	-0.08	0.36	0.02	0.44	0.40