Title
Multilevel Diagnostic Item Response Model for School Selection and Assessment

Author
Langi, Meredith Lindsay

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Multilevel Diagnostic Item Response Model for School Selection and Assessment

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Education

by

Meredith Lindsay Langi

2021

ABSTRACT OF THE DISSERTATION

Multilevel Diagnostic Item Response Model for School Selection and Assessment

by

Meredith Lindsay Langi
Doctor of Philosophy in Education
University of California, Los Angeles, 2021
Professor Minjeong Jeon, Chair

Performance-targeted interventions, typically based on student test performance, are an important tool in improving educational outcomes. These types of interventions are often applied at the school level, where low-performing schools are selected for participation. However, typical school effects methods for understanding school performance do not directly identify the low-performing schools that would benefit the most from additional support. Additionally, typical school effects methods do not differentiate school performance based on important aspects of the curriculum. This dissertation fills this gap in school effects methods by proposing the Multilevel Diagnostic Item Response (MD-IR) model. The MD-IR model is a multilevel, confirmatory mixture item response theory model that incorporates strategic constraints in order to differentiate schools, and students within schools, based on the aspects of the curriculum that would be most relevant for a performance-targeted intervention. By incorporating latent classes, the MD-IR model classifies schools as high- or low-performing, and as such, identifies schools most in need of support. The formulation of the MD-IR model is presented, along with a detailed empirical example demonstrating its application in the context of international educational development using data from PISA for Development. Results from the empirical example illustrate the utility of this model and its promise in filling this identification gap in the school effects literature.

The dissertation of Meredith Lindsay Langi is approved.

Mark Hansen

Chad Hazlett

Mike Seltzer

Minjeong Jeon, Committee Chair

University of California, Los Angeles

2021

To my daughter, Salote.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I'd like to thank my two advisers, Dr. Minjeong Jeon and Dr. Mark Hansen. Thank you both for your constant support and guidance throughout the entire PhD process, and for challenging me to think more deeply about methods and measurement. It has been an honor to work with both of you. I'd also like to thank Dr. Chad Hazlett and Dr. Mike Seltzer for serving as committee members.You have all been influential in my academic career and I am so grateful to have had the opportunity to learn from you.

I am also grateful to the SRM faculty for being thoughtful and supportive mentors at various stages of the program, and to my fellow SRM students for helping to create such a supportive community.

I also need to acknowledge the amazing copy editing skills of my mother, who has ensured that this document is truly readable. Thank you Mom, for the time you took to help get me over the finish line!

Finally, I am most grateful to my husband who has encouraged and supported me throughout my entire graduate school journey, no matter how long it all took. Thank you Sione, and Salote, for always reminding me what is most important.

VITA

EDUCATION

2007	Bachelor of Arts, Liberal Studies Concordia University, Irvine
2014	Master of Education, International Education Policy Harvard Graduate School of Education

WORK

2016-2020	Graduate Student Researcher National Center for Research on Evaluation, Standards, and Student Testing (CRESST),University of California, Los Angeles
2017-2020	Graduate Student Researcher Education Leadership Program University of California, Los Angeles
2016-2017	Graduate Student Researcher Professor Felipe Martinez University of California, Los Angeles
2014-2015	Research Analyst Education Accountability Project Harvard Graduate School of Education

PUBLICATIONS

Koretz, D., & Langi, M. (2018). Predicting Freshman Grade?Point Average from Test Scores: Effects of Variation Within and Between High Schools. *Educational Measure-*

ment: Issues and Practice, 37(2), 9-19.

Koretz, D., Yu, C., Mbekeani, P. P., Langi, M., Dhaliwal, T., & Braslow, D. (2016). Predicting Freshman Grade Point Average From College Admissions Test Scores and State High School Test Scores. *AERA Open*, 2(4).

Koretz, D., Jennings, J. L., Ng, H. L., Yu, C., Braslow, D., & Langi, M. (2016). Auditing for Score Inflation Using Self-Monitoring Assessments: Findings From Three Pilot Studies. *Educational Assessment*, 21(4), 231-247.

CHAPTER 1

Introduction

In 2015, after over a decade of worldwide commitment to enroll all students in schools, the United Nations and education leaders shifted the focus from enrollment, or quantity in education, to quality in education. Broadly, the UN shifted focus from the Millennium Development Goals to the Sustainable Development Goals (SDGs), with goal number 4 focused on education. This goal aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (United Nations, 2019). Notably, this goal puts a great emphasis on equity across all populations. The education community, with support from the UN and the World Bank, responded with the Incheon Declaration, which commits to "addressing all forms of exclusion and marginalization, disparities and inequalities in access, participation and learning outcomes" (UNESCO, 2015).

To address issues in equity, many efforts have focused on schools as the target level for interventions as this is an area in which governments have more control (Anderson, Milford, & Ross, 2009). The approach used to identify appropriate schools depends on the type of intervention or policy under consideration. Willms (2006) outlines five different approaches to targeting an intervention: (1) universal interventions, (2) SES-targeted interventions, (3) compensatory interventions, (4) performance-targeted interventions, and (5) inclusive interventions. Universal interventions are implemented for all schools, and an identification strategy is not necessary. Inclusive, compensatory, and SES-targeted interventions are determined by more observable categories, (such as community income level or students' disability status), although it should be noted that there are often complexities in measuring these as well.

Performance-targeted interventions, however, present a particularly challenging identification process as they are reliant on student test performance. Selecting schools to target requires creating categories based on performance of students and the impact schools have on that performance. The decisions required for creating these school categories should depend on the goal of the intervention as the need for categories can vary across different political, realistic, and value perspectives (Mehrans & Cizek, 2012). Whatever the goals of the intervention, the identification process should select schools by performance-targeted criteria that best match that of the intended intervention. For example, one performance-targeted intervention that is particularly common in international education is teacher professional development and training (e.g., Lucas, McEwan, Ngware, & Oketch, 2014). Education experts emphasize that professional development for teachers in targeted schools should link student learning to teacher practice and to do so, an emphasis should be placed on improving curricular development and instruction (e.g., Darling-Hammond, 2017). In this professional development intervention, the link to the curricular and instructional areas that are most in need of improvement should be included in the school identification process, such that educators participating are from schools that are most in need of improvement on these particular areas. In other words, the performance-targeted identification procedure should include not only targeted approaches for identifying schools, but also be based on targeted definitions of performance.

1.1 School Effects

Targeting schools for interventions is not only a matter of convenience for governments, but it can also lead to reduction of inequalities in learning outcomes. Figure 1.1 shows a framework that illustrates the complex relationship between student outcomes and schools. This framework was first proposed by Willms (2010) and has been adopted by PISA (OECD, 2010). It highlights heterogeneity in multiple prosperity outcomes by defining different subpopulations, such as gender or poverty level, and making a

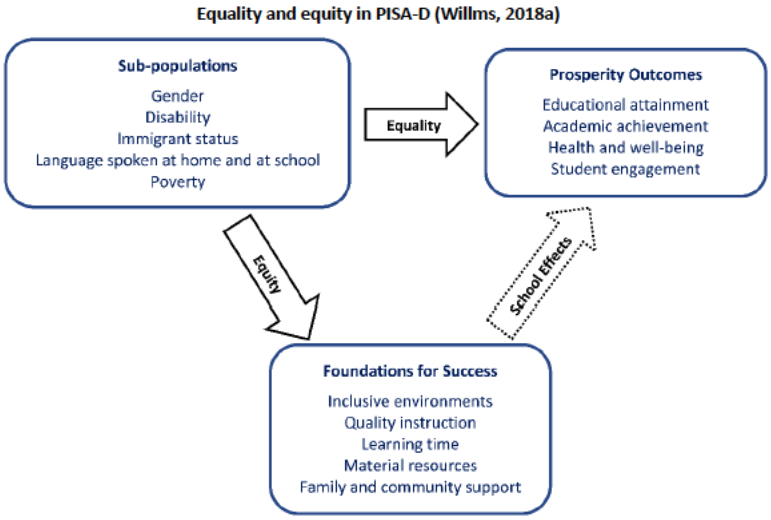

Figure 1.1: Willms (2010) and PISA framework for the relationship between subpopulations, outcomes, and resources.

key distinction between equality and equity. Equality is defined as differences (or lack thereof) in prosperity outcomes. For example, researchers may be interested in measuring differences in math performance by gender or other background characteristics. Equity is defined as differences in the foundations for success, or as differences in the resources needed to be successful on the prosperity outcomes. Studying equity, for example, involves studying the relationship between gender, or other characteristics, and the foundations for success that are necessary for achievement in mathematics.

In this framework, school effects are captured in the arrow between the foundations for success and the prosperity outcomes. The placement of this arrow indicates the belief that schools have a direct impact on prosperity outcomes. The definition of the term *school effects* used in the PISA framework is "the effect on academic performance of attending one school or another" (OECD, 2013). This effect is influenced by foundations of success that include student dimensions (e.g., family support), teacher dimensions (e.g., instructional quality), school dimensions (e.g., inclusive environments), and community dimensions (e.g., community support). The different types of interventions outlined by Willms (2006) all have important roles to play in influencing the different aspects of this framework and

improving outcomes for students. Performance-targeted interventions are particularly useful in the areas of teacher and school dimensions relating to instruction, curricular materials, and environments. Appropriately identifying schools for interventions in these areas can improve a schools' ability to impact student learning.

The term *school effects* can be controversial in education research. Debates are not centered around whether schools impact students, but on how to measure and define this impact. When a definition is agreed upon, policy debates continue as to whether or not it is appropriate to utilize school effects measures when making high-stakes decisions. Further, it becomes unclear whether that which is being measured is actually the school effect on the construct of interest, or some irrelevant variance, such as cheating or inflation (Koretz, 2017). Despite the controversy, there remains interest in understanding the impact of schools on student learning, as is demonstrated in the United States ESSA policies and international development work (e.g., Al-Samarrai et al., 2017; Andrabi, Das, & Khwaja, 2015; DuFour, 2004). More recent emphasis in school effects research and policies focuses on how to incorporate feedback and school improvement in the measurement and resulting classification of schools (e.g., CCSSO, 2017).

To summarize, measuring the impact of schools on student learning is an important step in implementing interventions needed to achieve the goals set out by the UN and the World Bank. The approach taken to measure the school effects and identify schools for interventions must match the goal of the intended intervention. In other words, the categorization procedure must take into account student and school performance on the curricular and instructional areas that are most in need of improvement and that are the aim of the performance-targeted intervention.

1.2 Current Approaches for Measuring School Effects

In recent years, thanks to advances in multilevel modeling, methods for measuring school effects have become more sophisticated. Despite these advances, the current methods

are not designed for classifying schools as they require additional steps to make these categorizations. Additionally, current school assessment approaches do not incorporate detailed curricular and instructional information that is needed to link school assessments to interventions. These limitations make it difficult to use the current methods for the use of quality performance-targeted interventions that could be useful in closing gaps in student performance. A full review of current methods and their limitations is provided in Chapter 2.

1.3 Research Aims

This research proposes a psychometric model that addresses the limitations in current approaches. The proposed Multilevel Diagnostic Item Response model emphasizes selection and identification of schools, while also incorporating relevant advances in psychometric modeling, (which are reviewed in Chapter 3). Specifically, this model is a multilevel item response model that incorporates mixture distributions at both student and school levels. It is a confirmatory model in that it utilizes existing theories for defining the latent groups prior to model estimation. This makes it useful in the context of school effectiveness because both students and schools can be classified into high- and low-performing groups, and then are appropriately selected for performance-targeted interventions. The model also enhances existing methods by incorporating additional information that links the assessment items to curricular domains. By taking advantage of this item information, classifications are then based on student performance across these different domains. The full introduction to the new model is given in Chapter 4.

1.4 Example: PISA for Development

A running example is incorporated throughout the following chapters in order to more clearly illustrate the benefits of the proposed model. It is based on a new PISA initiative, PISA for Development (PISA-D), that assesses student knowledge in low-income countries.

The goal of this example is to demonstrate the Multilevel Diagnostic Item Response (MD-IR) model in the context of international development. This example specifically looks at mathematical literacy and differences in school effects and student performance across key processes involved in mathematical problem solving. PISA has made efforts to ensure that the definition of mathematical literacy is nuanced and relevant for instructional practices (OECD, 2019). This type of nuanced information can provide education leaders with key data on student and school math performance that is useful for designing and implementing performance-targeted interventions, as well as for selecting schools with the highest need for participation in these interventions. Since resources are often particularly limited in low-income countries, this targeted selection is essential. While the PISA-D assessment is not designed for identifying individual schools (due to the sampling procedures), the running example provides as strong illustration of how the MD-IR model can provide information for educational planning. Hypothetical illustrations for identifying individual schools based on the MD-IR model are also included in this dissertation, with the recognition that the data are not designed for this purpose.

1.4.1 PISA-D Background

The PISA for Development initiative was designed to provide a more accessible test for countries that expect to perform poorly on the typical PISA assessment. Like the regular PISA, this assessment was administered to 15 year-old students. The participating countries include: Ecuador, Guatemala, Honduras, Cambodia, Paraguay, Senegal, and Zambia. Three subjects are assessed in each country: math, reading, and science. The PISA-D assessment was balanced between these three subjects, as opposed to the regular PISA, which assesses one main subject and two minor subjects in each cycle (OECD, 2019). Student and school background questionnaires are also included. The student questionnaire asks questions regarding the students' home life, school history, and physical and mental health. School questionnaires focus primarily on school climate and community context. This rich set of data allows for exploring the complexity of school

Table 1.1: School resource statistics for six countries participating in PISA-D

	Ecuador	Guatemala	Honduras	Cambodia	Paraguay	Senegal
N of Students	3715	3332	3131	3225	2837	3174
N of Schools	172	190	201	168	204	159
School Resources						
Extremely Low	12	24	36	47	15	46
Severely Low	26	30	35	40	41	36
Low	36	37	40	37	47	30
Moderate	43	39	22	32	58	33
High	55	58	62	12	40	14

effects discussed in Figure 1.1. Students and schools are sampled through a complex two-phase sampling procedure.

1.4.2 School Resource Data

Table 1.1 presents school resource statistics and sample sizes in six of the seven countries, (Zambia was excluded.) Sample sizes are similar across countries, with around 3,000 students and at least 160 schools in all countries. School context is different for schools within country, particularly in terms of the amount of available resources. The empirical example in this dissertation will focus on Cambodia. More details on the sample are available in Chapter 5, Section 5.2.3.

1.4.3 PISA-D and School Effects

Large-scale assessments such as PISA and TIMSS (Trends in International Mathematics and Science Study) have been used in order to study school effects in low-income countries. A large study across 25 economically diverse countries finds that schools matter in countries of all economic levels, but particularly so in poor countries and countries with large economic inequalities (Chudgar & Luschei, 2009). Prior World

Bank research comparing school effects in high- and low-income countries found that in low-income countries, schools may play an even more important role than family characteristics (Heyneman & Loxley, 1983). While this finding is debated (see for example, Huang (2010) in the Philippines and Bouhlila (2015) in the Middle East and North Africa,) school effects clearly play and important role in understanding student performance in low-income countries. As a new initiative, PISA-D and the broader expansion of large-scale international assessments to include low-income countries has the potential to expand research on school effects in these countries. The MD-IR example in this dissertation capitalizes on this opportunity by exploring differences in school effects within Cambodia.

1.4.4 Mathematics Literacy and Assessment

The PISA for Development mathematics assessment aims to measure students' mathematics literacy. PISA defines mathematical literacy as "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" (OECD, 2018, p. 51). The focus of this definition is on engaging with mathematics, particularly through the key verbs "formulate", "employ", and "interpret". These three verbs are considered the three processes that students engage in to be problem solvers in math (OECD, 2018). The formulating process refers to how well students are able to formulate a stated problem in a mathematical form. The employing process refers to the students' ability to perform the appropriate computations and manipulations to arrive at the correct solution. The interpreting process refers to how well the students can interpret the solution in the real-world context. In PISA-D, items are linked to each of these three processes, with 50% of items being employ items, 25% formulate items, and 25% interpret items. Each item is assigned to only one category.

True mathematical literacy requires the ability to utilize all three of these mathematical processes and their underlying competencies. However, students throughout the world

tend to have difficulties with various aspects of these processes. Field testing of PISA-D assessment items revealed differences in student performance by these processes, with formulating items being the most difficult (Stacey, 2015). Similarly, students in Indonesia struggled the most with formulate items, showing consistent errors in this domain (Kohar, Zulkardi, & Darmawijoyo, 2014). Even countries that typically perform highly on the general PISA math assessment, such as Finland and South Korea, struggle with various aspects of these processes (Turner, Blum, & Niss, 2015).

One possible explanation for the struggles students face is the fact that teachers themselves have deficiencies in explaining the problem solving process (Sáenz, 2009). Güler and Arslan (2019) find that many teachers could solve math problems, but were unaware of the processes and competencies they utilized in the process. Teaching the full mathematical modeling process is essential for student high-performance and therefore, efforts should be made to improve the teaching of the processes involved in solving real world math problems (Blum, 2015). The MD-IR example presented here explores differences in school performance across these mathematical processes in Cambodia.

1.5 Conclusion

The empirical example presented in Chapter 5 demonstrates how the proposed model brings together the study of school effects, mathematics literacy, and performance-targeted interventions for the purposes of international educational development. The unique combination of psychometric modeling approaches used in the MD-IR makes it a useful new tool in implementing interventions that can support countries as they strive to achieve the goal of equitable outcomes and learning opportunities for all.

CHAPTER 2

School Effects Methods Literature Review

Methods for studying the impact of schools on student learning vary widely across many different disciplines. In this review, I focus on statistical models that utilize multilevel modeling techniques, which are popular in the field of education research.

2.1 Hierarchical Linear Models

2.1.1 Basic Multilevel Model

Perhaps the most common way of measuring school effects is through two-level regression modeling, which is often labeled as hierarchical linear modeling (Raudenbush & Bryk, 2002), multilevel models, or mixed effects models, depending on the field. In this dissertation I refer to these models as hierarchical liner models (HLM) and present this method in detail as it forms the basis for many of the other approaches reviewed. I first present the unconditional two-level model, then discuss how it can be adapted for modeling the complexity of school effects.

The simplest two-level model as outlined in Raudenbush and Bryk (2002) is the unconditional model, where level one is modeled as,

$$Y_{jm} = \beta_{0m} + r_{jm}, \tag{2.1}$$

and level two as,

$$\beta_{0m} = \gamma_{00} + u_{0m}. \tag{2.2}$$

In school effects studies, Y_{jm} is typically a score for student j, who attends school m, on an achievement test (e.g., math). The intercept in Equation 2.1, β_{0m}, represents the mean score for school m. The level one error term is represented by r_{jm} and is normally distributed with mean zero and a variance of σ^2 that remains constant. In this model, student-level scores are based solely on the mean score of the school the student attends. Figure 2.1 shows the within-school distribution of student residuals (the lower row of distributions) for school M. Student j in school M has a positive residual, so that within school M, student j performs higher than the school's average.

In Equation 2.2, γ_{00} represents the grand mean of student scores and u_{0m} is the school-level random effect associated with school m. The random effect is assumed to have a mean of zero and a constant variance ϵ^2. To make the interpretation of u_{0m} clear, substitute Equation 2.2 into Equation 2.1 to get the combined model,

$$Y_{jm} = \gamma_{00} + u_{0m} + r_{jm}. \tag{2.3}$$

Equation 2.3 shows that student j's score can be decomposed into the grand mean of student scores, the unique contribution of school m that is the school's effect, and the student's residual. Figure 2.1 shows the distribution of school effects u_0 for all schools (the upper distribution). School L is located at the bottom of the distribution, such that the academic performance of students in school L will be lower, on average, compared to the average student score. Students in school M at the top of the distribution are expected to perform higher, on average, compared to the average student score. In this unconditional model, the school effect u_{0m} captures all aspects of the school's performance, including the impact of various school practices and the impact of the school's context. Disentangling school practices from context is discussed in the following section.

The variance components are an important aspect of the model in school effectiveness

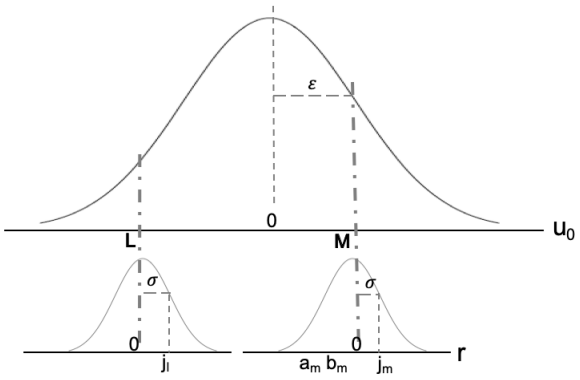

Figure 2.1: Distribution of school effects u_0 and within-school distributions of student residuals r.

studies. The between-school variance, ϵ^2, signals the extent to which schools influence student outcomes. If the variance is large, schools have a substantial impact on student outcomes. On the other hand, if the variance is small, the schools play a less important role (Raudenbush & Bryk, 2002). The within-school variance, σ^2, indicates the size of gaps within schools. A larger within-school variance suggests larger achievement gaps between students within schools. In this model, it is assumed that the within-school variance is equal for all schools. Figure 2.1 shows how the variance is equal for both school L and school M.

Studies of school effects typically include covariates at one or both levels of analysis. Raudenbush and Bryk (2002) emphasize that including student-level background variables improves the accuracy and precision of school effects estimates by reducing the level one error variance, r_{jm}. Including covariates also accounts for student and school characteristics, and as such, the school effect term, u_0, can be interpreted as an adjusted measure of performance that captures the school-level variation that is not accounted for by the covariates in the model. This means that the interpretation of the school effect term depends on the covariates included in the model (Grilli & Rampichini, 2009).

Adding covariates to Equation 2.1 yields,

$$Y_{jm} = \beta_{0m} + \beta_{1m}X_{1jm} + ... + \beta_{Qm}X_{Qjm} + r_{jm}, \tag{2.4}$$

where X_{qjm} represents the qth student background characteristic. Modeling the additional β coefficients gives,

$$\begin{aligned} \beta_{0m} &= \gamma_{00} + u_{0m} \\ \beta_{1m} &= \gamma_{10} \\ &\vdots \\ \beta_{Qm} &= \gamma_{Q0}. \end{aligned} \tag{2.5}$$

In Equation 2.5, the student-level intercept, β_{0m}, is allowed to vary across schools, just as in Equation 2.2. This model is the random-intercept model where the other regression slopes are assumed to be equal across schools. In other words, a school's effectiveness is captured in u_{0m}, which does not vary by the student characteristics that are included in the model and follows the same assumptions as in Equation 2.2. This assumption of equal regression slopes across schools can be relaxed, which will be discussed in a following section.

School-level covariates can also be added to the level two models,

$$\begin{aligned} \beta_{0m} &= \gamma_{00} + \gamma_{01}W_{1m} + ... + \gamma_{Sm}W_{Sm} + u_{0m} \\ \beta_{1m} &= \gamma_{10} \\ &\vdots \\ \beta_{Qm} &= \gamma_{Q0}. \end{aligned} \tag{2.6}$$

In Equation 2.6, the student-level intercept is no longer modeled simply as a function of the grand mean of student scores and a school-specific deviation from that mean. Instead, that intercept is modeled as a function of the grand mean, as well as specific school-level characteristics, such as average student SES (socioeconomic status) and average student prior achievement. By including these school characteristics, u_{0m} becomes the school

effect adjusted for school context.

2.1.2 Modeling Type A and Type B Effects

To elaborate on the definition of school effects provided above, Raudenbush and Willms (1995) give a further explanation of this term by splitting school effects into two different categories: Type A and Type B effects. A Type A effect is the difference in a students's observed performance and the expected performance of that same student were she to attend an average school. This type of effect is what a parent might consider when selecting a school for their child. In other words, it does not attempt to isolate a school's performance and practice separately from other contextual effects such as the social and economic characteristics of the school's community.

A Type B effect is different in that it aims to isolate the impact of the school's practice on students' learning. Practice is considered school leadership, curriculum, instructional approaches and quality, and resource utilization and would be considered foundations for success in Figure 1.1. These practices are separate from school context (Raudenbush & Willms, 1995), although the relationship between the two is typically quite strong. A Type B effect is the difference in a student's observed performance in her school and her expected performance if she had attended an average school *within the same context*. School and district administrators are often interested in this type of school effect. In practice, as is clearly shown in Figure 1.1, there is a relationship between context and practice (i.e., they covary).

Estimation of Type A school effects is feasible because it is not necessary to isolate the school effect from the contextual effect. On the other hand, Type B effects are impossible to isolate without some level of bias (Castellano, Rabe-Hesketh, & Skrondal, 2014; Raudenbush & Willms, 1995). Similarly, consistent estimation of Type A variance components is feasible, while consistent estimates for Type B effects are much more difficult (Raudenbush & Willms, 1995).

Type A effects. Type A effects do not aim to isolate school context from the school effects

estimate. In terms of the model, practice is represented by u_{0m}, and context effects are represented by the γ coefficients and the W_{Sm} covariates,

$$A_m = \gamma_{01} W_{1m} + ... + \gamma_{Sm} W_{Sm} + u_{0m}. \tag{2.7}$$

In Equation 2.7, the Type A effect is consistent for all students within school m. To clarify this, I present an example with one covariate at each level. At the student-level is student SES, and at the school-level is the school aggregated SES. The outcome variable is students' scores on the PISA-D mathematical literacy assessment. Note that this score can be created either as a sum score (i.e., totaling the number of correct responses) or using item response theory. Only a score on the general domain is typically used, without any information on sub-domains or mathematical processes (formulate, employ, and interpret).

Writing both the student and school levels in the same equation yields the combined model,

$$Y_{jm} = \gamma_{00} + \beta_{SES} SES_{jm} + \gamma_{S\bar{E}S} S\bar{E}S_m + u_{0m} + r_{jm}, \tag{2.8}$$

where Y_{jm} is student j's (who attends school m) math score. $\gamma 00$ represents the grand mean of all math scores. SES_{jm} is a variable indicating the student's SES level, and β_{SES} is the slope associated with student SES that is constant across all students and schools. $S\bar{E}S$ is the school mean SES that represents the school's context and $\gamma_{S\bar{E}S}$ represents the contextual effect of aggregated SES on individual student scores. u_{0m} represents school m's unique effect on math achievement adjusted for student and school SES, and r_{jm} represents the student-level residual. Assumptions for u_{0m} and r_{jm} are the same as in earlier models. At this point, the model for the Type A effect for student j in school m is given as,

$$\begin{aligned} A_{jm} &= \gamma_{S\bar{E}S} S\bar{E}S_m + u_{0m} \\ A_{jm} &= \text{contextual effect} + \text{school effect.} \end{aligned} \tag{2.9}$$

Type A effects can be estimated without bias using maximum likelihood estimation (Raudenbush & Willms, 1995).

This form of school effect is often considered as being useful to parents who are selecting a school for their children since parents may want to consider the impact of a school's context on their child's scores. However, it is possible that some interventions are a combination of performance-targeted and SES-targeted interventions. For example, a program that focuses on providing resources to schools may want to consider a school's Type A effect because taking the socioeconomic context into account is important for resource distribution.

Type B effects. In contrast with Type A effect, Type B effects aim to isolate the impact of the school separately from the contextual effects, which presents additional challenges in estimation. Ideally, if it could be assumed that student characteristics, such as student SES, are unrelated to the school effect, the Type B effect for school m would be written as,

$$\begin{aligned} B_m &= u_{0m} \\ B_m &= \text{school effect.} \end{aligned} \tag{2.10}$$

In Equation 2.10, the school relevant factors beyond the school mean of achievement and the included covariates, are captured in u_{0m}, (Grilli & Rampichini, 2009). However, it is not likely that the school effect is unrelated to the student background characteristics, resulting in a biased and inconsistent estimate of Type B effects. Additionally, Raudenbush and Willms (1995) show in more detail that the maximum likelihood estimates of the Type B variance components will be underestimated, as the estimate will provide a lower bound of the Type B component.

Value-added models. A common interpretation of school effects study results comes in the form of *value-added*. Value-added studies aim to isolate the school's (or in some

cases, the teacher's) additional contribution to student learning beyond what would be expected for that student based on prior performance and other characteristics. In other words, a school's value-added for student i in school j written as a combined model is,

$$\begin{array}{rcccl} Y_{jm} & - & (\gamma_{00} + \beta X_{jm} + \gamma W_m) & = & u_{0m} + r_{jm} \\ \text{observed outcome} & - & \text{expected outcome} & = & \text{value added} + \text{residual.} \end{array} \tag{2.11}$$

Notation in Equation 2.11 is the same as in previous equations. When shifting the interpretation, the school effect term u_{0m} is referred to as the school's value-added, and a school with a positive value-added score is considered "effective". Whether, and to what extent, to include covariates that control for context effects is a considerable point of debate in the field of education accountability. Additionally, value-added measures have limited reliability over time, making their application in high-stakes accountability policies problematic. A full review of value-added approaches is beyond the scope of this review, but for useful discussions and critiques refer to Koretz (2008), McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004), and Reardon and Raudenbush (2009).

2.1.3 Differential Effectiveness

Up to this point, the models have assumed uniform school effects across all schools. However, this assumption may not always hold true. In the HLM framework, the random-slope model allows for exploring differential effectiveness based on observable student characteristics. Specifically, the additional Q number of β coefficients from Equation 2.4 can be set to vary across schools,

$$\begin{aligned}
\beta_{1m} &= \gamma_{10} + u_{1m} \\
&\vdots \\
\beta_{qm} &= \gamma_{q0} + u_{qm} \\
&\vdots \\
\beta_{Qm} &= \gamma_{Q0} + u_{Qm}.
\end{aligned} \tag{2.12}$$

γ_{q0} represents the average slope across schools for student-level covariate q. u_{qm} represents the unique difference in that slope associated with school m. The u_{qm} components are assumed to have a mean of zero, constant variance, and are allowed to covary.

To illustrate, Equation 2.8 is adjusted to include the student-level covariate for prior math achievement. Note that student and school-aggregate SES could have been left in the model, but for simplicity, only one covariate is included. Writing as a combined model gives,

$$Y_{jm} = \gamma_{00} + (\gamma_{math} + u_{1m})math_{jm} + \gamma_{m\bar{a}th}m\bar{a}th_m + u_{0m} + r_{jm}, \tag{2.13}$$

where γ_{math} represents the average slope on math achievement for all schools, and u_{1m} represents the school specific deviation from this average slope. Type A and B effects are then modeled as follows,

$$A_{jm} = \gamma_{S\bar{E}S}S\bar{E}S_m + u_{0m} + u_{1m}math_{jm} \tag{2.14}$$

$$B_m = u_{0m} + u_{1m}math_{jm}. \tag{2.15}$$

Using this approach, it is possible to model the fact that school effectiveness may vary based on student characteristics. However, utilizing the results from these models for ranking or targeted interventions is difficult because the regression lines for each school cross at different points (Grilli & Rampichini, 2009). In other words, the rankings for schools can change drastically depending on the location of the math achievement scale.

Additional school-level covariates can be included in Equation 2.12 in order to control for school characteristics in estimating differential effects.

2.1.4 Examples of HLM Studies

The HLM approach outlined above is extremely popular for studying school effects, and as such, there are a plethora of examples in the literature. The relatively small number of examples presented here demonstrates the wide-ranging use and flexibility of the HLM framework across a variety of contexts. Willms and Somer (2001) provide an interesting example of using HLM to study school effects across numerous countries in Latin America. The authors demonstrate how two- and three-levels can be used to study school, classroom, and family effects, both within and between countries. Odden, Borman, and Fermanich (2004) use HLM to include less common school-level covariates in order to understand fiscal effects of different levels of investment in education.

Studies have also used primarily random-slope models to look at differential effectiveness, especially by gender and SES. Opdenakker and Van Damme (2006) use three-level models, with students nested in classrooms, which are in nested in schools, in order to look at differences in context and practice in public and Catholic schools. Strand (2010) explores differential effectiveness by ethnicity, gender, poverty, and prior achievement in the U.K. Kyriakides, Creemers, and Charalambous (2019) use random-slope models to emphasize the connection between differential effectiveness and the relationship between quality and equity in education. van Hek, Kraaykamp, and Pelzer (2018) analyzed data from PISA 2009 in OECD countries to determine whether school resources had differential effects on males and females. This study used cross-level interactions, an approach not outlined above, in order to study differential effectiveness. This allowed the authors to look at differential Type A effects, but not differential Type B effects.

2.1.5 Limitations

Clearly, the HLM framework is quite flexible and useful for studying the impact of schools on student learning, but some important limitations do exist. First, it is difficult to classify schools in need of a particular performance-based intervention. To create classifications, schools are often ranked based on their school effect or value-added score. Then, a certain percentage, such as the bottom 25% of schools in this ranking, might be selected for intervention. However, these rankings are typically unreliable and highly sensitive to the specification of the model (McCaffrey et al., 2004).

Second, most HLM studies use scores on a single dimension such as the general "math" dimension for the analysis. As such, less nuanced and actionable feedback is provided. In order to make use of the more nuanced information regarding sub-dimensions and cognitive processes, subsequent analyses beyond the single dimension must be performed.

A third limitation to the HLM approach is the need for the process to be conducted in two steps. The first step involves estimating student scores based on student responses to test items. This can be done in a number of ways, and the approach taken involves differing sets of limitations. If sum scores are used, school effect scores may be biased as a result of failing to address measurement error that is embedded in observed responses (Fox, 2004; Fox & Glas, 2001). Often a more modern approach is taken, and scores are estimated using IRT models. Scores are typically estimated without accounting for school clustering, which can lead to the attenuation of the relationship between the school effect and the observed variables (Mislevy, 1984; Pastor, 2003). In response to these concerns, researchers have recommended incorporating the IRT model as a measurement model into the multilevel framework. In doing so, latent ability scores (as opposed to sum scores or total scores) can be utilized to produce more accurate estimates of school effects (Fox & Glas, 2001). This is the approach discussed in the next section.

2.2 Multilevel Item Response Theory Modeling

2.2.1 The IRT Measurement Model

Introducing a measurement model based on IRT into the HLM framework is often referred to as Multilevel Item Response Theory (IRT) modeling. Broadly, models in the IRT framework describe the relationship between a pattern of item responses and the student's ability, taking into account the characteristics of test items (Fox, 2004). Multilevel IRT has been formulated for the inclusion of covariates in two primary ways across the literature (Kamata & Vaughn, 2011). The first formulation presented by Kamata (2001) uses hierarchical generalized linear models to present IRT as a two-level logistic regression model with items nested in students, which can then be generalized to additional levels to account for student nesting in schools. A second formulation presented by Fox and Glas (2001) also considers items nested within students. The formulation of multilevel IRT for school effects that is presented below is based on the Fox and Glas framework.

The multilevel IRT model incorporates a latent variable based on student responses to items in order to estimate student latent ability. Using the PISA-D example, this latent ability would be mathematical literacy. An additional subscript i is introduced to notate items. To model the probability that student j in school m responds correctly to item i,

$$\Pr(y_{ijm} = 1|\theta_{jm}, b_i) = \frac{\exp(\theta_{jm} + b_i)}{1 + \exp(\theta_{jm} + b_i)}, \tag{2.16}$$

where θ_{jm} is the student's latent ability, b_i is the location parameter for item i (i.e., the location where half of students would respond correctly to the item), and y_{ijm} is the student's response for item i. Note that y_{ijm} is the response on a single item, whereas Y_{jm} from Equation 2.1 is the full test score for the student. This particular formulation is the Rasch model, where all items discriminate equally among students (Rasch, 1960). Note that relaxing this assumption allows for the two-parameter logistic IRT model with an additional parameter. Other IRT models are also possible, but for simplicity, I will focus on the Rasch model.

Similarly to Equations 2.1 and 2.2, student latent ability can be modeled with multiple levels,

$$\theta_{jm} = \beta_{0m} + r_{jm}, \tag{2.17}$$

and

$$\beta_{0m} = \gamma_{00} + u_{0m}. \tag{2.18}$$

u_{0m} is still interpreted as the deviation of school m from the school mean scores, and can also be interpreted as the unique contribution of school m. Student-level residuals r_{jm} are normally distributed with a mean of zero and a variance, σ^2. School-level random effects are also distributed normally with a mean of zero and a variance of ϵ^2. Again, ϵ^2 represents the amount of influence schools have on student learning, and the within-school variance, σ^2, indicates the amount of equality within schools.

Under these assumptions, because there are no covariates yet in the model,

$$\theta_{jm} = r_{jm} + u_{0m}, \tag{2.19}$$

which represents a student-level latent component and a school-level latent component. Often in multilevel IRT, the components are rewritten so that $r_{jm} = \theta_{jm}$ and $u_{0m} = \theta_m$, then substituted into the the model giving,

$$\Pr(y_{ijm} = 1|\theta_{jm}, \theta_m) = \frac{\exp(\theta_{jm} + \theta_m + b_i)}{1 + \exp(\theta_{jm} + \theta_m + b_i)}, \tag{2.20}$$

so that both the student latent trait and the school latent trait are represented with similar notations. The formulation in Equation 2.20 will be used in later sections.

2.2.2 Type A and B Effects

Type A and B effects are modeled following the same reasoning as in the HLM framework. Relevant covariates are added to Equations 2.17 and 2.18,

$$\theta_{jm} = \beta_{0m} + \beta_{1m}X_{1jm} + ... + \beta_{Qm}X_{Qjm} + r_{jm}, \tag{2.21}$$

and

$$\beta_{0m} = \gamma_{00} + \gamma_{01}W_{1m} + ... + \gamma_{S}W_{Sm} + u_{0m}. \tag{2.22}$$

X_{qjm} represents student background characteristics, and β_{qj} represents the corresponding student-level regression coefficients. Similarly, W_{qm} represents school characteristics, and γ_{qm} represents the corresponding school-level coefficients.

2.2.3 Examples of Multilevel IRT Studies of School Effects

Multilevel IRT is a useful advancement in school effectiveness methodology and is gaining in popularity. However, it is not yet as widely used as the HLM approach without the measurement model that is outlined in Section 2.1. In fact, Hairon, Goh, Chua, and Wang (2017) highlight the benefits in applied research on teacher and school development and encourage researchers to utilize the approach more often. Many of the initial articles on multilevel IRT are primarily illustrative. Fox (2004) demonstrates the use of multilevel IRT for understanding the impact of schools on student mathematics learning in Dutch schools and how explanatory variables can be incorporated at each level. The analysis shows how the ranking of schools changes when the measurement model is included. Fox (2005) extends this work by introducing an IRT model for polytonomous responses. Pastor (2003) takes a different view of school effectiveness and illustrates the differences between schools on students' academic self-esteem.

Many of the school effectiveness studies utilizing multilevel IRT focus more on

explaining differences between groups, rather than for ranking or accountability purposes. One exception is Bacci and Caviezel (2011), who use a multilevel partial credit model to rank teachers based on student satisfaction. There are also a few studies that are examples of a more explanatory approach. Briggs (2008) uses a number of student- and school-level covariates to analyze group differences in science achievement. Höhler, Hartig, and Goldhammer (2010) study students' foreign language competence within- and between-schools, while providing an additional methodological advancement by combining multilevel IRT with multidimensional IRT. Sulis and Toland (2017) explore differences in mathematics achievement between classrooms in Italian schools, and consider whether differences between students are explained by gender and SES.

2.2.4 Assumptions and Limitations

Multilevel IRT modeling is based on a number of assumptions. First, the assumptions of the selected IRT model at the item level must be considered. The above formulation uses the Rasch model, which assumes equal discrimination for all items (Rasch, 1960). However, this may not be a valid assumption for the assessment, as is the case in many of the examples provided above, and another IRT model (e.g., two-parameter logistic model) may be more appropriate. Additionally, most IRT models used for looking at school effects represent a single dimension, such as math, and do not consider possible sub-dimensions of math skills or the cognitive processes required for solving items. In order to incorporate this additional information, an approach such as outlined in Höhler et al. (2010) would be necessary.

The multilevel IRT approach shares many of the same limitations as the HLM approach. While multilevel IRT addresses the limitation of the two-step HLM framework by incorporating the measurement model, it still relies on the assumption of equal variances within and across sites Pastor (2003). In fact, Figure 2.1 is essentially identical for multilevel IRT, where the X axis labels are θ_m in place of u_0, and θ_{jm} in place of r. In other words, this model assumes that both students and schools are sampled from

homogeneous populations. However, as explored in Figure 1.1, this assumption of a homogeneous population may not be valid. The assumption of equal variance at the within-level means that the model assumes all schools have the same achievement gap within schools. Not only is this assumption likely invalid, understanding whether some schools successfully reduce the achievement gap is worth investigating.

Finally, multilevel IRT also does not provide a solution to the challenges of classifying students and schools. When the goal of an analysis is to classify students and schools for performance-targeted intervention, additional steps must be taken, just as in the HLM framework.

2.3 Chapter Conclusion

Despite important methodological contributions to the field of school effectiveness research, more research is needed to ensure that the modeling approach addresses the complex challenges inherent in understanding the impact schools have on students. Additionally, more needs to be done in order to improve the classification process for grouping students and schools for performance-targeted interventions.

The next chapter reviews a number of models from the field of psychometrics that are theoretically and statistically related to multilevel IRT and the Multilevel Diagnostic Item Response model, but have not yet been used in school effectiveness research.

CHAPTER 3

Related Psychometric Models Review

The importance of strong methods to classify students and schools for use in performance-based indicators, and the complexity of the process illustrated in Figure 1.1 are well-known in the field of psychometrics. Of particular interest in psychometrics is the classification of respondents based on the belief that the distribution of respondents represents a finite number of latent (i.e., unobservable and inferred from the data) classes (Gnaldi, Bacci, & Bartolucci, 2016). In the case of school effectiveness research, school-level latent classes could represent levels of school performance, based on student achievement data. The benefits of classifying schools based on performance include the ability to target reforms and interventions, as well as to understand the characteristics of the higher performing schools.

The psychometric models reviewed in this section represent existing methods that primarily aim to classify respondents into latent classes based on their responses to test items. Each approach has strengths and weaknesses in terms of its capability to account for latent heterogeneity, while providing useful information for policy makers and practitioners. Additionally, each model has a theoretical and statistical relationship to the proposed Multilevel Diagnostic Item Response model that will be presented in Chapter 4.

The first type of model is an extension of IRT models, termed Mixture IRT, which aims to improve the modeling of heterogeneity in IRT. The next set of models are termed Cognitive Diagnostic Models (CDMs) or Diagnostic Classification Models (DCMs), and these have important similarities to the proposed model regarding classification

of respondents and the use of item information. The next two models are historically influential in cognitive diagnostic modeling and in the development of the Multilevel Diagnostic Item Response model: (1) the Log Linear Test Model (Fischer, 1973) and (2) the Saltus Model (Wilson, 1989). Each of these models is reviewed below.

3.1 Mixture Item Response Theory

3.1.1 Rationale

As mentioned in Chapter 2, the application of IRT models requires a number of assumptions. Mixture IRT aims to address the key assumption that the IRT model holds across the population, regardless of latent class. Mixture IRT relaxes this assumption by incorporating a latent class analysis into the model, which allows the IRT model to vary across classes, but to hold within a class (Gnaldi et al., 2016; Rost, 1990; Smit, Kelderman, Flier, et al., 2000). Typically, the models incorporate a finite number of mixture (normal gaussian) distributions in order to infer group membership based on observations from multiple latent classes (Marcoulides & Heck, 2013; B. O. Muthén, 1989; von Davier, 2010). The focus on latent classes is different from multiple-group models, which allow for the IRT model to differ by groups that are based on observable background characteristics. The benefit of combining the IRT model with latent class analysis is that students are assigned to classes, but student ability is also allowed to vary within a class. This is possible because both continuous and categorical latent variables are estimated, fulfilling the practical need for diagnosis (categorical variables) and comparison of schools and students (continuous variables).

3.1.2 Student-level Formulation

Due to the complexity of mixture IRT models, I begin by outlining the student-level formulation, where items are nested in students. Similarly to the previous IRT section,

mixture IRT models can include multiple parameters such as in the two-parameter, three-parameter, and generalized partial credit models. For simplicity, I continue to focus on the Rasch IRT model.

The probability of student j responding correctly to item i is conditional on the student's latent ability θ_{jg} and their latent class assignment, $C = (1, ..., g, ..., G)$,

$$\Pr(y_{ij} = 1|\theta_{jg}, C = g) = \frac{\exp(\theta_{jg} + \beta_{ig})}{1 + \exp(\theta_{jg} + \beta_{ig})}. \tag{3.1}$$

β_{ig} is the item location parameter for latent class g. Importantly, $\theta_{jg} \sim \mathcal{N}(\mu_g, \sigma_g^2)$, which highlights another key difference between typical IRT models and mixture IRT models. The group level subscript in the distributions indicates that the means and variances are allowed to vary across latent classes. In other words, some student groups may perform higher, on average, compared to others. They may also have larger gaps between low and high performing groups within a class.

The marginal probability model across latent classes is,

$$\Pr(y_{ij} = 1|\theta_j) = \sum_{g=1}^{G} \Pr(C = g)\Pr(y_{ij}|\theta_{jh}, C = g), \tag{3.2}$$

where $\Pr(C = g) = \pi_g$, or the probability of belonging to class g, and $\sum_{g=1}^{G} \pi_g = 1$. The probability of belonging to class g can be interpreted as the proportion of respondents in class g where no explanatory variables are included in modeling π_g.

Covariates can be included for predicting either the probability of person j belonging to class g, or for predicting the latent trait θ_{jg} (Li, Jiao, & Macready, 2016). Technical reasons for including covariates include more accurate parameter estimates and latent class assignment (e.g., Lubke & Muthén, 2005).

3.1.3 Multilevel Formulation

In order to be useful for school effectiveness studies, the multilevel IRT model can be extended to include a school level. Incorporating additional levels follows much of the same logic as in multilevel IRT, but with the relaxed assumption of the IRT model holding across latent classes. Latent classes are incorporated in both the student-level and the school-level of the model, where $C = (1, ..., g, ...G)$ now becomes the notation for student-level classes, and $K = (1, ..., h, ..., H)$ is for school-level classes. Following the same logic from the multilevel IRT section, the student latent trait, θ_{jm} from Equation 3.1 can be decomposed into the student component, still θ_{jm}, and a school component, θ_m. Now the conditional model, conditional on student group and school group, can be formulated as,

$$\Pr(y_{ijm} = 1|\theta_{jmgh}, \theta_{mh}, C = g, K = h) = \frac{\exp(\theta_{jmgh} + \theta_{mh} + \beta_{igh})}{1 + \exp(\theta_{jmgh} + \theta_{mh} + \beta_{igh})}. \tag{3.3}$$

In this model, it is possible for the location parameter, β_{igh}, to vary both by the student-level classes and the school-level classes. Again, the distributions of the latent traits at both levels are allowed to vary by classes. In other words, $\theta_{jmgh} \sim \mathcal{N}(\mu_{gh}, \sigma^2_{gh})$, and $\theta_{mh} \sim \mathcal{N}(\mu_h, \sigma^2_h)$. By allowing the between-school variance components to differ across groups, this model accounts for the possibility that variation in school effects differs by school latent groups. Allowing for different within-school variance components accounts for the possibility that schools in each group may have differing levels of achievement gaps depending on the student group.

Specifying the marginal probability model is more easily shown in two equations:

$$\Pr(y_{ijm} = 1|\theta_{jm}, \theta_{mh}, K = h) = \sum_{g=1}^{G} \Pr(C = g|K = h)\Pr(y_{ijm}|\theta_{jmgh}, \theta_{mh}, C = g, K = h), \tag{3.4}$$

and

$$\Pr(y_{ijm} = 1|\theta_{jm}, \theta_M) = \sum_{h=1}^{H} \Pr(K = h)\Pr(y_{ijm}|\theta_{jm}, \theta_{mh}, K = h). \quad (3.5)$$

In Equation 3.4, $\Pr(C = g|K = h) = \pi_{gh}$ is the probability of a student belonging to class g, conditional on the school class membership. In other words, the probability of belonging to different student classes depends on the class of school in which the student attends. In Equation 3.5, $\Pr(K = h) = \pi_h$ is the probability of a school belonging to class h.

Vermunt (2008) highlights that a number of special cases can exist in this formulation. One special case the author provides is when the student-level group membership probabilities do not depend on school-level group membership, but the parameters of the IRT model do vary based on the school-level membership. In other words, the school-level classes capture the common variation in responses within a school. Another special case is when student-level group membership probabilities do depend on school-level group membership, but the parameters of the IRT model do not. In other words, the association between item responses is fully captured by the student groups, and the school groups capture the association between students (Vermunt, 2008). Finally, covariates can also be included at both levels in order to improve class assignments (Gnaldi et al., 2016; Vermunt, 2010)

3.1.4 Confirmatory versus Exploratory Applications

Typically, mixture IRT models are used in an exploratory manner to understand the number of latent classes as well as the distribution of the latent variable in the population (e.g., B. O. Muthén, 1989). Examples of research using multilevel mixture IRT to account for school clustering and school impact have primarily used an exploratory approach. Vermunt (2007) uses mixture IRT to determine whether items function differently across school groups. To do this, the author introduces school-level classes, but not student-level classes. Cho and Cohen (2010) look for differential item functioning (DIF) at

both the student and school levels using a Bayesian estimation framework. Gnaldi et al. (2016) extend the earlier work of Bartolucci (2007) by extending Latent Class IRT (LC-IRT) to include multiple levels and multidimensional traits in order to examine whether subgroups show distinct response styles, and whether expected student scores are characterized by schools. These examples are exploratory because they do not have a pre-determined theory guiding the number of classes. Instead, they compare models with different numbers of classes at each level in order to determine the number of classes that lead to the best model fit.

Some examples of confirmatory mixture IRT models do exist, but not in the field of school effectiveness. For example, Mislevy and Verhelst (1990) use mixture IRT models to identify a fixed number of solution strategies employed by test takers, based on substantive theories of different strategy types. Bolt, Cohen, and Wollack (2002) apply a mixture Rasch model to explore test speededness by comparing a "speeded" class with a "nonspeeded" class.

3.1.5 Limitations

The flexibility of mixture IRT models allows for successful modeling of student and school heterogeneity. The use of a continuous latent trait allows for within-class comparisons and an understanding of dispersion within and across schools. Incorporating a categorical latent class variable provides the opportunity for classifications of students and schools. As such, multilevel mixture IRT models address an important limitation outlined in the previous chapter for estimating school effects with HLM and multilevel IRT. However, without adopting an a priori theory for interpreting the latent classes, classification assignments are difficult to interpret in the context of targeting policy and program interventions. Additionally, the complexity of these models means that they can be quite difficult to estimate. Finally, multilevel mixture IRT does not incorporate curricular information, and therefore, the estimated scores and classifications are based on a general dimension such as mathematical literacy, as opposed to a more nuanced definition that

includes cognitive processes.

3.2 Diagnostic Classification Models

3.2.1 Rationale

Up to this point, the models reviewed have focused on a continuous latent trait. In the case of mixture IRT, a categorical variable was added in order to classify students and schools based on latent classes. In this section, I review a modeling framework, called Diagnostic Classification Models (DCMs; sometimes referred to as CDMs), because these models are growing in popularity in the field of education for the purposes of providing diagnostic feedback. DCMs are confirmatory latent class models. The goal of latent class analyses is to classify respondents into unobserved categorical latent classes (Lazarsfeld & Henry, 1968). More specifically, the latent classes are based on attribute profiles, where the attribute profile represents the students' standings on a number of tested attributes. These attributes can represent various subdimensions or cognitive functions, such as those measured by PISA-D mathematics. Under the DCM framework, students can be classified as mastering or not mastering each of the cognitive processes (formulate, employ, and interpret). In this case, there are $2^3 = 8$ possible attribute profiles, one of which is where a student has mastered the formulate process, but not the other two. This type of feedback can be useful to teachers who are aiming to improve instruction for their students. In the case of this student, the teacher would want to focus on employing strategies and interpreting results.

Incorporating the variety of available information about the items into the model is a key component of DCMs. Having this information, as well as a substantive theory regarding the latent classes (attribute profiles) makes these confirmatory models. The next section provides the formulation of a general DCM, followed by discussions of extensions that have relevance for school effectiveness research.

3.2.2 Formulation: Log-linear Cognitive Diagnosis Model

Research around DCMs is growing rapidly, and summaries of the field of diagnostic models already exist (see for example, Rupp & Templin, 2008). Therefore, I limit this discussion to one general DCM model, the Log-linear Cognitive Diagnosis Model (LCDM; Henson, Templin, & Willse, 2009) which encompasses other well-known DCMs (e.g., DINA and the DINO models). Again, the LCDM is a constrained latent class model that utilizes categorical variables to classify students based on theoretically determined attribute profiles. In an exam with A number of attributes, the attribute profile of student j is given by $\boldsymbol{\alpha}_j = (\alpha_{j1}, ..., \alpha_{ja}, ...\alpha_{jA})$. In an exam measuring the cognitive processes of PISA-D, student j from above who has mastered the first, but not the second and third attributes, would have an attribute profile of $\boldsymbol{\alpha}_j = [1, 0, 0]$.

In order to link the test items to the specific subdimensions or pre-specificed test information, a design matrix called the Q-matrix is included in the item-level model. For an exam with I items and A attributes, the Q-matrix is an $I \times A$ matrix, with I rows and A columns.

Here, the probability of a correct response on item i for student j is conditional on student j's attribute profile $\boldsymbol{\alpha}_j$, rather than the continuous latent trait of θ_j as in previous latent models:

$$\Pr(y_{ij} = 1|\boldsymbol{\alpha}_j) = \frac{\exp(\lambda_{0i} + \boldsymbol{\lambda}_i^T \mathbf{h}(\mathbf{q}_i, \boldsymbol{\alpha}_j))}{1 + \exp(\lambda_{0i} + \boldsymbol{\lambda}_i^T \mathbf{h}(\mathbf{q}_i, \boldsymbol{\alpha}_j))}, \tag{3.6}$$

where the intercept for this item, $\lambda_{i,0}$ represents the log-odds of a correct response for a student who has not mastered any of the A attributes that are tested by this item. ˘$_i$ is the regression parameter for item i, $\mathbf{q}_i$ is the vector of Q-matrix entries indicating whether the item measures each attribute, and $\boldsymbol{\lambda}_i^T \mathbf{h}(\mathbf{q}_i, \boldsymbol{\alpha}_j)$ is a vector of linear combinations that indicate whether or not the parameter is present for student j on item i. For a more in depth formulation of the conditional model, including a discussion of main effects and interaction terms for attributes, see Henson et al. (2009).

Now, let there be $C = [1, ..., g, ...G]$ possible latent classes where each class is represented by an attribute profile and $G = 2^A$. The marginal probability model can be expressed as,

$$\Pr(y_{ij} = 1) = \sum_{c=1}^{G} \Pr(C = g)\Pr(y_{ij}|C = g), \tag{3.7}$$

where $\Pr(C = g) = \pi_c$, and π_c is the proportion of students who belong to that particular latent class (i.e., have a given attribute pattern). As in previous models, $\sum_{c=1}^{C} \pi_c = 1$.

The LCDM is a complex model with a large number of parameters and subscripts. More in-depth information on the formulation of the LCDM, including how specific constraints are placed to specify other popular DCM models, is presented in Henson et al. (2009). Another model presented by von Davier (2008) is even more general than the LCDM in that the LCDM model can be considered a special case of this general diagnostic model (GDM; von Davier, 2014). The GDM is also based on latent class analysis and can be thought of as a general framework for confirmatory multidimensional item response models (von Davier, 2010). For the purposes of this research, the key aspect of both of these models and their related frameworks is the use of a Q-matrix to link items to more nuanced aspects of test items in order to classify students into categorical latent classes.

3.2.3 Multilevel DCMs

Research extending DCMs to multiple levels has been limited, and therefore, not yet used in school effectiveness research. At the time of writing, two examples of early work in using DCMs in a multilevel framework have been presented. First, von Davier (2010) presents a Hierarchical General Diagnostic Model (HGDM) as an extension to the single-level GDM (von Davier, 2008). Specifically, the HGDM accounts for clustering by modeling the probability of latent group membership conditional on cluster (e.g., school) membership. It estimates class-specific profiles and parameters, as well as school-level cluster proportions (i.e., the proportion of students in each class in a school). The author

uses an example in language testing to demonstrate how modeling student clustering provides information that improves the classification of students into high- and low-proficiency groups. While understanding differences in the distribution of the latent attributes across schools is important, it does not classify schools directly from the model. Finally, student- and school-level covariates are not incorporated in the HGDM, making it difficult to isolate school impact from other background characteristics.

W. C. Wang and Qiu (2019) introduce a different approach to multilevel DCMs. This modeling approach uses the LCDM at the item level and can incorporate student and school background covariates to understand their effects on attribute mastery. The authors argue that a major benefit of using the multilevel DCM is that the standard errors of the covariates are estimated accurately, which leads to appropriate statistical tests of the covariates and improves the accuracy of student classifications. To do this, two approaches are provided: (1) the latent continuous variable approach and (2) the multivariate Bernoulli distribution approach. The first approach is outlined here because much of its formulation is quite similar to that of the HLM provided in the previous chapter.

Assume a continuous trait α_a^* underlies the categorical trait α_a. If α_a^* is larger than some threshold, the trait is considered mastered, and $\alpha_a = 1$. If α_a^* is smaller than that threshold, then $\alpha_a = 0$, and the trait is considered not mastered. It is assumed that the vector $\boldsymbol{\alpha}^* = (\alpha_1^*, ..., \alpha_A^*)$ follows a multivariate normal distribution that has a mean vector $\mu^T = (\mu_1, ..., \mu_A)$ and a variance for all attributes fixed to one.

Now, let α_{jma}^* denote the continuous latent trait for student j in school m on attribute a. Then, α_{jma}^* can be modeled as with Q number of student covariates,

$$\alpha_{jma}^* = \beta_{0ma} + +\beta_{1ma}X_{1jma} + ... + \beta_{Qma}X_{Qjma} + r_{jma}, \tag{3.8}$$

where X_{qjma} is the qth student covariate, β_{0ma} the intercept, and β_{qma} the related regression slope. r_{jma} is the residual and follows a multivariate normal distribution with means of zero and a variance/covariate matrix Σ. At the school-level,

$$\beta_{qjm} = \gamma_{0a} + \gamma_{1a} W_{1ma} + ... + \gamma_{Sma} W_{Sma} + u_{ma}, \tag{3.9}$$

where W_{sma} is the school-level covariate, γ_{0a} the intercept, and γ_{sma} the school-level regression slope. u_{ma} also follows a multivariate normal distribution with means of zero and a variance/covariance matrix Ω.

The benefits of this model are that it is understandable within the HLM framework, and that it can be estimated with available software. However, as the number of attributes increases, so does the dimensionality of the integration, making estimation quite slow. Additionally, the model does not provide school-level classifications.

3.2.4 Limitations

DCMs are very useful in providing student-level classifications that can result in nuanced feedback regarding student strengths and weaknesses. Modeling categorical latent variables based on item responses eliminates the need to determine cut-scores and classifications that are complicated with continuous latent variable models. Additionally, the DCMs utilize important and relevant information that is often provided by test makers. However, DCMs have a number of limitations in terms of their utility for school effectiveness research. First is the limited research on DCMs in multilevel settings. The new approach outlined above illustrates some early work in this area, but much more effort is needed in order for multilevel DCMs to be useful in classifying schools for policy and interventions. Second, research has also been limited on the inclusion of covariates in DCMs. Ayers, Rabe-Hesketh, and Nugent (2013) illustrate this using student-level covariates, but not school-level. Future work beyond the W. C. Wang and Qiu (2019) paper is needed in this area. Finally, because DCMs only use categorical latent variables, they are unable to provide comparison data within latent classes.

3.3 Log Linear Test Model

3.3.1 Rationale

One of the early models to make use of subdimensions or cognitive functions attributed to test items is the Log Linear Test Model (LLTM; Fischer, 1973). The LLTM is an important adaption of the Rasch model because it considered the impact of cognitive dimensions on the attributes of the item. In the development of this model, Fischer (1973) argued that the model can be useful in instructional research if the subject is made up of tasks or items that are a combination of particular types of cognitive operations. The example originally proposed by Fischer groups items based on different domains in calculus, and the author uses the LLTM to test that the seven domains exist as hypothesized.

3.3.2 Formulation

The formulation of the LLTM makes an important change to the Rasch model in Equation 2.16 (without the subscript m for the schools). Specifically, the LLTM decomposes the item location parameter of item i,

$$\beta_i = \sum_{a=1}^{A} \beta_a X_{ia}, \tag{3.10}$$

where X_{iq} is an indicator variable denoting whether item i belongs to item group a, and β_a is the effect item group a has on the location parameter. The item groups are determined a priori by substantive theory regarding the subject matter being tested. For example, the three PISA-D cognitive functions (formulate, employ, and interpret) could influence the location of the items.

3.3.3 Limitations

This model differs from cognitive diagnostic models because the item factors do not vary by respondent, and therefore, there is no classification of respondents based on these

factors (Hartz, 2002; Stout, 2007; von Davier, 2009). Additionally, the LLTM rests on the strong assumption that the location parameter is predicted fully by the item characteristics, although this assumption can be relaxed (Jeon, Draney, & Wilson, 2015). For additional information on the importance of the LLTM and its influence on subsequent models (e.g., MLGM; Embretson, 1984), see Hartz (2002).

3.4 Saltus Model

3.4.1 Rationale

One model that was influenced by the LLTM is the Saltus model (Wilson, 1989). The Saltus model was initially developed in order to understand cognitive leaps in development and to assign students to Piagetian type stages, although its use can be extended beyond this original purpose (Jeon, 2018; Jeon et al., 2015). The benefit of the Saltus model is that it incorporates categorical latent classes and a continuous latent trait, and therefore, can be useful for both classification and comparisons within classes. This makes it similar to mixture IRT, but it differs in that it is confirmatory with a pre-specified number of latent classes. Another benefit is that the Saltus model incorporates item information regarding subdimensions or cognitive functions, based on the approach taken in the LLTM. This allows it to provide more nuanced information beyond a score on the latent trait.

3.4.2 Formulation

In the Saltus model, the probability of student j responding correctly to item i is conditional on the latent trait $\theta_{j(g)}$ and latent group membership $C_j = h$, and is modeled as,

$$\Pr(Y_{ij} = 1|\theta_{j(h)}, C_j = h) = \frac{\exp(\theta_{j(h)} - \beta_i + \sum_a \tau_{ja} b_{ia})}{1 + \exp(\theta_{j(h)} - \beta_i + \sum_a \tau_{ja} b_{ia})}. \tag{3.11}$$

β_i represents the item location parameter. $\theta_{j(g)}$ similarly represents the ability of student

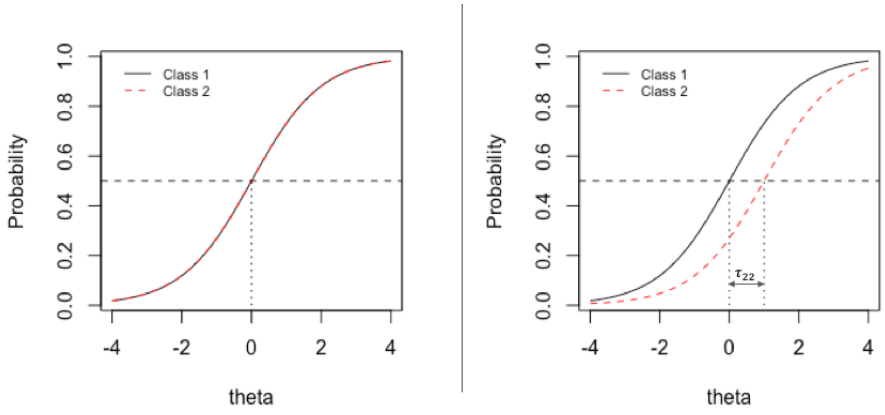

Figure 3.1: Item response curves for two latent classes for the reference item group (left) and for the focal item group (right). τ_{22} represents the difference in item difficulty between class 1 and class 2. Theta is student mathematical ability.

j (who is in class g) on the construct of interest, and is distributed with $\theta_{j(g)} \sim \mathcal{N}(\mu_g, \sigma_g^2)$. The additional and key parameters of the Saltus model are based on the ideas in the LLTM model in that items are divided into groups based on theoretical aspects of the items. b_{ia} is an indicator variable of whether item i belongs to item group a, where $(a = 1, ..., A)$. τ_{ja} represents the difference in item location for students in class g on item group a, compared to the reference class.

Figure 3.1 represents item response curves for two latent classes when there are two item groups. In the Saltus model, one group of items is set as the reference group. The item response curve on the left of Figure 3.1 is the response curve for an item in the reference item group. In this case, there is no difference between classes on the items in these groups. The response curve on the right is for an item in the focal item group. On this item, there is a difference in item location between the two latent classes that is shown as τ_{22}, and this indicates that items in this group are more difficult for students in class 2.

The choice to decompose the item parameter by item groups is based on the assumption that students in different latent groups perform differently on certain types of

items. This is useful when the substantive theory suggests that item difficulty changes systematically based on student class membership (Jeon et al., 2015).

The marginal probability across latent classes is modeled by,

$$\Pr(Y_{ij} = 1|\theta_j) = \sum_{g=1}^{G} \Pr(C = g)\Pr(Y_{ij} = 1|\theta_{j(h)}, C_j = h). \tag{3.12}$$

where $\Pr(C = g) = \pi_g$, which is the probability of a student belonging to class g. In the original Saltus model, the number of student latent classes is equal to the number of item groups (i.e., $H = G$). However, this can be relaxed in a more generalized version of the Saltus model presented in Jeon et al. (2015). For identification, one latent class is set as the reference class with a mean of $\theta = 0$, or the item location parameters are set as $\sum \beta_i = 0$.

The original Saltus model is based on the Rasch model. Jeon (2018) extends this basic model by incorporating the IRT discrimination parameter and polytonomous responses. The Saltus model has also been explored for use in standards setting (Draney & Jeon, 2011).

3.4.3 Limitations

By simultaneously using categorical latent class variables for classification and continuous latent traits for within-class comparisons, as well as incorporating more specific item information, all of which are guided by theory, the Saltus model contains key components needed for a useful model in understanding school impacts on student learning. However, the Saltus model is only formulated for the student-level and cannot provide school-level information.

3.5 Chapter Conclusion

Overall, a number of interesting models exist to help researchers understand student responses and item characteristics, as well as the fact that student characteristics influence

these responses. However, no existing models provide clear classifications at both student- and school-levels while also providing useful and nuanced feedback. The next chapter presents the development of the Multilevel Diagnostic Item Response model, which incorporates key components of the Saltus model, and draws on the multilevel analyses in the other modeling frameworks.

CHAPTER 4

Multilevel Diagnostic Item Response Model

4.1 Research Motivation and Aims

4.1.1 Motivation

The aim of this research is to address the need for a school effects model that classifies schools into categories based on average student performance, as well as classifying students within schools. For the purposes of performance-targeted interventions, the categories need to be based on overall performance, as well as on relevant aspects of the curriculum. In other words, for particular interventions, it is useful to select students and schools based on more than a single score. Rather, it may be useful to define the latent classes by overall performance and differences on a certain attribute of interest. In addition to these key goals, it is also useful for the model to be able to account for differences in achievement gaps within schools.

Currently, no tool is available that addresses all of these goals. To do so, the model must use a multilevel framework, as in multilevel IRT, include both categorical and continuous latent variables, as in mixture IRT, and incorporate item attribute information, as in DCMs or the Saltus model. The Multilevel Diagnostic Item Response (MD-IR) model proposed here addresses these concerns. The MD-IR model can be viewed as a multilevel mixture IRT model with strategic theoretical constraints on the latent classes and the item attribute groupings. It can also be viewed as a multilevel extension of the Saltus model. As such, the development of the MD-IR model is based on advanced psychometric modeling such that it addresses a gap in the available models in the school effectiveness

literature.

4.1.2 Key Components of MD-IR

The MD-IR model is a multilevel model that provides estimates of continuous and categorical latent variables, at both the student- and school-levels. The inclusion of the continuous latent variable is the key component for considering this model a school effects model. At the school-level, the continuous normal latent variable represents the school effect, and its variance represents the amount of impact schools have on student achievement. At the student-level, the continuous normal latent variable represents student achievement, and its variance represents achievement gaps within schools. In these respects, the MD-IR model is similar to multilevel IRT.

In addition to the continuous latent variable, the MD-IR also includes a categorical latent variable in the form of latent classes for students and schools. The addition of the categorical latent variables is similar to mixture IRT, and allows for students and schools to be classified based on item performance. A key difference between typical mixture IRT applications and the MD-IR model is that the MD-IR model utilizes a confirmatory approach in determining the number of estimated latent classes. A second key difference between mixture IRT and the MD-IR is that the latent classes are defined by more than the average performance on the general dimension. In the MD-IR, the classes are also defined by differences in performance on an attribute of interest.

The MD-IR model makes use of important attribute information that is typically included by test developers. In the case of the PISA-D, the MD-IR model can incorporate the cognitive functions required for mathematical literacy, and as such, can differentiate students and schools based on performance on the overall trait, and the cognitive functions. This is similar to the Saltus model, except that the MD-IR model is able to do this for both students and schools. In other words, the average gain of a particular student class on items of a particular attribute does not have to be the same if the students' schools belong to different school classes.

Before fully introducing the MD-IR model, the next section first provides notation, definitions, and the context for presenting the model. Next, the formulation of the model is given, along with other technical details. Last, we provide a discussion or possible extensions for common applications.

4.2 Notation and Context

Let y_{ijm} be the response on item $i(i = 1, ..., I)$ for student $j(j = 1, ..., J_m)$ in school $m(m = 1, ..., M)$. Generally, item responses can be polytonomous or dichotomous and are assumed to be conditional on student latent achievement θ_{jm} and school latent effect θ_m, both of which are continuous traits. It is also expected that students clustered within schools are similar in ways that are attributed to the school in which they attend. Students are assumed to belong to G number of categorical latent classes, denoted by $C = (1, ..., g, ..., G)$, and that schools belong to H number of school-level latent classes, denoted by $K = (1, ..., h, ..., H)$, both of which are unknown and inferred from the data. For both levels, G and H are defined prior to model estimation by substantive theory or practical interest. It is assumed that θ_{jm} and θ_m are made up of G and H number of normal distributions, respectively.

Items are also believed to belong to one of A number of attributes. Attributes are also defined a priori and can represent subdomains of the continuous latent trait of interests (e.g., formulate, employ, and interpret in the PISA-D math assessment). Item location parameters, β_i, are expected to be influenced by the item's attribute, such that τ_{gha} represents the difference in location parameter of items measuring attribute a, and that this difference varies by student group g and school group h. This will be discussed more fully below.

For simplicity and clarity, the MD-IR model is presented under a relatively simple context. The formulation presented below assumes binary response data. It focuses on only two latent classes at both levels, such that $H = G = 2$. As in the context of

performance-targeted interventions, we are looking to differentiate between schools and students in need of support and those that do not. Additionally, there are only two attributes, $A = 2$, and items only measure one attribute at a time.

4.3 Formulation

To formulate the MD-IR model, let C_{jm} denote the student class for student j in school m. Also let K_m denote school m's school-level class. The conditional probability of student j in school m responding correctly to item i is given as,

$$\Pr(y_{ijm} = 1|\theta_{jm(gh)}, \theta_{m(h)}, C_{jm} = g, K_m = h) = \frac{\exp(\theta_{jm(gh)} + \theta_{m(h)} - \beta_i + \tau_{gha} b_{ia})}{1 + \exp(\theta_{jm(gh)} + \theta_{m(h)} - \beta_i + \tau_{gha} b_{ia})}. \tag{4.1}$$

The parameters are defined as follows:

- $\theta_{jm(gh)}$ is the achievement of student j, who attends school m. $\theta_{jm(gh)} \sim \mathcal{N}(\mu_{gh}, \sigma^2_{gh})$. The student class means and variances differ across groups, with one student-level group set as the reference class (discussed below). Allowing the student-level variance to differ across groups allows for the possibility that within-school gaps could differ across latent classes.

- $\theta_{m(h)}$ is the school effect of school m. $\theta_{m(h)} \sim \mathcal{N}(\mu_h, \sigma^2_h)$. The school class means and variances differ across groups, with one school-level group set as the reference class. Allowing the school-level variance to differ across classes allows for the possibility that school classes may differ in their impact.

- β_i is the item location parameter. Note that β_i is not called "difficulty". The interpretation of this parameter in relation to item difficulty will be discussed below. Also note that β_i is purposely equal across all latent classes, which is a key difference from typical mixture IRT models.

- τ_{gha} represents the difference in item location for students in class g who attend a school in class h on an item measuring attribute a, compared to students in the reference class. This is the key component that differentiates the latent classes and will be discussed more fully in Section 4.6.2.

- b_{ia} is an indicator variable indicating whether item i measures attribute a. One attribute is set as the reference group.

More details on interpreting these parameters and the reference classes are discussed in Section 4.6.

The marginal probability with respect to the latent classes is given as the probability of a correct response conditional on student achievement and school effect,

$$\Pr(y_{ijm} = 1|\theta_{jm}, \theta_m) = \sum_{h=1}^{H} \pi_h \Pr(y_{ijm}|\theta_{jm}, \theta_{m(h)}, K_m = h), \tag{4.2}$$

where $\pi_h = \Pr(K_m = h)$ is the probability of school m belonging to class h.
$\Pr(y_{ijm}|\theta_{jm}, \theta_{mh}, K_m = h)$ is the average conditional success probability for school m given student-level class proportions. It can be expressed as,

$$\Pr(y_{ijm} = 1|\theta_{jm}, \theta_{mh}, K_m = h) = \sum_{g=1}^{G} \pi_{g|h} \Pr(y_{ijm}|\theta_{jm(gh)}, \theta_{m(h)}, C_{jm} = g, K_m = h), \tag{4.3}$$

where $\pi_{g|h} = \Pr(C_{jm} = g|K_m = h)$, which is the probability of a student belonging to class g, conditional on the school class membership. In other words, the probability of belonging to different student classes depends on the student-level class proportions in the school that the student attends. $\Pr(y_{ijm}|\theta_{jm(gh)}, \theta_{m(h)}, C_{jm} = g, K_m = h)$ is the success probability for student j in school m given in Equation 4.1.

4.4 Maximum Likelihood Estimation

Estimation of model parameters can be conducted using maximum likelihood. The likelihood of the observed data, assuming local independence of items and where $\mathbf{\Phi}$ is a vector of all parameters in the model, is defined as,

$$L(\mathbf{\Phi}|\mathbf{Y}) = \prod_{m=1} \int_{\theta_m} \left[\sum_{h=1}^{H} \pi_h \prod_{j=1}^{J_m} \int_{\theta_{jm}} \left[\sum_{g=1}^{G} \pi_{g|h} \prod_{i=1}^{I} \Pr(y_{ijm}|\ \theta_{jm(gh)}, \theta_{mh}, C = g, K = h) \right] g(\theta_{jm(gh)}) d\theta_{jm}. \right] h(\theta_{m(h)}) d\theta_m \tag{4.4}$$

$\mathbf{Y}$ is all item responses for students within schools, stacked on top of each other. $\Pr(y_{ijm}|\ \theta_{jm(gh)}, \theta_{mh}, C = g, K = h)$ is the measurement model given in Equation 4.1. $g(\theta_{jm(gh)})$ is the student-level achievement distribution for student class g, and $h(\theta_{m(h)})$ is the school-level school effects distribution for school class h. π_h and $\pi_{g|h}$ are as defined above. To maximize the likelihood function in Equation 4.4, use Mplus version 8 is used (L. Muthén & Muthén, 2019).

4.5 Identification

A number of constraints need to be made for identification of the model. I suggest the following constraints based on the original Saltus model (Mislevy & Wilson, 1996; Wilson, 1989) as they ease the interpretation of the parameters (see below). First, one latent class at both the student- and school-levels is selected as the reference class, such that the means of the continuous latent traits, θ_{jm} and θ_m, are fixed to zero for these classes. The difference between the constraints in the MD-IR model and the original Saltus model is that the second school-level latent variable, θ_m, must be considered as well. Second, constraints are needed for the τ_{gha} parameter. One attribute is selected to be the reference attribute, such that the value of τ_{gha} is zero for all latent classes on the reference attribute. These constraints together mean that for the reference class, no attributes have higher or lower difficulty, and for the reference attribute, there are no differences between latent

classes (Jeon, 2019).

4.6 Parameter Interpretation

In order to illustrate an application of the MD-IR model, imagine a hypothetical scenario where we are interested in identifying *low-performing students* within schools who are in need of additional support in mathematics. We are also interested in identifying *low-performing schools* with high proportions of these students. This leads to two latent classes at each level. At the student-level, one class represents students in need of support, and are therefore selected for the intervention, while the other represents higher performing students who do not need support, and are therefore not selected. Similarly, the school classes represent schools in need of support and those schools that do not. In this hypothetical scenario, imagine that the particular mathematics intervention focuses on the cognitive functions used in solving math items, particularly on whether students can set up a math problem based on the given scenario. As such, an assessment like the PISA-D could be used where items are assigned to one of two possible item groups: formulating items and employing/interpreting items.

4.6.1 Latent Classes

Latent classes are defined by the combination of the student-level latent trait and the school-level effect (as in Equation 2.19), as well as the τ parameters discussed in the next section. With two classes at each level, there are four possible combinations of classes for students: (1) low-performing students in low-performing schools, (2) high-performing students in low-performing schools, (3) low-performing students in high-performing schools, and (4) high-performing students in high-performing schools. Figure 4.1 shows these four class combinations. Viewed this way, it is clear that the number of total classes could be considered as $T = G \times H = 2 \times 2 = 4$. In other words, within each of the H school-level classes, there are G student-level classes.

School / Student	Low Performing	High Performing
Low Performing	[1]	[3]
High Performing	[2]	[4]

Figure 4.1: Latent classes for student and school levels

For this hypothetical scenario, we are interested in determining which schools belong to the *low-performing* school class. These schools would be in the left column of Figure 4.1, and the students within these schools would be in classes 1 and 2. Additionally, at the student-level, we are interested in determining which students within schools fall into the *low-performing* student class. These students are classes 1 and 3 in Figure 4.1 (i.e., top row of student classes).

4.6.2 Item Difficulty and Differences by Group

In applications of IRT, the β_i parameter is often interpreted as the item difficulty. To make use of this interpretation for the MD-IR model, it is helpful to see how the β_i and τ_{gha} parameters can be rewritten as a reparameterization of the multilevel mixture Rasch model. Jeon (2018) demonstrates this relationship in the single-level case, showing that the Saltus model can be written as a reparameterization of a single-level mixture Rasch model.

First, note that all the τ_{gha} parameters from Equation 4.1 can be represented as a $T \times A$ matrix. In this illustration, the τ matrix is,

$$\begin{bmatrix} \tau_{(11)1} & \tau_{(11)2} \\ \tau_{(21)1} & \tau_{(21)2} \\ \tau_{(12)1} & \tau_{(12)2} \\ \tau_{(22)1} & \tau_{(22)2} \end{bmatrix}. \tag{4.5}$$

Following the constraints outlined in Section 4.5, one attribute is set as the reference attribute, and as such, all $\tau_{gh1} = 0$ for all latent classes. This implies the assumption that there are no differences in performance between the latent classes on items in the reference attribute group (Jeon, 2019). Additionally, one of the T latent classes is set as the reference class, and $\tau_{(22)2} = 0$ for that reference class. This implies a second assumption, that for the reference latent class, no attributes show higher or lower difficulty. Under these assumptions and constraints, the matrix of τ parameters is,

$$\begin{bmatrix} 0 & \tau_{(11)2} \\ 0 & \tau_{(21)2} \\ 0 & \tau_{(12)2} \\ 0 & 0 \end{bmatrix}. \tag{4.6}$$

Now, to show that τ_{gh} can be interpreted as the difference in item location for student class g and school class h, the reparameterization of the model as the multilevel mixture IRT model is given. Equations 4.7 through 4.10 show the conditional models for all four classes for items in attribute 2 (the non-reference attribute):

$$logit(\Pr(y_{ijm} = 1|\theta_{jm(gh)}, \theta_{m(h)}, C = 1, K = 1) = \theta_{jm(11)} + \theta_{m(1)} - \underbrace{\beta_i + \tau_{(11)2} b_{i2}}_{=\beta^*_{i(11)}}, \tag{4.7}$$

$$logit(\Pr(y_{ijm} = 1|\theta_{jm(gh)}, \theta_{m(h)}, C = 2, K = 1) = \theta_{jm(21)} + \theta_{m(1)} - \underbrace{\beta_i + \tau_{(21)2} b_{i2}}_{=\beta^*_{i(21)}}, \tag{4.8}$$

$$logit(\Pr(y_{ijm} = 1|\theta_{jm(gh)}, \theta_{m(h)}, C = 1, K = 2) = \theta_{jm(12)} + \theta_{m(2)} - \underbrace{\beta_i + \tau_{(12)2} b_{i2}}_{=\beta^*_{i(12)}}, \quad (4.9)$$

$$logit(\Pr(y_{ijm} = 1|\theta_{jm(gh)}, \theta_{m(h)}, C = 2, K = 2) = \theta_{jm(22)} + \theta_{m(2)} - \underbrace{\beta_i + \tau_{(22)2} b_{i2}}_{=\beta^*_{i(22)}}. \quad (4.10)$$

In Equation 4.10, $\tau_{(22)2} = 0$, and therefore, $\beta^*_{(22)} = \beta_i$. For the other three classes, β^*_i is equal to the item location plus a difference, τ_{gh2}, that is unique to each group, (recall that b_{ia} is an indicator variable, and as such, $b_{i2} = 1$ when items belong to attribute two). In other words, β^*_i is the item difficulty. Figure 4.2 shows the item response curve for an item in attribute 2, with differences across the four latent classes. The curve for class 4 that represents the reference group is shown with the solid black line. The additional three classes are shown with colored dashed-lines and each group's corresponding τ_{gh2} parameter is shown as the difference in location for each response curve.

If we imagine that the reference class is high-performing students in high-performing schools, and that attribute two is math formulating items, then Figure 4.2 shows that low-performing students in low-performing schools have a relative disadvantage on formulating items compared to those higher reference students. The other two classes (high-students in low-schools and low-students in high-schools) have a relative disadvantage as well, but it is not as large as the low-performing class. Capturing this difference in item difficulty provides a useful distinction between student and school latent classes that goes beyond multilevel mixture IRT.

4.6.3 Latent Traits and Distribution Parameters

4.6.3.1 School Effect Distributions

School effects mean. Figure 4.3 shows the hypothetical distributions of the student and school latent traits in a format that is similar to Figure 2.1 in the section on HLM. The

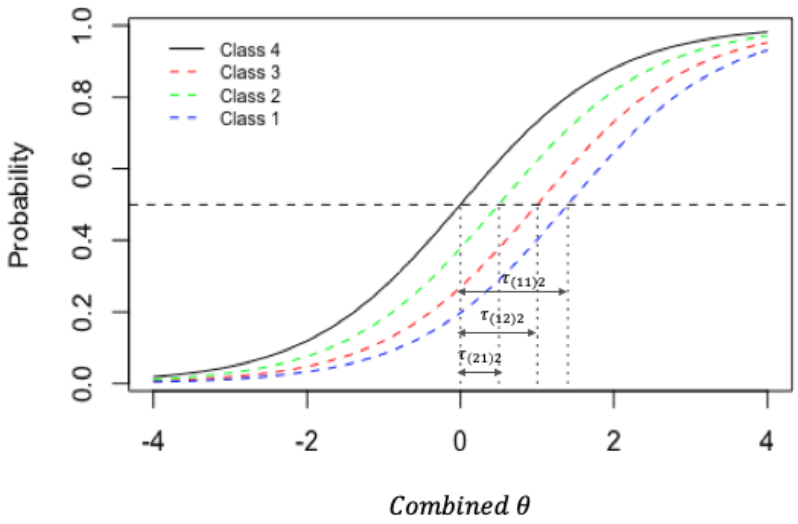

Figure 4.2: Item response curves for an item measuring attribute 2 for four (two at each level) latent classes. The IRCs for the non-reference classes are different from that of the reference class by the corresponding $\tau_{(gh)2}$.

top level shows the distribution for the two school-level latent classes, where the mean of one class, in this case the higher-performing class, is set to zero. The mean of the non-reference group at the school level, μ_1, is the difference in average school effectiveness between schools in need of support and those that are not.

School effects variance. The between-school variance of θ_m, σ^2_m, represents the amount of impact schools have on student math scores. In this example, the (hypothetical) estimated variance for the schools in the low-performing class is $\hat{\sigma}^2_1 = 1$, and the estimated variance for the high-performing school class is $\hat{\sigma}^2_2 = 0.8$. This means that in the low-performing class, the school a student attends matters slightly more than school attendance in the higher-performing class. It is also possible to find the reverse, or no difference between classes. Understanding the difference in impact between classes is highly useful in determining the type of intervention for lower-performing schools. It is also highly possible that whether or not the variance is different by class depends on context, (e.g., low-income countries having more drastic differences across classes).

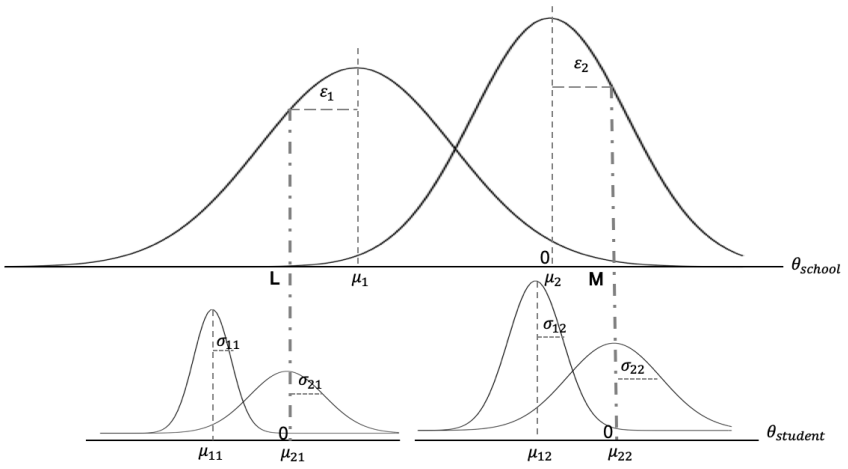

Figure 4.3: Population distributions for school- and student-level latent traits in the MD-IR model. θ_{school} represents school effects, and $\theta_{student}$ represents within-school student achievement scores. One school-level class and one student-level class have a mean set to zero while variances are freely estimated.

4.6.3.2 Student Achievement Distributions

Student achievement mean. The lower level of Figure 4.3 represents the distribution of the student-level latent trait. Within each school class, the higher-performing student-level class has a fixed mean of zero. This means that the interpretation of μ_{11} and μ_{12} is the difference between the means of the high-performing students and the low-performing students, within the low-performing school class and the high-performing school class, respectively.

Student achievement variance. The within-school variance of θ_{jm} represents the academic gap within a school. Continuing with the hypothetical example, imagine σ^2_{11} (the student-level variance for student class one and school class one) is estimated as 0.6, and $\sigma^2_{1,2}$ (the student-level variance for student class two and school class one) is estimated as 1. This would mean that the within-school gap is smaller for low-performing students in low-performing schools than it is for the same class of students in high-performing schools.

4.6.4 Class proportions

The prior probabilities of class membership in mixture models are often referred to as mixing proportions. π_h in Equation 4.2 can be interpreted as the proportion of schools in a particular school class. In the hypothetical example, if $\pi_{h=1} = .30$, then 30% of schools are classified into the low-performing school class. $\pi_{g|h}$ from Equation 4.3 can be interpreted as the proportion of students in a particular class, conditional on school class. For example, if $\pi_{g=1|h=1} = .40$, and $\pi_{g=2|h=1} = .60$, then 40% of students in low-performing schools are classified as low-performing, while 60% of students in the same school class are classified as high-performing.

Further examples of parameter interpretations are provided in Chapter 5.

4.7 Student and School Scores and Class Probabilities

Providing individual student and school classification probabilities is essential for the diagnostic use of the MD-IR. The estimation of the probability of belonging to each class also provides confidence in the classification of students and schools. To obtain latent class probabilities, maximum a posteriori (MAP) is used (e.g., Dias & Vermunt, 2008; B. O. Muthén & Muthén, 2010).

Individual student and school scores, θ_{jm} and θ_m, can also be estimated using the MD-IR. To obtain student- and school-level scores, an empirical Bayes prediction method, often termed expected a posteriori (EAP) in IRT modeling, is used. Interpretations of these latent traits is essentially the same as in multilevel IRT. In other words, θ_{jm} is student j's overall math achievement. θ_m is school m's effect, or contribution, to student scores. When no covariates are included in the model, the school effect includes both context and school practice effects.

4.7.1 Diagnostic Feedback

The goal of obtaining individual school and student classifications and scores is to provide diagnostic feedback to education leadership. This is possible at multiple levels using the MD-IR model. At the district level, state level, or possibly even the country level, the leadership is likely to be interested in the estimated model parameters, particularly the proportion of schools and students classified in the *low-performing* categories. Leadership will also be interested in whether there are differences in item difficulty for certain groups on important aspects of the curriculum. Beyond the model parameters, leaders who aim to identify schools for performance-targeted interventions will need to determine which schools are classified as *low-performing*, and will therefore need individual school classifications.

Leadership at the school level is more likely interested in the specific results for their own school. Specifically, they will want to know to what class their school was assigned, and with what probability. They would also be interested in the results of continuous latent school effect because it provides information on the school's standing within the school class. At the student-level, schools would be interested in the within-school classification of students at their own school. This would allow school leadership to identify students in need of additional support and target relevant programs and support.

4.8 Chapter Conclusion

The Multilevel Diagnostic Item Response model offers a unique approach to studying schools and their impact on student learning. The MD-IR is multilevel, which allows for both student-level and school-level information. It also includes a categorical latent variable for student and school classifications, as well as a student and school continuous latent variable. By incorporating the continuous latent traits, the model is able to provide estimates of these variance components, giving information on the impact of schools in different classes, as well as the achievement gaps within schools in different classes.

Finally, by incorporating item attribute information, the model classifies students based on important aspects of the curriculum, as well as on the more general domain of interest.

CHAPTER 5

Empirical Example: PISA for Development

5.1 Introduction

The goal of this chapter is to provide an example of the application of the Multilevel Diagnostic Item Response model in the context of international education development. Specifically, this chapter demonstrates how the key components of the MD-IR model that were discussed in Chapter 4 can be used to bring together the study of school effects, performance-targeted interventions, and mathematical literacy. The example provided in this chapter uses data from the PISA for Development (PISA-D) assessment that was administered in Cambodia.

The first key component of the MD-IR model is that it incorporates a continuous, normal latent variable at the school level that represents the school effect, and its variance represents the heterogeneity in the impact schools have on student achievement. Studying school effects in the context of international development is important because, as mentioned in Chapter 1, schools are often found to play an even larger role in student outcomes in low-income countries compared to those in higher income countries (Chudgar & Luschei, 2009). This means that not only is it simpler for governments to target schools for interventions, but it is also an effective method of improving student outcomes. By focusing on schools as the target level of intervention, countries can improve equity in the outcomes that are demonstrated in Figure 1.1.

The second key component of the MD-IR model is the incorporation of categorical latent classes at the student and school levels. This component is key for identifying

which schools will participate in performance-targeted interventions in order to improve student outcomes, and is particularly important in the context of international development because resources are often scarce for implementing universal interventions. The confirmatory nature of the latent classes allows for definitions of the classes a priori. In this example, we will be looking to identify low-performing schools in Cambodia, and low-performing students within those schools, for participation in a hypothetical mathematics-based intervention. Using this performance-based approach allows for us to target best the schools that are likely to see the most improvement from participation. While the PISA-D assessment is not designed for individual scoring, the identification aspect of the model will be presented in this example for illustrative purposes.

The third component of the MD-IR model is the use of attribute information provided by item developers and the estimation of differences in item difficulty. This information is what allows us to study differences in mathematical literacy based on defined item groups. In this example, it allows us to differentiate latent classes based on differences in performance on the first of the three cognitive processes: formulate, employ, and interpret. This illustration focuses on formulate items for two reasons. First, formulating a math problem from a real world situation is the first step in problem solving. Second, formulate items were the most difficult for students in the field test trials across multiple countries (Stacey, 2015). More information on this mathematical process is discussed in the following section.

To bring these three components together, we will imagine a hypothetical scenario where we are aiming to identify low-performing schools that would benefit from additional support in the area of formulating mathematical problems in real life situations. This chapter is organized as follows. First, I review the details of the PISA-D math assessment, as well as demographic and descriptive performance statistics for students and schools in the sample. Then, I present the specific research goals used to guide the analysis. Next, I outline the methods, which are followed by the presentation of the results from the MD-IR model. Then, I present some additional analyses to support the

conclusions drawn from the MD-IR results. Lastly, I close the chapter with a discussion of implications and limitations in this specific analysis.

5.2 Data: PISA-D Mathematics Assessment

5.2.1 Background

As discussed previously, the PISA-D initiative was designed to provide education leaders in low-income countries with more relevant information by creating a more accessible test. The PISA-D assessment incorporates more easy test items than the original PISA that was designed for OECD countries, with 60% of test items being considered "easy". Seven participating countries have available data, but this example will focus on Cambodia. The data include tests of 15 year-old students who are still enrolled in school, as well as students who are no longer attending school. The out-of-school data are essential for understanding achievement in these countries since many students drop out of school prior to age 15. However, these data are not relevant for understanding school performance and are therefore not used in the present study. Three subjects are assessed with PISA-D: mathematics, reading, and science, but this example will focus on the mathematics assessment.

Student and school background questionnaires are also included. For the purposes of this analysis, student socioeconomic status, student urbanicity, school resource status, and school urbanicity are used and follow the definitions defined by PISA-D. In all PISA assessments, student socioeconomic status (SES) is referred to as Economic, Social and Cultural Status (ESCS) and is a composite based on information regarding home possessions (e.g., devices and books in the home) and other measures of home environment gathered from the the student questionnaire (OECD, 2019). The measure in the original PISA does not adequately capture lower levels of SES typically found in low-income countries. Therefore, the PISA-D measure of student SES was adapted to provide additional levels of poverty and comprises four levels: extremely poor, severely poor, poor, and not

poor. School resource status is similarly a composite based on information reported by principals on the school questionnaire regarding school infrastructure, school facilities, and the availability of instructional resources. The measure has five levels: extremely low, severely low, low, moderate, and high. School and student urbanicity are dichotomous variables indicating rural or urban status. The purpose of including these demographics is to understand better the students and schools that are identified as low-performing, and to consider the relationship between student populations and mathematics achievement as discussed in Chapter 1. In this case, we expect to see differences in math performance, as well as differences in latent classifications, based on these demographics sub-groups.

5.2.2 Mathematics Assessment

The PISA-D mathematics assessment is a unidimensional math assessment that aims to measure students' mathematics literacy. PISA defines mathematical literacy as "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" (OECD, 2018, p. 51). The three key verbs, "formulate", "employ", and "interpret", are considered the three processes that students engage in to be problem solvers in math (OECD, 2018). The emphasis in the definition of mathematics literacy is on engaging in mathematics through these different processes, yet established school effects approaches do not differentiate performance in these areas. The formulating process refers to how well students are able to formulate a stated problem in a mathematical form. The employing process refers to the students' abilities to perform the appropriate computations and manipulations to arrive at the correct solution. The interpreting process refers to how well students can interpret the solution in the real-world context. True mathematical literacy requires the ability to utilize all three of these mathematical processes (OECD, 2018). Figure 5.1 shows a sample formulate item from PISA-D. The task in this item involves determining the correct method for changing Singapore dollars to South African rand, given a specific

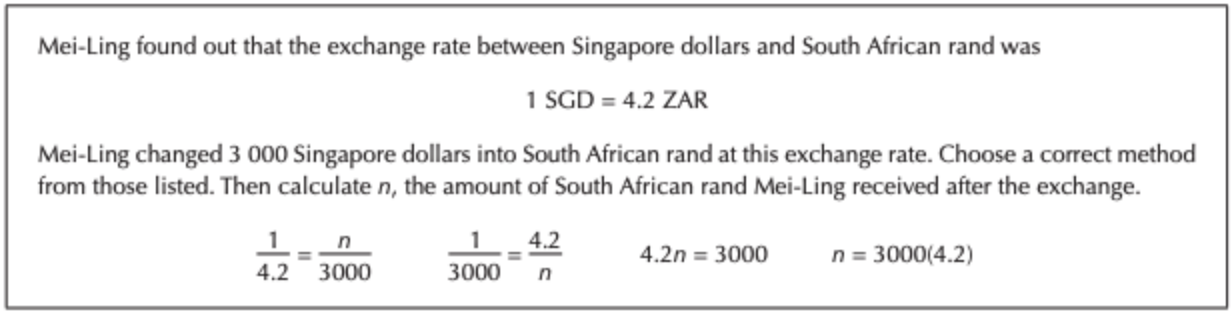

Figure 5.1: Sample PISA-D mathematics item that includes the formulate cognitive process in the first step.

exchange rate. This first formulating step is distinct from the calculation in the second step that would be considered part of the employing process. Determining the correct approach to solving a problem such as this one is an essential step in solving other real world math problems.

In PISA-D, the goal is to include math items of which 50% are employ items, 25% are formulate items, and 25% are interpret items. Each item is assigned to only one category. There are a total of 62 possible math items across 8 different booklets. Items are matrix sampled across students and administered via paper and pencil. Items are primarily dichotomously scored, with some polytonomous items included. For this analysis, all polytonomous items are re-scored as binary (full credit or no credit) in order to demonstrate the context outlined in Chapter 4.

5.2.3 Sample Data and Demographics

As mentioned, school context in Cambodia differs greatly from that of OECD countries. Before differentiating schools and students based on student math performance, it is useful to understand better this context for schools and for students.

Overall, the students taking the PISA-D assessment in Cambodia are low-SES, as indicated by poverty levels of extremely poor, severely poor, or poor. Student sample statistics are presented in Table 5.1. There are 3,225 students in 168 schools taking the

Table 5.1: Student sample size and demographic counts and proportions for students taking the PISA-D mathematics assessment in Cambodia.

	N	Proportion
Total Sample Size	3225	1.00
Gender		
Female	1709	0.53
Male	1516	0.47
Urbanicity		
Urban	951	0.29
Rural	2274	0.71
Poverty Level		
Extremely Poor	95	0.03
Severely Poor	1436	0.45
Poor	1047	0.32
Not Poor	647	0.20

PISA-D math assessment. Fewer students are considered extremely poor ($p = 0.03$), but large proportions are considered severely poor ($p = 0.45$) and poor ($p = 0.32$). Only 20% of students tested in Cambodia are considered not poor ($p = 0.20$). A small majority of 15 year-old students are female ($p = 0.53$), and the majority live in rural locations ($p = 0.71$).

Similarly, the majority of schools are low- to extremely-low resourced schools, with only approximately a quarter of schools having moderate or high levels of resources ($p = 0.19$ and $p = 0.07$, see Table 5.2). Most schools are located in rural areas ($p = 0.73$), which typically have fewer resources than their urban counterparts. Not only do schools differ on school resource levels, but they also differ on the proportion of students within schools that have high levels of poverty. The average level of extremely poor students within schools is small ($mean = 0.03$), but varies across schools ($SD = 0.07$). On the other extreme, the average level of students who are not poor within a school is larger ($mean = 0.18$), but varies even more widely from school to school ($SD = 0.22$).

The average total school size is also included in Table 5.2. This number represents the average number of students within schools as reported by principals on the school questionnaire. The average school size is approximately 1,029 students, with some schools as small as 17 students, and others as large as 5,111 students. Schools across these different contexts are likely to differ in their levels of impact on student achievement scores, as well as differ in their effectiveness of instruction on the cognitive processes.

For all PISA studies, including PISA-D, a two-stage stratified sampling design is used. In the first stage, schools are sampled from a sampling frame with probabilities proportionate to their size (OECD, 2019). Students are then sampled from within these schools in the second sampling stage. The average sample size per school in this analysis is approximately 19 students. The largest school sample is 29 students, while the smallest is 1 student. Very few schools (7 schools) have fewer than 5 students. Selection probabilities vary in the sampling process, and non-response exists at both the school and student levels (OECD, 2019). Therefore, the sampling weights that are provided with the PISA-D data are incorporated throughout the analysis at both levels. Incorporating sampling

Table 5.2: School sample sizes and demographic statistics for schools assessed with the PISA-D mathematics assessment in Cambodia. Top panel includes sample N counts and proportions for school demographics. Bottom panel includes means and standard deviations of within-school student poverty proportions, full school size, and sample size within schools.

	N	Proportion
School Sample Size	168	1.00
Urbanicity		
Urban	45	0.27
Rural	123	0.73
Resource Level		
Extremely Low	47	0.28
Severely Low	40	0.24
Low	37	0.22
Moderate	32	0.19
High	12	0.07
	Mean	SD
Within School Poverty		
Extremely Poor	0.03	0.07
Severely Poor	0.47	0.25
Poor	0.31	0.17
Not Poor	0.18	0.22
Total School Size	1028.72	964.66
Sample Size Within School	19.2	7.31

Table 5.3: Example Schools A, B, and C demographic information and average proportion correct on PISA-D all math items and on items from the formulate cognitive process.

	School A	School B	School C
N sample	19	19	25
School Size	655	669	256
Urbanicity	Urban	Rural	Urban
Resource Level	Extremely Low	Low	Moderate
Total Proportion Correct			
Mean	0.15	0.27	0.57
SD	0.19	0.17	0.16
Rank	151	75	2
Formulate Proportion Correct			
Mean	0.10	0.11	0.44
SD	0.21	0.19	0.30
Rank	72	62	2

weights allows for the model parameters to be interpreted as estimates of the student and school populations in Cambodia.

Due to the complex sampling procedures, the PISA-D data are not appropriate for individual school scores, particularly in a high-stakes context. However, for the purposes of illustration, we will explore the performance of three example schools that have larger samples in the data. Table 5.3 shows demographic information for these example schools, Schools A, B, and C, that we will follow throughout this example. School A is a large, urban school with extremely low resources. School B is a large, rural school, with low resources. School C is a medium-sized, urban school with moderate resources. The details of the performance of these three schools will be discussed in the following section.

Table 5.4: Item statistics for each PISA-D mathematics cognitive process—employ, formulate, and interpret—and the full mathematics assessment.

	Employ	Formulate	Interpret	All Items
Number of Items	28	12	22	62
Number Per Student				
Mean	10.49	4.53	8.48	23.49
SD	3.69	2.54	3.25	7.91
Avg Proportion Correct	0.27	0.15	0.35	0.29
$>$ .50 Correct	0.11	0.00	0.23	0.13
$\leq$ .20 Correct	0.29	0.67	0.18	0.32

5.2.4 PISA-D Math Item Statistics

In Cambodia, 8 of the possible 12 math booklets were administered to students, with 62 overall math items: 28 employ items, 12 formulate items, and 22 interpret items. The number of items is not equal for all students. Individual students saw an average of 23.49 items overall, 10.49 employ items, 4.53 formulate items, and 8.48 interpret items. Item information and statistics are presented in Table 5.4.

All items are very difficult. Consistent with field test findings, the formulate items are the most difficult. The average proportion correct in the sample for individual items is 0.29 across all items, but only 0.15 for formulate items. On the other hand, the average item proportion correct was 0.27 and 0.35 for employ and interpret items, respectively. Very few items have more than 50 percent of students responding correctly. In fact, 0 percent of items in the formulate cognitive domain had more than 50 percent of students responding correctly, compared to 11 percent of employ items and 23 percent of interpret items. Moreover, 67 percent of formulate items had less than 20 percent of students responding correctly. Clearly, formulate items are particularly difficult for students. This is especially concerning because formulating situations into mathematical problems is an essential first step for utilizing math in real world situations.

5.2.5 Student Performance

Individual student scores show the same patterns in terms of difficulty by domain. Table 5.5 shows average student scores (as the proportion correct) for the total math score, as well as by each cognitive process. Individual student proportions are calculated as the number correct divided by the number of items for each process that the student was administered. The average student scores that are presented in Table 5.5 are weighted averages of these individual student proportions, weighted based on the total student weight that includes the probability of school and student sampling weights for each individual student (OECD, 2019). This means that the student level rows of Table 5.5 can be interpreted as the estimated mean proportion correct (and its standard deviation) for students in Cambodia. The assessment is very difficult. The average total proportion correct for all students is very low ($p = 0.28$), but with a sizable spread in scores ($SD = 0.18$). Formulate items are by far the most difficult ($p = 0.16$) and has the largest spread of scores ($SD = 0.24$) of any of the other cognitive processes. Interpret items are the easiest ($p = 0.36, SD = 0.24$), and employ items are in the middle ($p = 0.28, SD = 0.21$). These large standard deviations indicate that there are students performing at levels far below their peers, and that they could benefit from additional support.

On average, student performance differs based on urbanicity and poverty levels. Students in urban schools outperform their rural peers on all cognitive processes by a sizable amount. Overall, urban students have an average proportion correct of 0.36 ($SD = 0.19$), while rural students have an average proportion correct of 0.25 ($SD = 0.17$). The standard deviation of the employ and interpret processes are similar for urban and rural students, but larger for urban students on the formulate process. Students who are not poor also outperform their peers who are considered poor, severely poor, or extremely poor in all processes (see Table 5.5). This pattern of performance across these demographic subgroups is consistent with subgroup differences in other low-income countries (e.g., Willms, 2010).

Table 5.5: Weighted average proportion correct on the three cognitive processes and the full assessment. Top portion includes student-level weighted means and standard deviations for the full sample and by student urbanicity and poverty. Bottom portion includes school-level weighted means and standard deviations for the full sample and by school urbanicity and resource level.

	Employ		Formulate		Interpret		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Full Student Sample	0.28	0.21	0.16	0.24	0.37	0.24	0.29	0.18
Student Urbanicity								
Urban	0.36	0.21	0.24	0.28	0.47	0.23	0.37	0.18
Rural	0.26	0.21	0.13	0.21	0.34	0.23	0.26	0.17
Student Poverty								
Extremely Poor	0.18	0.15	0.09	0.19	0.27	0.17	0.18	0.12
Severely Poor	0.25	0.19	0.13	0.21	0.33	0.23	0.25	0.16
Poor	0.29	0.21	0.16	0.24	0.39	0.24	0.30	0.19
Not Poor	0.36	0.22	0.26	0.28	0.48	0.24	0.38	0.20
Full School Sample	0.24	0.10	0.12	0.11	0.31	0.12	0.24	0.10
Within School SD	0.17	0.05	0.17	0.09	0.19	0.05	0.13	0.04
School Urbanicity								
Urban	0.35	0.10	0.25	0.13	0.46	0.11	0.37	0.10
Rural	0.22	0.08	0.10	0.08	0.28	0.10	0.21	0.08
School Resource Level								
Extremely Low	0.23	0.09	0.12	0.08	0.30	0.09	0.23	0.08
Severely Low	0.20	0.06	0.08	0.08	0.24	0.10	0.19	0.07
Low	0.23	0.08	0.10	0.08	0.33	0.11	0.24	0.08
Moderate	0.29	0.10	0.18	0.11	0.38	0.12	0.30	0.10
High	0.44	0.13	0.34	0.18	0.55	0.11	0.44	0.12

5.2.6 School Performance

Scores continue to reflect this pattern when aggregated at the school level. As with the student averages, school averages are weighted averages based on school sampling weights provided by PISA-D (OECD, 2019). The average school mean was 0.24 ($SD = 0.10$) for all items and 0.12 ($SD = 0.11$) for formulate items. Figure 5.2 shows histograms of the total score and the scores for each cognitive process. The school mean total score is shown in the top left histogram. School mean scores are quite low overall, with a small number of schools having an average above 0.50. School means for employ items are shown in the top right histogram and show a similar pattern to the overall mean. Mean interpret scores are shown in the bottom right histogram. These items were somewhat easier for students. Finally, mean formulate scores are shown in the bottom left histogram of Figure 5.2. This histogram reflects the challenging nature of these items as it is skewed positive with clear floor effects. It is possible that there are qualitative differences between the schools who are getting scores of zero, or essentially zero, and schools who are getting higher scores. It is also possible that there are differences in item difficulty for students in these lowest schools compared to the higher schools. By using the MD-IR model, we are able to model these potential differences and to focus the analysis on the formulating process.

School averages also differ based on school demographics. Urban schools have an average school mean of 0.36 ($SD = 0.10$) compared to the rural average school mean of 0.24 ($SD = 0.08$). Schools with high levels of resources also outperform schools at all lower levels of resources, on average (see Table 5.5). Schools at higher levels of resources have a larger standard deviation ($SD = 0.12$), such that some of these high-resourced schools perform at levels similar to schools with fewer resources. While the standard deviations for lower-resourced schools are not quite as large, it is still possible that some of these schools perform at higher levels more similar to that of the higher-resourced schools. Therefore, in determining how to select schools for participation in a specific instructional intervention, school SES may not be the best classification criteria. Using a model such as the MD-IR for selection based on performance may be a better selection

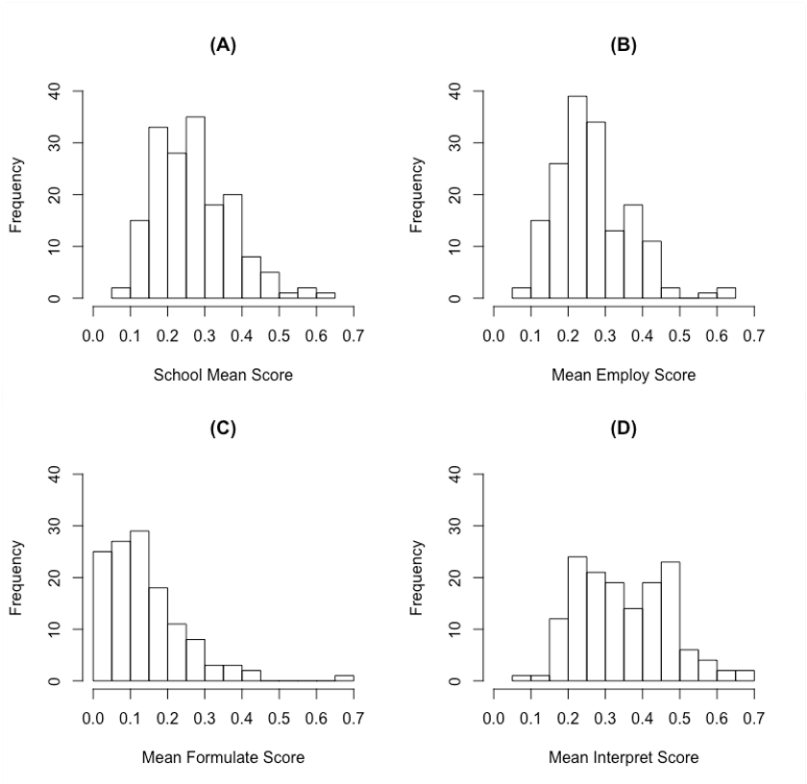

Figure 5.2: Histograms of school average proportion correct for full PISA-D math assessment and each cognitive process. (A) Full math average proportion correct. (B) Employ items average proportion correct. (C) Formulate items average proportion correct. (D) Interpret items average proportion correct.

approach when improved performance is the goal.

The within-school standard deviations are also presented in Table 5.5. These show variation of student scores within schools by each domain and overall. Within schools, interpret scores show the largest variation with the mean standard deviation of 0.21. Formulate and employ scores are slightly lower at 0.19 and 0.18, respectively. For employ, interpret, and total scores, the spread of the average standard deviation is not large. The spread for formulate scores, however, is larger with a standard deviation of 0.09. Large differences on scores within schools indicates that there are large within-school gaps between the high- and low-performers. This suggests that there are larger differences in within-school gaps for the formulate process than for the other processes or for the total score.

The left panel of Figure 5.3 shows a scatterplot of average school scores plotted against the within-school variance. The correlation between overall score and the association within variance is 0.59. As shown in the scatterplot, schools with lower scores tend to have a lower within-school variance, likely due to the floor effects associated with the low scores. The impact of the floor effects on the variance is clear in the right scatterplot of the mean formulate scores by within-school formulate variance. Here, the zero, or almost zero, scores have essentially no variance, while the higher scores have a more sizable variance. The MD-IR model is able to capture these meaningful differences by estimating unique variances across both student and school latent classes.

5.2.7 Schools A, B, and C

Descriptive results clearly indicate that there are schools and students that perform at lower levels compared to their counterparts. For Schools A, B, and C, we would like to determine whether these schools are low-performing schools that are in need of additional support on the formulating process. We would also like to provide these schools with clear information regarding their relative performance to other schools, as well as information on their within-school student performance.

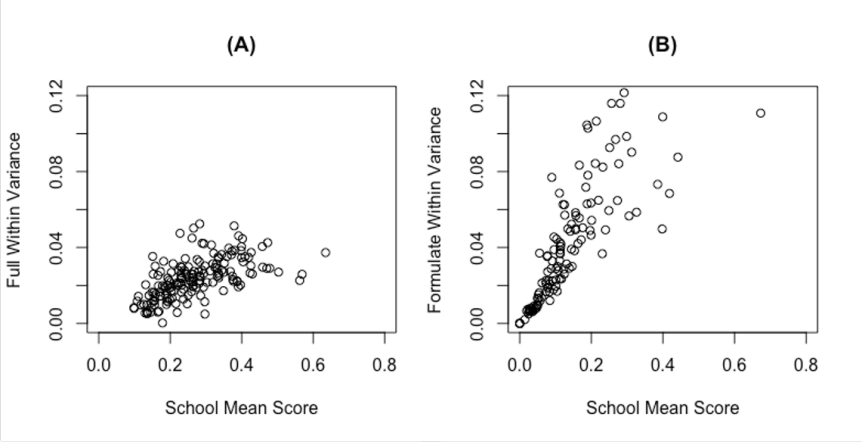

Figure 5.3: Plots of school mean scores and associated within-school variance. (A) School mean total score by school mean total within-school variance. (B) School mean formulate score by school mean formulate within-school variance.

Based on these descriptive results, School A performs substantially below the sample average, with a total average proportion correct of 0.15 ($SD = 0.19$). It is ranked 151st out of all 168 schools on the total assessment (see Table 5.3). However, when looking at the average formulate proportion correct, School A performs better. On the formulate process, School A has an average proportion correct of 0.10 ($SD = 0.21$) that is only slightly below the sample mean and has a ranking of 72. School A provides an example of an urban school that does not perform as highly overall as its other urban counterparts, but is somewhat stronger on the formulate process, relative to other schools.

School B is a rural school that has an overall performance that is slightly above the sample average, with a total average proportion correct of 0.27 ($SD = 0.12$) and a ranking of 129th. When looking at the proportion correct on formulate items, School B is also close to the sample average, with an average formulate proportion correct of 0.11 ($SD = 0.19$). School B provides an example of a school that is in the middle overall, as well as on the formulate process.

School C, also an urban school, has a high total average performance of 0.57 ($SD = 0.16$) and a high formulate average score of 0.44 ($SD = 0.30$). The spread of scores in

School C is the highest of the three example schools, particularly so in the formulate process. School C also has the 2nd highest average performance on both the total assessment and on the formulate process. School C provides an example of a high-performing school that also has a high spread of students in the formulate process.

The performance of these three schools follows the trend of higher-resourced schools performing higher than lower-resourced schools, but not necessarily the trend based on urbanicity. The proportions correct do provide some insight into the performance of these schools, but do not provide direct classifications for either students or schools.

5.2.8 Summary

To summarize, while all items are difficult for students, formulate items are the most difficult across all student and school demographic groups. Consistent with other education research in low-income countries, urban and high-resourced schools outperform rural and lower-resourced schools, on average. However, this does not tell the full story of school performance, as some higher-resourced or urban schools may perform at a lower level similar to their lower-resourced or rural counterparts. For this reason, using the descriptive results from this section alone makes it difficult to identify schools for participation in the hypothetical intervention. Finally, within-school student performance shows some differences across schools, as indicated by spread on the within-school standard deviations of student performance. Understanding how student performance within low-performing schools is different from their high-performing counterparts offers useful insights for the implementation of the formulate intervention.

5.3 Research Goals

The overarching goal of this analysis is to provide an example of how the MD-IR can be used to identify low-performing schools in Cambodia that are in need of support on the formulate cognitive process. The specific research goals that guide this analysis are:

1. To identify low-performing schools that are most in need of support on the formulate cognitive process.

2. To understand better these low-performing schools in terms of within-school student performance, compared to high-performing schools.

3. To understand better these low-performing schools in terms of school context and student demographics, compared to high-performing schools.

To address these goals, we will use the MD-IR model to identify low-performing schools based on the formulate items of the PISA-D assessment. In addition to these three research goals, we will continue to follow the three schools, Schools A, B, and C, throughout the analysis for the example of identification. We will be looking to see which of these schools would be identified as most in need of support and likely to benefit most for a performance-targeted intervention. The following section discusses the methods used to achieve these research goals.

5.4 Methods

Prior to applying the MD-IR model to address the research goals, an additional analysis is conducted that applies the standard, existing approach for studying school effects. Specifically, the first analysis applies a multilevel IRT model in order to discuss the differences between the existing approach and the proposed MD-IR model. The details for the multilevel IRT analysis and the MD-IR analysis are discussed below.

5.4.1 Multilevel IRT

5.4.1.1 Model and Assumptions

The first model fit is the multilevel Rasch model that is presented in Chapter 2, Equation 2.20. Recall that in this model, the probability of a correct response is conditional on

student ability (θ_{jm}) and school effectiveness (θ_m), and is a function of item difficulty (β_i). A key assumption in this model is that the school effect and student math achievement are assumed to be made up of one distribution for all schools and one distribution for all students, respectively. Because of this assumption, the analysis based on the multilevel Rasch model does not directly address the research goal of identifying low-performing schools. Additionally, the traditional multilevel Rasch model does not incorporate item information, meaning that we cannot differentiate schools and students based on the formulate cognitive process. Instead, the goal of the multilevel Rasch analysis is to provide useful information regarding the overall distribution of school effects and how this effectiveness varies across schools. It also provides estimates of the distribution of student overall math achievement, giving us insight into student performance.

5.4.1.2 Estimation

Estimation of model parameters is conducted using full-information maximum likelihood estimation with robust standard errors in Mplus version 8 (L. Muthén & Muthén, 2019). Estimated scores for student ability and school effectiveness are estimated as expected a posteriori (EAP) scores, as is commonly done in many IRT analyses (B. O. Muthén & Muthén, 2010). Sampling weights are included for each level of the analysis, meaning both the student-level weight and the school-level weight are used. The purpose of including sampling weights is to allow for the interpretation of the model parameters as estimates of the student and school population in Cambodia.

5.4.2 MD-IR

5.4.2.1 Model and Assumptions

The application of the MD-IR model follows the formulation discussed in Chapter 4 in Equation 4.1. To address the first research goal, two latent classes are incorporated at the school level, defined a priori as low-performing and high-performing school classes.

Each of these latent classes has its own distribution of school effects, with the mean of the reference class fixed to zero. The reference school class is the low-performing school class, meaning that the estimated mean of the high-performing school class is the estimated difference in average school effect between the two school-level latent classes. The variance is estimated for both school classes.

For the second research goal, two latent classes are incorporated at the student level. The combination of these student and school classes results in four possible student-level classes, as presented in Figure 4.1. The four latent classes are: (1) low-performing students in low-performing schools, (2) high-performing students in low-performing schools, (3) low-performing students in high-performing schools, and (4) high-performing students in high-performing schools. The means of the two reference classes, the low-performing schools in each school class, are fixed to zero. The variance is estimated for all four student-level latent classes.

Since our goal in this example is to identify low-performing schools and low-performing students that are differentiated based on formulating items, the τ parameters that are incorporated in the MD-IR model are specified to represent the difference in item difficulty on formulating items between a reference latent class and the other latent classes. This is done (1) by specifying two item groups a priori that are defined using existing PISA-D item information, and (2) by specifying a student-level reference class. The first of the two item groups, the target item group, is composed of the formulate items. The second of the item groups, the reference item group, is composed of the other items (i.e., employ and interpret items). The student-level reference class is the class that is the reference class at both school and student levels. In this case, the reference class is the low-performing student class within the low-performing school class. Recall from Chapter 4 Section 4.6.2 that a number of assumptions are specified for the τ matrix. Based on these assumptions, the model is specified such that there are assumed to be no differences on non-formulating items for any of the latent classes. This assumption ensures that we are differentiating the latent classes based on only the formulating items,

which is of interest for our hypothetical intervention. The second assumption is that there are no differences on formulate items for low-performing students in low-performing schools compared to the other items. This specification allows us to estimate the τ parameters for the other three classes as the difference in item difficulty, or the advantage, that students in the higher-performing classes have on formulate items, compared to the lowest-performing class.

Finally, to address the third research goal, a post-hoc analysis is conducted where school resource levels and urbanicity are compared across high- and low-performing school classes that are determined from the MD-IR model. Student demographics for poverty and location are also compared across within-school latent classes, providing more insight into the context of the low-performing schools.

5.4.2.2 Estimation

Like the multilevel Rasch model, full information maximum likelihood estimation is conducted for parameter estimation. Details regarding the likelihood of the observed data for the MD-IR model can be found in Chapter 4 Section 4.4. The MD-IR analysis is also conducted in Mplus version 8 and uses a modified EM algorithm for numerical integration with 15 integration points. For estimating the MD-IR, as in any mixture modeling analysis, multiple starting values are used in order to avoid local maxima solutions. Specifically, the analysis is conducted with 400 random starting values, with 50 optimizations at the final stage. Estimated scores at both the student and school level are also EAP scores. Latent class posterior probabilities are estimated, and individual class membership is based on the most likely class. Student-level and school-level sampling weights are included, again with the goal of interpreting model parameters as estimates for the student and school population in Cambodia.

5.4.2.3 Model Fit

Assessing whether the MD-IR shows reasonable fit to the data presents a challenge. Currently, there are no established goodness-of-fit measures for complex latent variable models with non-continuous data. In order to provide some evidence of reasonable fit, two approaches are taken here. The first approach is to assess the relative fit by comparing relative fit indices with competing, baseline models. Three fit indices are used: Akaike information criterion (AIC), Bayesian information criterion (BIC), and sample-size adjusted BIC. Two models are selected as possible competing models. The first competing model is the multilevel Rasch model. This model is selected because it is a well-established model, and is the standard approach within an IRT framework for exploring school effects. Additionally, the multilevel Rasch model can also be thought of as a single-class MD-IR, where there is only one latent class at each level, and the item group parameters are equal to zero. The second competing model is a multilevel mixture IRT model that is discussed in Chapter 3 Section 3.1. Comparisons to this model are useful because it offers even more flexibility than the MD-IR model by estimating different item difficulty parameters for each latent class.

The second approach is to assess absolute fit of the MD-IR model by assessing its predictive validity. To do this, simulated datasets similar to the original data are generated based on the estimated parameters of the MD-IR model. Then using these datasets, key aspects of the original data are compared to those of the replicated datasets. For example, since we are interested in school-level scores on the full assessment and on the formulate items, the average proportion correct for both in the original data is compared to each replicated dataset. Small differences between these values indicate reasonable fit of the MD-IR to these data. This approach is comparable to Bayesian posterior predictive checking (Jeon, De Boeck, Li, & Lu, 2020) and similar approaches have been used in other maximum likelihood contexts (e.g., Jeon, De Boeck, & van der Linden, 2017; C. Wang, Fan, Chang, & Douglas, 2013).

5.5 Results

5.5.1 Multilevel Rasch Analysis

School effects. Results from the multilevel Rasch model indicate that schools vary in their impact on student learning ($\hat{\epsilon} = 0.432$, $SE = 0.06$). Estimates for the full sample and scores by student and school demographics are presented in Table 5.6. The right plot of Figure 5.4 shows the school effects distribution. These results indicate that there are less effective schools in terms of overall mathematics instruction, and that some schools would benefit from additional support. However, this approach does not directly identify which schools should participate in an intervention. If this standard approach is used to select schools for a performance-based intervention, a threshold would need to be applied to this distribution to determine participation among schools. Additionally, this school effect score represents school impact on overall math achievement, but does not differentiate based on formulate items.

Additionally, school effectiveness in mathematics does vary by context. Urban schools are somewhat more effective than rural schools, and highly-resourced schools are more effective than lower-resourced schools, on average. However, the standard deviations of both groups indicate that some schools in these categories are performing at levels lower than that of their rural and lower-resourced counterparts, making these demographic categories imprecise in terms of selected schools for performance based interventions.

Student math achievement. Results also indicate that there are differences within schools on math achievement. Within-school differences are moderate as shown by the estimated within-school variance ($\hat{\sigma}^2 = 0.762$, SE = 0.04). The left plot of Figure 5.4 shows the distribution of within-school student math achievement across all students. In this model, the estimate of within-school variance is the same across the full school population, meaning that it is assumed that this level of within-school variance is true across all schools. However, based on the scatterplot in Figure 5.3, schools with higher average scores may actually have slightly larger within-school variances. Consistent with the

Table 5.6: Multilevel Rasch distributional estimates of student math achievement and school effects. Left portion shows student math achievement for the full assessment and by student urbanicity and poverty level. Right portion shows school effects for the full sample and by school urbanicity and resource level.

Student-Level Math Achievement			School-Level School Effect		
	Mean	Variance		Mean	Variance
Full Sample	0	0.762	Full Sample	0	0.432
		(0.04)			(0.06)
Urbanicity	Mean	SD	Urbanicity	Mean	SD
Urban	0.026	0.755	Urban	0.782	0.565
Rural	-0.011	0.684	Rural	-0.005	0.521
Poverty Level			Resource Level		
Extremely Poor	-0.228	0.584	Extremely Low	0.044	0.540
Severely Poor	-0.039	0.683	Severely Low	-0.208	0.477
Poor	0.018	0.711	Low	0.276	0.526
Not Poor	0.108	0.754	Moderate	0.590	0.613
			High	0.981	0.561

Note: Standard errors are in parentheses.

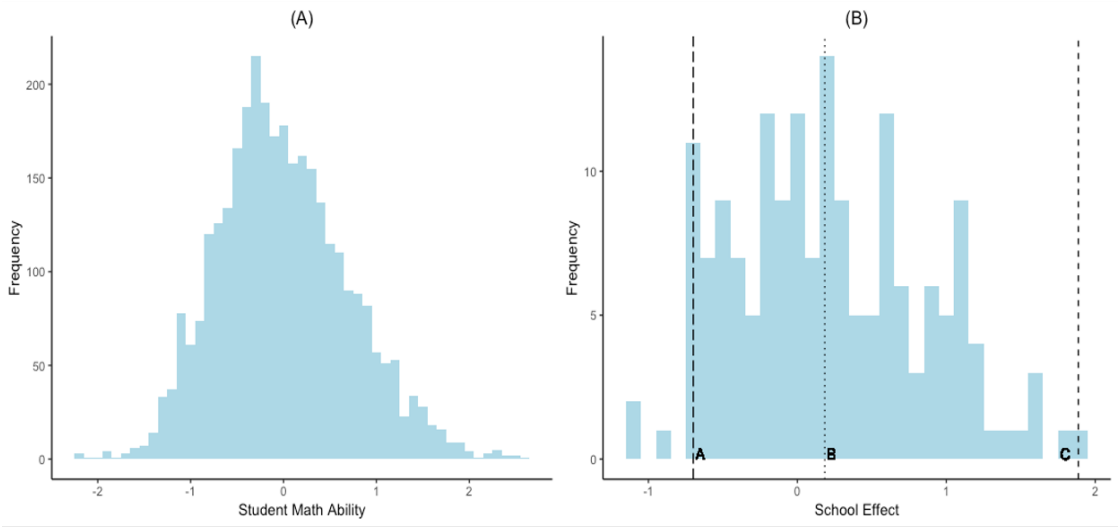

Figure 5.4: Histograms based on two-level Rasch model for (A) student math ability and (B) school effects. School effect scores for Schools A, B, and C are indicated with vertical dashed lines.

results from the previous section, student and school performance varies by context. At the student level, urban students and students who are not poor outperform rural and poor students, on average, in overall math ability.

Item Difficulty. Consistent with previous results, the majority of these items have high estimated item difficulty. Table 5.7 presents average item difficulty, the standard deviation, the minimum and the maximum for formulate items, other items (i.e., employ and interpet items), and for the overall assessment. Across the full assessment, the average estimated item difficulty is 1.91 (SD=1.68). Formulate items are the most difficult, with an average item difficulty of 3.14 (SD=2.10). Not only is the average high for the formulate process, some individual formulate items have extreme levels of difficulty. Estimated item difficulties for all items and their associated standard errors are presented in Appendix A.

Schools A, B, and C. Using the results from the multilevel Rasch model, we are able to find the estimated school effects for Schools A, B, and C, and the average student-level estimated ability within each school. School A is the lowest performing of these three schools, with the lowest school effect score ($\hat{\theta}_A = -0.699$, see Table 5.8 and Figure

Table 5.7: Average item difficulty, standard deviation, minimum, and maximum for items in the formulate cognitive process, the combined other item group (i.e., employ and interpret items), and the overall assessment, based on the multilevel Rasch model.

	Mean	SD	Min	Max
Formulate	3.14	2.10	1.13	8.10
Other Items	1.61	1.44	-1.34	5.78
Overall	1.91	1.68	-1.34	8.10

Table 5.8: School effect, school effect rank, and average student math score estimates based on multilevel Rasch model for Schools A, B, and C.

	School A	School B	School C
School Effect	-0.699	0.184	1.887
School Rank	157	82	1
Student Scores			
Mean	-0.065	0.027	0.131
SD	1.026	0.863	0.718

5.4) that is well below the mean school effect. In terms of its impact on student math scores, it is ranked 157th of 168 schools. Within School A, the average student-level estimated ability is also lowest, and the within-school gap is the highest of the three schools ($\theta_{j\bar{A}} = -0.065; SD = 1.026$). School B has a higher estimated school effect score than School A ($\hat{\theta}_B = 0.184$), and is ranked 82nd in terms of school impact. Within School B, the average student-level estimated ability is close to the mean within-school ability, and has moderate within-school gaps ($\theta_{j\bar{B}} = 0.027; SD = 0.863$). School C has the highest estimated school effect ($\hat{\theta}_C = 1.887$), is the highest ranked school in terms of school impact, and has the highest within-school average ability ($\theta_{j\bar{C}} = 0.131; SD = 0.718$).

Summary. The analysis using the multilevel Rasch model provides insight into school effectiveness and student math performance in Cambodia. Results indicate that there are differences in school effectiveness, and that some schools are less effective than other

schools. The results also indicate that there are differences in student performance within schools on mathematical ability. This information is very useful in understanding the impact schools have on student performance. However, the goals of the multilevel Rasch analysis differ from those outlined in Section 5.3, and based on this approach, we are not directly able to identify the low-performing schools that would benefit from an intervention specifically designed around formulating math problems. To do this, we now turn to the analysis using the MD-IR model.

5.5.2 Multilevel Diagnostic Item Response Model

The MD-IR model is useful in this scenario because it allows us to identify directly low-performing schools based on formulating items in order to select the schools that would most benefit from intervention on the formulating process. This section presents results from the MD-IR analysis, beginning with interpretations of the parameter estimates, followed by comparisons of overall probabilities of correct responses by latent classes. Lastly, evidence supporting reasonable model fit will be presented.

5.5.2.1 MD-IR Parameter Estimates

Table 5.9 presents parameter estimates for the distributions of the latent classes, and Tables 5.10 and 5.11 present estimates of item location and differences by latent class, respectively. Each of these is reviewed in turn.

Class proportions. Overall, 62% of schools were identified as low-performing schools. These schools are identified as the schools that would most benefit from the intervention on formulating mathematics problems. The majority of students, 56%, are classified in these low-performing schools, with 31% of these students classified as high-performing students in low-performing schools and the remaining 25% classified as low-performing students in low-performing schools. The sizable percentage of high-performing students within low-performing schools indicates that there are differences within these low-

Table 5.9: Distributional estimates and proportions for high- and low-performing student classes within high- and low-performing school classes, based on the MD-IR model.

School Class	Class 1 High		Class 2 Low	
Student Class	Class 1 High	Class 2 Low	Class 1 High	Class 2 Low
Student Ability Mean (μ_{gh})	1.210	0.000	0.402	0.000
	(0.08)		(0.13)	
Student Ability Variance (σ^2_{gh})	0.382	0.280	1.023	0.039
	(0.06)	(0.07)	(0.10)	(0.08)
School Effect Mean (μ_h)	0.394	0.394	0.000	0.000
	(0.11)	(0.11)		
School Effect Variance (σ^2_h)	0.295	0.295	0.195	0.195
	(0.04)	(0.04)	(0.02)	(0.02)
Proportion (π_{gh})	0.237	0.204	0.305	0.254

Note: Standard errors are in parentheses.

performing schools on student performance on formulate items. The remaining 44% of students are classified in high-performing schools, with 24% as high-performing students in high-performing schools, and 20% as low-performing students in high-performing schools. Table 5.9 presents the estimates of π_{gh} for each latent class.

School effect distributions. Estimates of the school effect distributions help us to understand the differences between the high- and low-performing schools in terms of their impact on student achievement. Estimates for school effect distributions are shown in Table 5.9, and the school-level distributions in the sample are shown in Figure 5.5. Schools are labeled as high or low depending on the comparative estimate of the school effect. The high-performing school class is labeled as class 1 ($\mu_{h=1}$ and $\sigma^2_{h=1}$), and the low-performing school class is labeled as class 2 ($\mu_{h=2}$ and $\sigma^2_{h=2}$). For brevity, estimated parameters are

written here without the h in the subscript.

There is a small difference in the mean school effect between the two latent classes ($\hat{\mu}_1 = 0.394, SE = 0.11$). This small difference indicates two things. First, since the difference is small, we can conclude that much of the difference between the high-performing and low-performing schools is due to differences in formulating items. Second, while the difference is small, any significant difference indicates that there are some differences in effectiveness in overall mathematics that are not captured in formulate items and are attributable to other processes. The variance components for both school-level latent classes are similar and relatively small. The variance component for the low-performing school class is the smallest ($\hat{\sigma}_2^2 = 0.195, SE = 0.02$), and is slightly larger for the high-performing school class ($\hat{\sigma}_1^2 = 0.295, SE = 0.04$). This suggests that the heterogeneity of effectiveness within school classes is similar for both high- and low-performing schools. Figure 5.5 shows the overlapping histograms of the school effect distributions. The histogram also includes dashed lines showing the estimated model parameter for the class means, and the dotted lines show the sample means based on most likely latent class assignment. Small discrepancies between these lines are expected.

Student score distributions. Estimates of student performance within school classes help us to understand how student performance differs within high- and low-performing schools. Table 5.9 also provides parameter estimates for student-level latent classes. As was discussed in Chapter 4 and shown in Figure 4.3, student-level distributions differ by school latent classes. Figure 5.6 shows student-level histograms for each school-level latent class. Student classes are labeled as high or low based on the comparative within-school estimated student ability. Combined with school latent classes, there are four student-level classes: high students in high schools ($\mu_{g=1,h=1}$; $\sigma^2_{g=1,h=1}$), low students in high schools ($\mu_{g=2,h=1}$; $\sigma^2_{g=2,h=1}$), high students in low schools ($\mu_{g=1,h=2}$; $\sigma^2_{g=1,h=2}$), and low students in low schools ($\mu_{g=1,h=2}$; $\sigma^2_{g=1,h=2}$). For brevity, estimated parameters are written without the g and h subscripts.

In low-performing schools, there is a small but significant difference between high-

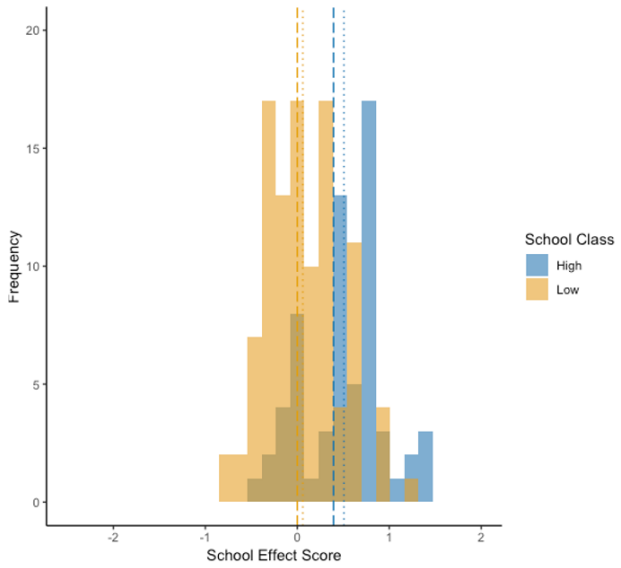

Figure 5.5: Histogram of school effects by school-level latent class based on results from the MD-IR model. High-performing schools are light blue and low-performing schools are gold.

and low-performing student classes, on average ($\hat{\mu}_{2,2} = 0.402, SE = 0.13$). There is substantial spread for the high-performing class ($\hat{\sigma}^2_{1,2} = 1.023, SE = 0.10$), but essentially no spread for the low-performing class ($\hat{\sigma}^2_{2,2} = 0.039, SE = 0.08$). This lack of spread is shown clearly in Figure 5.6, and is due to the difficulty of the assessment and to the fact that the students in this lowest-performing class are hitting the floor of the exam. For these students, the test is providing very little information that differentiates them from the other low-performing students, resulting in scores for all students in this low-performing class that are at, or very near, the fixed mean of zero. The large spread for the high-performing students in this low-performing school class also suggests an interesting finding. Specifically, there is a group of high-performing students who have noticeably lower scores compared to the other high-performing students. This can be seen in the bi-modal distribution of the high-performing school class. This group of high-performing students even has lower estimated math ability scores compared to the low-performing students. This unexpected result is due to the limited range and

the fixed mean of zero of the low-performing class. Additionally, these students differ from both the other high-performing students and from the low-performing students in interesting ways. They differ from the other high-performing students because they are responding correctly to fewer items across the entire assessment. However, they also differ from low-performing students because they respond correctly to some, albeit not many, formulate items. The average proportion correct for these students on formulate items is 0.063. Students in the low-performing class respond correctly to essentially no formulate items. The average proportion correct for low-performing students on formulate items is 0.005. Since the model is specified to differentiate latent classes based on formulate items, this group of students is classified as high-performing, as opposed to low-performing, because they do respond correctly to a small number of formulate items. In other words, in the low-performing school class, students are essentially differentiated based on whether or not they respond correctly to any formulate items. Clearly, many of the students in the high-performing student class within low-performing schools would also benefit from interventions on formulate items, making the school level an appropriate level of intervention for this context.

In high-performing schools, there is a large difference between high- and low-performing students ($\hat{\mu}_{2,1} = 1.210, SE = 0.08$). The spread within classes is small for both classes, but slightly larger for the high-performing class ($\hat{\sigma}^2_{1,1} = 0.382, SE = 0.06$; $\hat{\sigma}^2_{2,1} = 0.280, SE = 0.07$). The larger difference suggests that in high-performing schools, the difference between high- and low-performing students is not fully due to the formulate items, but also to the other cognitive processes. This difference is captured in the right panel of Figure 5.6, where we can see that the two classes in high-performing schools are relatively similar in size, and the estimated mean and sample class means for both classes are very similar, indicating reasonable fit and classification.

As discussed in Chapter 2, the within school variance component from the multilevel IRT model provides researchers with information regarding within-school achievement gaps. In the MD-IR model, information regarding within-school achievement gaps comes

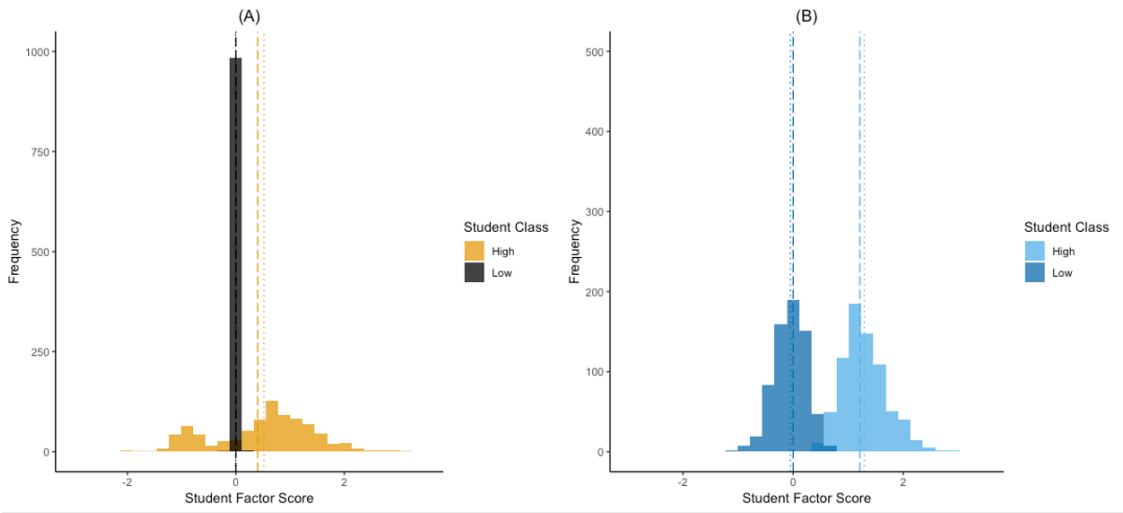

Figure 5.6: Histograms of student math ability for both school latent classes. Dashed lines show estimated model parameters for class means. Dotted lines show sample means based on latent class assignment. (A) Histogram for low-performing schools. Low-performing students in low-performing schools are shown as black, and high-performing students in low-performing schools are shown as gold. (B) Histogram for high-performing schools. Low-performing students are shown as dark blue, and high-performing students are shown as light blue.

from the mean difference between within-school latent classes, as well as from the variance components. When considering within-school gaps, looking at both mean differences and variances provides a clearer picture than when looking only at the variance component. This is because, while the variance components within student-level latent classes may be small, the mean difference could be large. In this example, this is the case for high-performing schools. The smaller variance components suggest that inequities within schools are small, but the larger mean difference between these classes indicates otherwise. The opposite is true for the low-performing school class. In this school class, the mean difference is small, but the variance component for the high-performing group is large.

Item location and overall difficulty. This section presents estimates and interpretations of item locations and overall item difficulties. Table 5.10 presents summary statistics for

Table 5.10: Average item location, standard deviation, minimum, and maximum for items in the formulate cognitive process, the combined other item group (i.e., employ and interpret items), and the overall assessment, based on the MD-IR model.

	Mean	SD	Min	Max
Formulate	4.63	2.20	2.41	9.77
Other Items	1.99	1.42	-0.90	6.13
Overall	2.50	1.90	-0.90	9.77

the item location parameter (β_i) for the MD-IR model. A full table of all item location parameters and associated standard errors is shown in Appendix A. As expected based on the preliminary analyses, the formulate items are the most difficult, with extremely high average location parameters and extreme item difficulties at the upper range of the distribution. The other items are not nearly as difficult, on average, but do have some highly difficult items at the extremes.

Estimates for the τ_{gh} parameter and its impact on overall item difficulty and response probability for formulate items are included in Table 5.11. Item response function (IRF) plots for each latent class for a hypothetical average item ($\beta_i = 2$) in the formulate item group are shown in Figure 5.7. Recall from Section 4.6.2 on item difficulty in Chapter 4, that for an overall reference class, the τ_{gh} parameter is fixed to zero, and the overall item difficulty is equal to the item location. The overall reference class is the class that was the reference at both the student and the school levels, which in this case is the class of low-performing students in low-performing schools. This means that the τ_{gh} parameter represents the amount of advantage that students in the higher classes have on formulate items, compared to the lowest class.

Within low-performing schools, formulate items are easier for high-performing students compared to low-performing students ($\hat{\tau}_{1,2} = 1.12, SE = 0.32$). Formulate items are also much easier for high-performing students in high-performing schools, compared to the low-performing students in low-performing schools ($\hat{\tau}_{1,1} = 1.66, SE = 0.24$). On

Table 5.11: MD-IR estimates for the τ_{gh} parameter, as well as the overall item difficulty (combined τ_{gh} and item location) on a hypothetical formulate item with an item location of $\beta_i = 2$. Probabilities of a correct response by latent class for a student at the mean of the math ability distribution, and within a school at the mean of the school effect distribution.

School Class	High		Low	
Student Class	High	Low	High	Low
τ_{gh}	1.66	0.40	1.12	0.00
	(0.24)	(0.32)	(0.32)	–
Hypothetical Formulate Item, $\beta_i = 2$				
Overall Item Difficulty	0.34	2.00	0.88	2.00
Probability of Correct Response	0.64	0.09	0.24	0.06

the other hand, the estimated difference in item difficulty on formulate items between low-performing students in high-performing schools and the low-performing students in low-performing schools is small and not statistically significant ($\hat{\tau}_{2,1} = 0.40, SE = 0.32$). The IRF in Figure 5.7 shows clearly how the formulate items are easier for the two high-performing student classes (blue and gold lines) compared to the reference class (black line). As is clear from the IRF, it is the high-performing student classes in both high- and low-performing schools that have the advantage on formulate items.

5.5.2.2 Overall Probabilities

In order to help make sense of these many results, it is useful to look at the overall probability of a correct response based on the different item groups, by class. To do this, we can consider the response probability for the average student in the average school, of each respective latent class, and compare these across classes and across item groups. In other words, we can plug in the means of the student-level and school-level distributions, the appropriate τ parameter, and selected item location parameters into Equation 4.1 in order to calculate the probability of a correct response on an item. For

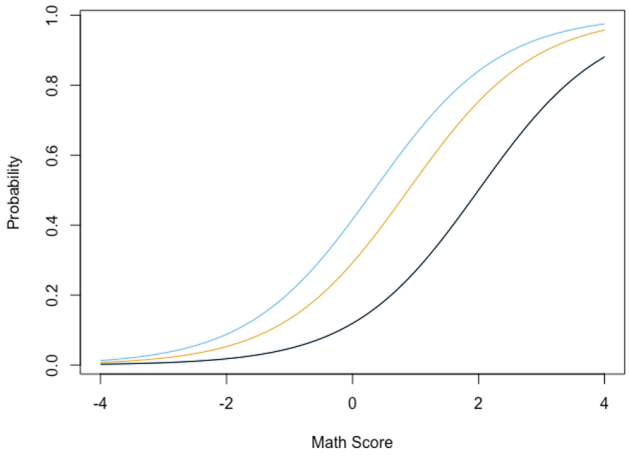

Figure 5.7: IRF plots from MD-IR model for a hypothetical formulate item with item location equal to two ($\beta_i = 2$). The IRF for the reference class, the low-performing students in low-performing schools, is the black line. The IRF for low-performing students in high-performing schools is equal to the IRF for the reference class. The IRF for high-performing students in high-performing schools is light blue, and the IRF for high-performing students in low-performing schools is gold.

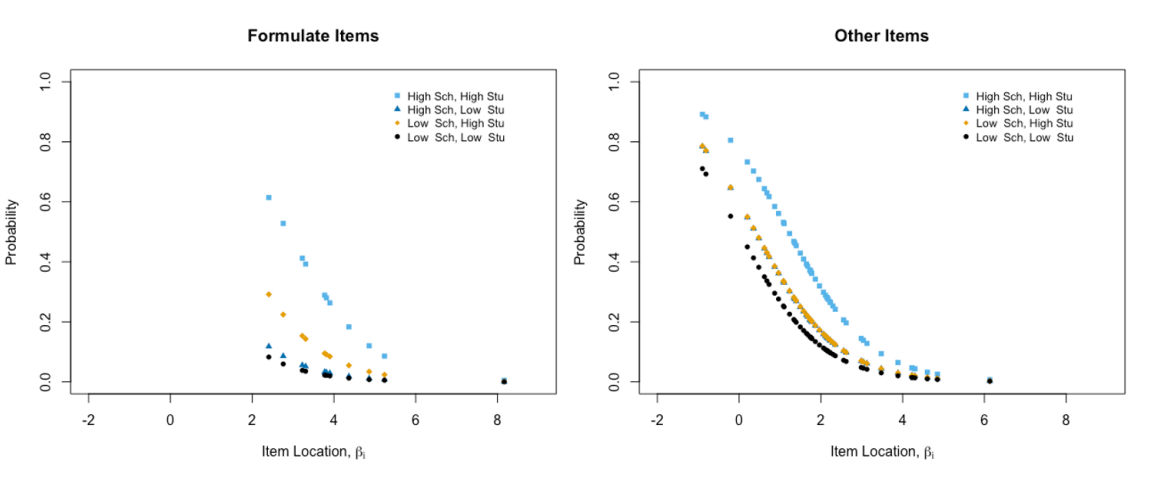

Figure 5.8: Overall probability plots for average students in average schools of each latent class, for the formulate (left) and other item (right) groups. Low-performing students in low-performing schools are black circles, high-performing students in low-performing schools are gold diamonds, low-performing students in high-performing schools are blue triangles, and high-performing students in high-performing schools are light blue squares.

example, consider a hypothetical formulate item with an item location parameter of 2 (i.e., $\beta_i = 2$). The response probability for an average low-performing student in an average low-performing school is very low, equal to 0.06 (see Table 5.11). In contrast, the response probability for an average high-performing student in an average high-performing school is relatively high, equal to 0.64. This approach allows us to see that, despite the fact that there is no difference in item difficulty for low-performing students in high-performing schools compared to low-performing students in low-performing schools, there is still an overall difference in response probability (0.09 compared to 0.06), due to the higher effectiveness of the high-performing schools.

Now we can consider the overall probabilities across the range of item locations in each of the item groups. Figure 5.8 plots the calculated probabilities for average students at each of the item locations in these item groups, separated by latent class. At the school

level, the small differences in average school effectiveness suggest that most of what differentiates the two school classes is captured in the formulate items. At the student level, the differences are small in the low-performing school class, but are larger in the high-performing school class, suggesting that not all of the student-level differences are captured in the formulate items. This becomes clear when looking at the overall probability of a correct response for students at the means of each latent class, on each of the two item groups. The left panel of Figure 5.8 shows the overall probability for formulate items. The number of items and the range of item locations are smallest in this focal item group. Despite this smaller range, we can clearly see differentiation among the latent classes on these formulate items. Most notable is the difference between the high-performing students in high-performing schools compared to the other classes, including the high-performing students in low-performing schools. The fact that the difference between low-performing students in each school class is not large suggests that the difference among schools is more attributable to the higher-performing students.

The right panel of Figure 5.8 shows the overall probabilities for the other items. While these items do not show as much differentiation as the formulate items, the mean differences do still capture some differences in probability by latent class. Again, the high-performing students in high-performing schools are the most clearly differentiated from the others, including high-performing students in low-performing schools. This indicates that while formulate items do differentiate well among latent classes, the other item groups may also differentiate among classes, but to a lesser degree.

5.5.2.3 Goodness-of-fit

Relative Fit. Results from the relative fit comparisons suggest reasonable fit of the MD-IR model to the data, compared to the first competing baseline model. Recall that two competing models were selected for comparison. The first is the multilevel Rasch model that is the most widely used IRT approach to studying school effects. All three fit indices, AIC, BIC and adjusted BIC, indicate better fit for the MD-IR model than the multilevel

Table 5.12: Relative model fit by AIC, BIC, and adjusted BIC for multilevel Rasch and MD-IR models.

	Multilevel Rasch	MD-IR
Classes Per Level	1-class	2-Class
AIC	70961.512	70742.667
BIC	71350.548	71204.648
Adjusted BIC	71147.193	70963.163

Rasch model (see Table 5.12). Since we would typically expect to see improved fit in a more flexible model, this provides us with minimal confidence of the reasonable fit to the data.

The second competing model that was selected was the multilevel mixture IRT. Unfortunately, when attempting to run the multilevel mixture IRT model with the PISA-D data from Cambodia, the model would not converge. This suggests a problem with empirical identification and provides some evidence that the data do not support this model. For this reason, Table 5.12 only presents results for the comparison with the multilevel Rasch model.

Absolute Fit. Since the relative fit comparisons only provide minimal confidence, an absolute fit approach is also used to evaluate goodness-of-fit. The second approach evaluates the predictive validity of the MD-IR model. 200 datasets were generated based on the MD-IR model parameters. The proportion correct on each subdomain at the within- and between-levels, as well as the proportion correct per item, are compared to the observed data weighted distributions as presented in Table 5.5. Small discrepancies would provide minimal confidence of reasonable fit of the MD-IR model to these data.

Table 5.13 shows the differences in each subdomain for the MD-IR model between the observed and generated data. At the student level, there are small differences in average proportion correct in all cognitive processes. The largest differences at the student level come from the top of the distribution, where a small number of students perform well on

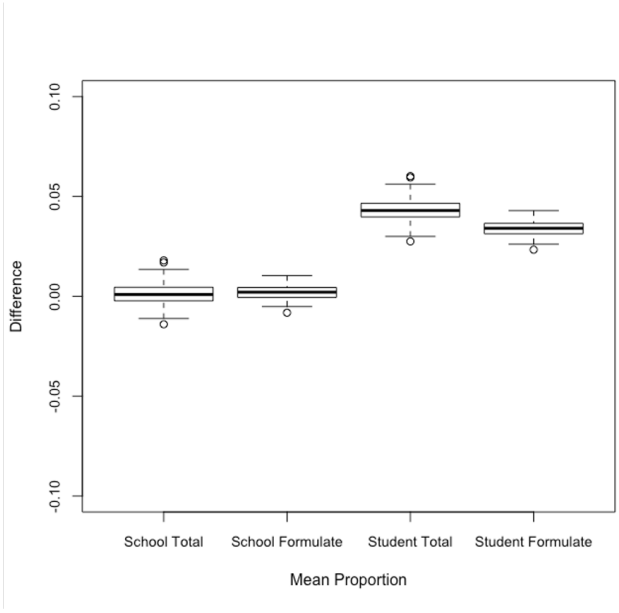

Figure 5.9: Distributions of differences in mean proportion correct on average school aggregate proportion correct, and on average student proportion correct, for total and formulate scores.

this assessment. At the school level, differences are quite small, indicating that the MD-IR model fits reasonably well, particularly at this level. The largest difference at the school level is also due to an outlier school at the top of the distribution of school averages on formulate scores (see Figure 5.2). While we would expect the differences to be smaller since they are aggregated, the results are encouraging as the MD-IR model aims to offer strong interpretations at the school level. Figure 5.9 shows the distributions of differences between the observed means and the generated means at both levels for each replication.

Fit results at the item level give us a more detailed look at absolute fit. On average, the differences for each cognitive process are small (see Table 5.13). Formulate items have the smallest average difference, while interpret items have the largest difference. The individual items that have the largest differences are the easiest items. This is because items that are extremely difficult have very low proportions correct, and therefore, limited space for larger differences to occur. To account for this, we can also divide the difference

Table 5.13: Difference between predicted data and observed data on proportions correct at the student and school levels, on within-school standard deviations, and on average item difficulty, all for each cognitive process and the overall PISA-D mathematics assessment.

	Employ	Formulate	Interpret	Total
Student Level				
Mean	0.039	0.034	0.049	0.043
SD	0.054	0.075	0.055	0.026
Min	0.000	0.000	0.000	0.000
Max	0.125	0.162	0.050	0.150
School Level				
Mean	0.006	0.002	-0.001	0.001
SD	0.003	-0.010	0.006	-0.006
Min	-0.032	-0.003	-0.057	0.000
Max	0.097	0.197	0.048	0.076
Within School SD				
Mean	0.056	0.080	0.051	0.028
SD	0.008	0.039	0.006	-0.005
Min	0.008	0.000	0.054	-0.012
Max	0.053	0.145	0.094	0.016
Item Difficulty*				
Mean	0.066	0.049	0.080	0.068
SD	0.033	0.030	0.030	0.033
Min	0.001	0.000	0.024	0.000
Max	0.137	0.088	0.141	0.141

Note: *Item difficulty here is defined as the proportion of students responding correctly to an item within the processes.

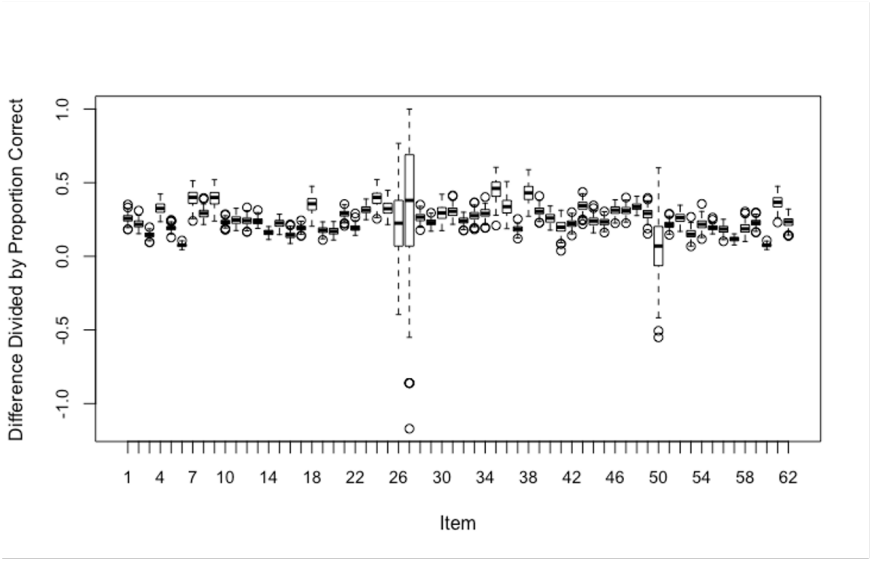

Figure 5.10: Distribution of differences in item proportion correct between replicated and original data, divided by the original item proportion correct, across all replications for all 62 PISA-D mathematics items.

for each replication by the actual proportion correct. Figure 5.10 shows the distributions for each item after making this adjustment. After accounting for item difficulty, only a few items with extreme difficulty values show poor fit.

Overall, these two approaches suggest that we can have at least minimal confidence that the MD-IR has reasonable fit to these data. With reasonable fit established, we can now explore whether the results provide expected conclusions regarding student and school context.

5.5.3 Evidence Supporting Model Validity

5.5.3.1 Proportions Correct by Latent Class

In order to explore the reasonableness of results, we can look at the observed proportion correct based on the most likely latent class classification from the MD-IR model and compare classifications by demographics. Overall, results from these comparisons support

the conclusions drawn from the MD-IR model. Table 5.14 presents the proportion correct on each subdomain and overall for school classes and for students in schools. At the school level, the high-performing class has higher average performance across all processes compared to the low-performing school class. There are also differences across latent classes at the student level. These differences are particularly striking on the formulate items. In the low-performing student classes, the differences between the other processes and the formulate domain are quite large, whereas in the high-performing classes, the differences are smaller. This provides support for the fact that formulate items do differentiate well among classes. However, there are smaller differences on the other processes, particularly between the highest performing class, suggesting that the assumption that there are no differences on the other items may not be ideal. The limitations around this assumption will be discussed in the conclusion section.

5.5.3.2 Demographics by Latent Class

Comparing school and student demographics by latent classes allows us to address the third research goal of understanding the schools and students in the low-performing school class. It also allows us to confirm whether the latent classes from the MD-IR make sense given our earlier analyses. As would be expected, there is a clear relationship between school context and school classification. Table 5.15 presents the proportion of high- and low-performing schools by poverty and urbanicity. Consistent with earlier findings that lower-resourced schools perform more poorly than their highly-resourced counterparts, 70% of the 47 schools with extremely low resources, and 88% of the 40 schools with severely low resources are classified as low-performing schools. On the other hand, 86% of moderately-resourced schools, and 83% of highly-resourced schools are classified as high-performing. However, it is worth noting that 30% of extremely low-resourced schools are classified as high-performing. An SES-targeted intervention, as opposed to a performance-targeted intervention would miss the fact that some low-resourced schools may not need support in teaching formulate math skills. Additionally,

Table 5.14: Average proportion correct and standard deviations for high- and low-performing school and student classes, for each cognitive process and overall PISA-D math assessment.

	Employ	Formulate	Interpret	Overall
School Classes				
High				
Mean	0.34	0.24	0.45	0.36
SD	0.21	0.28	0.24	0.19
Low				
Mean	0.23	0.09	0.30	0.23
SD	0.19	0.18	0.23	0.16
Student Classes				
High School, High Student				
Mean	0.46	0.42	0.57	0.49
SD	0.19	0.27	0.19	0.14
High School, Low Student				
Mean	0.22	0.05	0.31	0.22
SD	0.17	0.09	0.20	0.12
Low School, High Student				
Mean	0.29	0.20	0.37	0.29
SD	0.22	0.22	0.26	0.19
Low School, Low Student				
Mean	0.18	0.01	0.24	0.17
SD	0.14	0.03	0.17	0.09

Table 5.15: Proportion and counts of schools in each resource level and urbanicity categories that are classified in each high-performing and low-performing class.

	High		Low		Total
	N	Proportion	N	Proportion	N
Resource Level					
Extremely Low	14	0.30	33	0.70	47
Severely Low	5	0.13	35	0.88	40
Low	14	0.38	23	0.62	37
Moderate	28	0.86	4	0.14	32
High	10	0.83	2	0.17	12
Urbanicity					
Urban	34	0.76	11	0.24	45
Rural	30	0.24	93	0.76	123

only 24% of the 123 rural schools are classified as high-performing, compared to 76% of the 45 urban schools. This is also consistent with the previous finding that urban schools typically outperform rural schools.

There is also a clear relationship between student demographics and student classifications. Table 5.16 presents demographics by latent class at the student level. In both high- and low-performing schools, extremely and severely poor students are more likely to be classified as low-performing, but not always by sizable amounts. Similarly, it is only in high-performing schools that not poor students are classified as high-performing at much higher rates. Differences by gender are small in high-performing schools, but much larger in low-performing schools. Specifically, female students are more likely to be classified as low-performing if they attend a low-performing school. Differences in urbanicity are similar to that of the school level, since for most students, their urbanicity matches that of their school.

Overall, these differences in classification rates by school and student context support

Table 5.16: Proportion and counts of students in each poverty level, urbanicity, and gender categories that are classified in each high- and low-performing student class within high- and low-performing schools.

	High Schools				Low Schools				
	High Students		Low Students		High Students		Low Students		Total
	N	Prop	N	Prop	N	Prop	N	Prop	N
Poverty Level									
Extremely Poor	10	0.10	17	0.18	30	0.32	38	0.40	95
Severely Poor	215	0.15	244	0.17	431	0.30	546	0.38	1436
Poor	251	0.24	230	0.22	262	0.25	304	0.29	1047
Not Poor	272	0.42	181	0.28	97	0.15	97	0.15	647
Gender									
Female	342	0.20	376	0.22	427	0.25	564	0.33	1709
Male	349	0.23	303	0.20	424	0.28	440	0.29	1516
Urbanicity									
Urban	409	0.43	361	0.38	86	0.09	95	0.10	951
Rural	318	0.14	318	0.14	751	0.33	887	0.39	2274

the validity of the findings from the MD-IR model. While it is very concerning for equity, higher classifications of rural and low-income students into low-performing student and school classes are consistent with findings from the well-established multilevel Rasch approach and with trends in low-income countries. In practice, these findings would impact delivery and implementation of the hypothetical intervention as urbanicity and resource level are important considerations in any intervention.

5.5.4 Individual School Classifications

Now that some confidence in the results has been established through fit and validity analyses, we can now look at individual classifications for the three example schools, Schools A, B, and C. School classification, its associated probability of this classification, school effects, and within-school student performance for each of these three schools are presented in Table 5.17.

School A. Recall that School A is the lowest performing of these three schools in terms of the total proportion correct (see Table 5.3) and the estimated school effects from the Rasch model. Also recall, that the average proportion correct on formulate items was somewhat higher, relative to other processes, but still not higher than the sample average. However, despite the relative strength in the formulate items, School A is still classified as a low-performing school with a probability of 92%, and has an estimated school effect score lower than that of the mean for the low-performing schools ($\hat{\theta}_{A,low} = -0.59$).

Within School A, the highest proportion of students are classified as low-performing (58%). At the student level, the classification probabilities are given such that the sum of all four possible student classes sums to one. Within school A, the average classification probability for the low-performing class is 52%. The low-performing student class within School A performs below the average of low-performing classes in low-performing schools, on average ($\hat{\theta}_{\bar{j},stulow,A,schlow} = -0.11; SD = 0.21$). The rest of students are classified as high-performing with an average probability of 73%. The high-performing student class outperforms the average of the high-performing student class in low-

Table 5.17: School class, classification probability, estimated school effect, and within-school student performance statistics for Schools A, B, and C.

	School A	School B	School C
School Class	Low	Low	High
Probability	0.92	0.79	1.00
School Effect	-0.59	0.27	1.42
Student Performance			
High Student Class			
Proportion	0.42	0.67	0.80
Average Probability	0.73	0.63	0.84
SD Probability	0.21	0.13	0.13
Average Math Achievement	0.91	0.55	1.11
SD Math Achievement	1.33	0.92	0.46
Low Student Class			
Proportion	0.58	0.33	0.20
Average Probability	0.52	0.47	0.74
SD Probability	0.04	0.07	0.16
Average Math Achievement	-0.11	0.11	0.13
SD Math Achievement	0.21	0.36	0.70

performing schools ($\hat{\theta}_{\bar{j},stuhigh,A,schlow} = 0.91; SD = 1.33$).

The classification with confidence of School A in the low-performing school class provides additional validation of the MD-IR model. Throughout the other analyses, School A is consistently one of the lower-performing schools, and this remains the same based on the results of the MD-IR.

School B. School B represents a school that is average performing in terms of proportion correct on both the full assessment and the formulate items, and average in the estimated school effects from the Rasch model. Based on the MD-IR, School B is classified as a low-performing school, but with less certainty than School A. The probability of classification in this low-performing class is 79%. This lower certainty is expected and common when scores fall closer to the overlaps of the distributions. School B's estimated school effect score is higher than the average low-performing school mean ($\hat{\theta}_{B,low} = 0.27$).

Within School B, a majority of students are classified as high performing (67%) with an average probability of 63%. The high-performing student class within School B performs above the average of high-performing classes in low-performing schools, on average ($\hat{\theta}_{\bar{j},stuhigh,B,schlow} = 0.55; SD = 0.92$). The remaining 33% of students are classified as low-performing with an average probability of 47%. These students perform slightly above the average of low-performing students in low-performing schools ($\hat{\theta}_{\bar{j},stulow,B,schlow} = 0.11; SD = 0.36$).

School B provides an example of a school that is classified with less confidence compared to most other schools in the analysis. This result for School B is not surprising given results from the previous analyses as School B is consistently in the middle of the distribution. In any classification procedure, there will be schools (and students) whose classification is less certain. As such, seeing this uncertainty from the MD-IR model is consistent with other classification approaches.

School C. School C is a higher performing school. It is classified as high performing with over 99% probability. Its estimated school effect is much higher than that of the average for high-performing schools ($\hat{\theta}_{C,high} = 1.417$).

Within school C, 80% of students are classified as high-performing students. The estimated mean student ability for this class is slightly lower than the average for high-performing students in high-performing schools ($\hat{\theta}_{j,stu\bar{h}igh,C,schhigh} = 1.106; SD = 0.456$). The remaining 20% of students perform above the average for low-performing students in high-performing schools ($\hat{\theta}_{\bar{j},stulow,C,schhigh} = 0.127; SD = 0.699$).

School C provides an example of a school that is classified as high-performing with a high level of confidence. Throughout the analyses, School C was at the higher end of the distribution in scores and school effectiveness. As such, the classification in the high-performing class helps to further validate the MD-IR model.

5.6 Simulation Study

A final step in ensuring we can have confidence in the results provided by the MD-IR is a small simulation under the same conditions as the PISA-D data from Cambodia. To assess precision in parameter estimates, parameter bias and root mean square error (RMSE) are calculated. To assess accuracy of the classifications, classification accuracy and kappa's coefficient are calculated for both student and school levels.

5.6.1 Data Generation

Data are generated under the assumption that the MD-IR model is the true model in the population. Specifically, 100 datasets are generated using the same sample sizes as the Cambodia data, (i.e., 3,225 students in 168 schools). School effect and ability distributions are assumed to follow the same distributions by latent class as estimated in the previous section. Estimated item locations and τ parameters are used to set the item difficulties in each latent class. Estimation settings for each model were the same as described for the MD-IR analysis.

Table 5.18: Average parameter bias and RMSE for school effect latent class distribution estimates, for high- and low-performing schools.

	Mean		Variance	
	Bias	RMSE	Bias	RMSE
High Class	-0.019	0.191	0.004	0.036
Low Class	–	–	0.003	0.034

5.6.2 Results

Of the 100 replications, 91 converged replications were used. 9 replications did not converge because no correct responses were generated for one item that was extremely difficult in the original data. Specifically, item 36 has an extremely high estimated item location that was used in generating the data sets ($\hat{\beta}_{36} = 9.765$; see Table 5.10 for formulate maximum difficulty and item 36 in Table A.1). Since no correct responses were generated for this item, and therefore the replications could not converge, these replications were excluded from the analysis.

5.6.2.1 Parameter Recovery

Results from the simulation are encouraging. At the school level, bias and RMSE are small for both the estimated means and variances (see Table 5.18), giving us confidence that we can use the results at this level to understand better the difference between the high- and low-performing school classes. Figure 5.11 shows the distribution of parameter bias for each estimated parameter. The parameter bias for the estimated high-performing school mean (labeled as "Hi:SchMean") has a small average bias (-0.019) and a small spread across the distribution. The same holds true for the variances for both school latent classes. The average bias is very small for the high-performing class (0.004; "Hi:SchVar") and the low-performing class (0.003; "Lo:SchVar") variances, with even smaller spread across the distribution.

Table 5.19: Average parameter bias and RMSE for student math achievement latent class distribution estimates, and difference in item difficulty estimates, for each student-level latent class within school-level latent classes.

	Mean		Variance		Tau	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
High Sch, High Stu	0.004	0.038	-0.005	0.048	-0.023	0.229
High Sch, Low Stu	–	–	0.001	0.007	-0.004	0.038
Low Sch, High Stu	-0.008	0.078	0.001	0.000	-0.028	0.281
Low Sch, Low Stu	–	–	-0.002	0.017	–	–

Student level results are also encouraging, and are shown in Table 5.19. Average bias for the estimated mean is very small for the high school and high students (0.004; "Hi,Hi:StuMean"), and for the low school and high students (-0.008, "Lo,Hi:StuMean"). Bias for each of the variances are also very small with small spread in the distribution. Bias in each of the estimated differences in item difficulty are slightly larger, but are still quite small. The average bias range is from -0.028 to -0.004, with somewhat larger spread across the replications.

For most items, item location parameters are well recovered with low average bias. Figure 5.12 shows the distributions of parameter bias for all 62 items. Three items with extremely high levels of difficulty have small averages, but large ranges in the distributions. Other items have small averages and small ranges, giving confidence on the item location parameters estimated in the MD-IR.

5.6.2.2 Classification Accuracy

In order to assess the accuracy of school and student classifications, the average classification accuracy and kappa coefficient were calculated at the school level, the student level, and the combined levels. Classification accuracy is defined as the proportion of schools or students correctly classified in each replication. The average classification presented

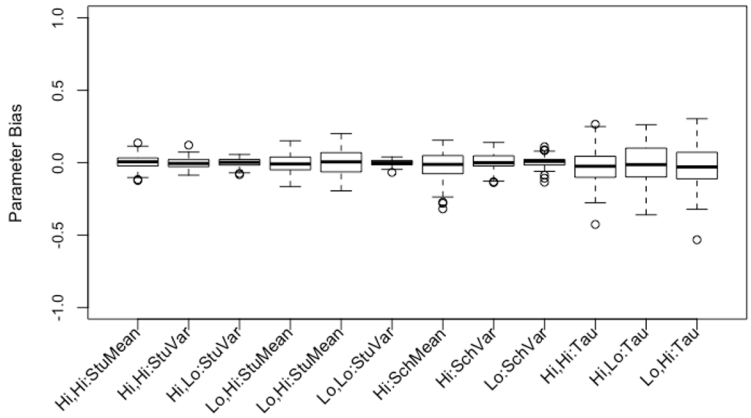

Figure 5.11: Parameter bias distributions for estimated school effects by latent class, student math achievement by latent class, and τ parameters by latent class.

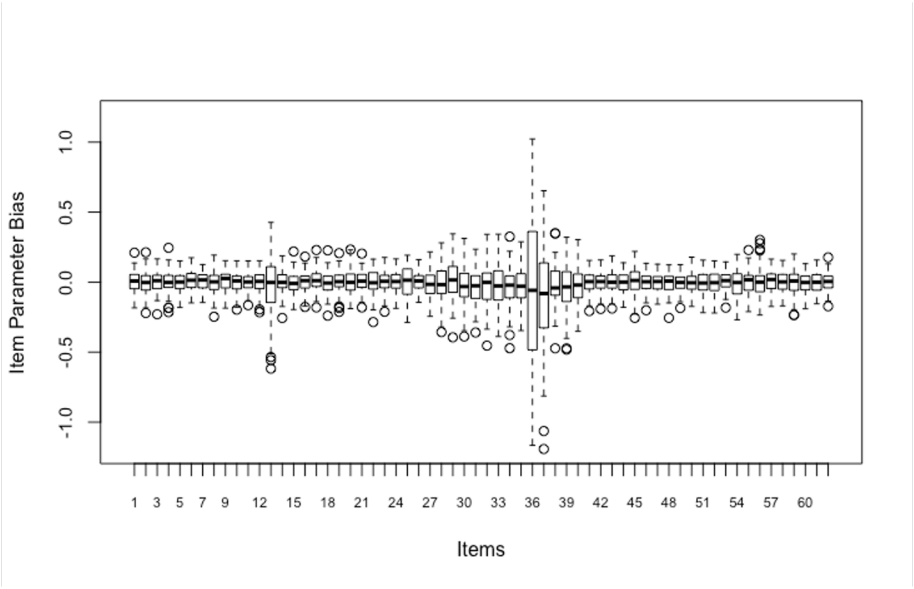

Figure 5.12: Item parameter bias distributions for 62 items across each replication.

in Table 5.20 is the average of these across each replication. Classification accuracy at the school level is very strong at 0.92. The kappa coefficient, which shows the level of agreement while accounting for the possibility of agreement by chance, is also very high at 0.83. This high accuracy gives us good confidence that we have well identified the low-performing schools for the intervention.

Classification at the student level is also high. Two student level classifications are presented in Table 5.20. First is the overall student classification which represents a student's high- or low-performing classification, regardless of school classification. Second is the overall student within-school classification, which is the correct classification rate for classifications of students in school classes (e.g., low-performing students in low-performing schools). This second classification rate is slightly lower, as it has two possibilities for error, one at the student level and one at the school level.

In order to determine whether some classes show better classification rates than others, the classification accuracy rates by true latent classes are also shown in Table 5.20. Students in high-performing schools show high classification rates for both high- and low-performing students. On the other hand, rates within low-performing schools are somewhat lower, particularly so for the high-performing students in low-performing schools. This is likely due to the low overall performance and floor effects for students in this school class.

5.6.3 Simulation Summary

Overall, simulation results provide additional confidence in the results found in the PISA-D study. Parameter recovery is strong for the latent class distribution estimates and for the estimated τ parameters. Classification accuracy is also very strong, particularly at the school level. This is very encouraging as the primary goal in this example is to identify low-performing schools based on the formulating cognitive process. Results suggest that we can have confidence in the classifications based on the MD-IR model.

Table 5.20: Average classification accuracy and kappa coefficient for all schools and all students, and classification accuracy by school and student latent classes.

	Classification Accuracy	Kappa
Overall School	0.92	0.83
Overall Student	0.83	0.66
Overall Student Within School Classes	0.77	0.69
High Class	0.90	
Low Class	0.93	
Student Classes		
High Sch, High Stu	0.83	
High Sch, Low Stu	0.82	
Low Sch, High Stu	0.61	
Low Sch, Low Stu	0.79	

5.7 Chapter Conclusion

The primary purpose of this chapter was to demonstrate an application of the MD-IR model in the context of international education development. Specifically, this chapter has demonstrated how the key components of the MD-IR model can be used to bring together the study of school effects, performance-targeted interventions, and mathematical literacy. To do this, we imagined a hypothetical scenario where the goal was to identify low-performing schools that would benefit from additional support on formulating mathematical problems in real life situations. Three specific research goals were addressed, each of which is discussed in more detail in this section. This section will also discuss limitations from this analysis and potential directions for addressing these limitations.

The first research goal was to identify low-performing schools based on the formulate cognitive process. 62% of schools in these data were identified as low-performing schools. Of the three example schools that we followed throughout this chapter, two schools,

Schools A and B, were identified as low-performing. The second research goal was to understand student performance within these low-performing school classes. Through the within-school student distribution estimates, we learned that differences in student overall math achievement are smaller in low-performing schools, compared to differences in high-performing schools. However, there are some students classified as high-performing within low-performing schools who would also benefit from support on formulating problems. The fact that students were differentiated based on whether they responded correctly to any formulate items versus no formulate items suggests that even some high-performing students would benefit from an intervention that is targeted at the school level. Through the additional τ parameter, we also learned that the high-performing students in low-performing schools have an advantage over low-performing students on formulate items. Taken together, these findings suggest that the differences between high- and low-performing students in the low-performing school class are primarily coming from a difference in the formulate process. As such, interventions for these schools that focus on the formulate process should also emphasize the need to support the low-performing students.

The third research goal was to understand the school and student context of schools in the low-performing class. The majority of schools in the low-performing class are extremely low-resourced, severely low-resourced, and low-resourced schools. Only 6 moderately- and highly-resourced schools were classified as low-performing schools. Similarly, the majority of schools in the low-performing class are rural schools. These gaps between well-resourced and less-resourced schools, and urban and remote schools, have implications for the implementation and success of a performance-targeted intervention. In this case, any targeted intervention would also need to ensure that the schools have the resources needed to carry out any recommended improvements. Yet despite the fact that low-resourced schools are over-represented in the low-performing category, a number of these schools were categorized as high-performing. This is encouraging in that there are some schools that are effective, despite the challenges of having fewer resources.

Students who attend the low-performing schools are also disproportionately poor and rural. There are also larger inequities within low-performing schools. Specifically, poor students are more likely to be categorized as low-performing when attending low-performing schools. The same is true for female students. Females who attend a low-performing school are more likely to be classified as low-performing. In the case of the hypothetical intervention, emphasis should be placed on reaching the lowest performing students across all demographic groups.

Two main limitations exist in this analysis. First, the test is extremely difficult, leading to floor effects and limited variance among the lowest-performing students. This could be partially due to the fact that the polytonomous items were dichotomized, limiting the possibility for partial credit. Future extensions of the MD-IR model, discussed in Chapter 6, could address this concern. Second, the current formulation is limited to two item groups. While the hypothetical scenario was only focused on formulate items, results indicate that there could be differences among the classes on other cognitive processes as well. Another extension of the MD-IR model could address this limitation by allowing for additional item groups.

Overall, this analysis has allowed us to identify the low-performing schools based on the formulate cognitive process, and has given us insights into the context of these schools. A performance-targeted intervention is one approach for improving outcomes and equity for all students, and the MD-IR model is a useful tool in identification of participants for these interventions.

CHAPTER 6

Conclusion

Throughout the world, local and national governments hope to improve schools in order to ensure quality education for all. In order to meet this challenge, governments need to implement interventions that target specific populations related to the goals of the improvement plans. Performance-targeted interventions are particularly difficult when it comes to identifying schools for participation in school-level interventions. This dissertation has highlighted the need for a modeling approach that supports these types of interventions, and has presented the MD-IR model as an option for identifying schools based on student assessment performance. This chapter reviews the contributions of the MD-IR model which was outlined in Chapter 4, and was demonstrated by the example in Chapter 5. It also discusses the limitations of the model in its current formulation, and suggests future extensions that could address these limitations.

6.1 Contributions of the MD-IR model

The current approaches to estimating school effects that were discussed in Chapter 2 do not provide direct classifications of schools and students based on assessment performance. HLM and multilevel IRT models also do not incorporate curriculum information or allow for differences in within-school achievement gaps based on school effectiveness. By incorporating approaches from other current psychometric models as discussed in Chapter 3, the MD-IR model is able to address these limitations, and contributes to the literature of school effects modeling by offering an alternative approach that may be used in settings where classifications are needed. Each of these contributions of the MD-IR

model is reviewed, and the implications in the Cambodia PISA for Development context are discussed.

6.1.1 Classification of schools and students

The first contribution of the MD-IR model is that it is a confirmatory model that classifies schools based on an a priori theory. Specifically, the introduction of a pre-specified number of latent classes allows for the classification of both students and schools into groups based on performance levels on pre-defined item groups of interest. In the example from PISA for Development in Cambodia, we were interested in identifying the schools that need the most support in teaching the formulate cognitive process in order to solve math problems in real world contexts. The example demonstrates how we can identify schools with certain levels of probability as high-performing or low-performing. By identifying schools using this approach, district, state, or national leaders can best identify schools for performance-targeted interventions that are related to the particular curricular area that the intervention will address.

Schools can be a particularly useful target level of intervention because education leaders often have easier access to schools than to individual students. Additionally, as is shown in the PISA-D example of low-performing schools, both high- and low-performing students within low-performing schools may need support and can benefit from improved instruction. This type of performance-targeted support can also have implications for improving equity outcomes. As is shown in Figure 1.1, inequities can exist when certain student sub-populations have systematic differences in their foundations for success, which in turn, can lead to inequalities in prosperity outcomes for some students. We see this in the PISA-D example. In Cambodia, rural schools and schools with extremely low resources are more likely to be categorized as low-performing schools, with lower overall student mathematics achievement. In this case, a performance-targeted intervention with identification based on the MD-IR model can support reducing inequities due to poverty and location.

In addition to identifying schools, the MD-IR model is also able to classify students as high-performing or low-performing on formulate items. This addition allows us to understand better the differences between high- and low-performing schools. In the PISA-D example, the addition of student classes allowed us to learn that there are wider within-school gaps in high-performing schools compared to low-performing schools. Additionally, student classes allow school leaders to identify students who may need the most support in particular curricular areas. Utilizing latent classes at both levels can allow for tailored interventions depending on the context. In Cambodia, the student-level classes supported the idea of school-level interventions as the most appropriate for improving outcomes on the formulate process. However, another scenario could reveal a different need. For example, if a district has a large number of schools that only have a few low-performing students, then perhaps a performance-targeted intervention as enrichment or tutoring makes more sense for these schools than a full school intervention. The information provided by the MD-IR model allows for these types of context specific adjustments.

6.1.2 Incorporation of curriculum information

Another important contribution of the MD-IR model is that it incorporates curricular information to provide structure in a mixture item response theory approach, allowing for latent classes to be differentiated by their performance on key curricular areas. Specifically, the inclusion of the τ parameter allows for differences in item location across classes based on specific item information that is confirmatory and is determine based on theory. In the PISA-D example, schools were differentiated based on student performance on formulate items, allowing us to identify the low-performing schools that need the most support on the formulating process. This parameter also allows us to understand the size of the differences in difficulty on these items. For example, we were able to learn that high-performing students in both low- and high-performing schools had an advantage compared to low-performing students in low-performing schools, but that this advantage

was particularly large for the high-performing students in high-performing schools. These results supported the need for an intervention focused on the formulate process. However, if the τ parameters in the PISA-D example were non-significant or very small, we might have concluded that an intervention focused on the formulating process is not the best approach for improving outcomes, and could have considered another theory for differences based on another curriculum area.

6.1.3 Allowance for differences in achievement gaps

The MD-IR model also allows for differences in achievement gaps across school classes. In both HLM and multilevel IRT models reviewed in Chapter 2, the distribution of student scores within schools, and school effect scores between schools, are assumed to be equal across the population. By incorporating latent classes in a mixture model framework, we can relax this assumption and allow for differences in gaps within schools. The differences in achievement gaps are captured both in the difference between student-level means and the student-level variance of the within-school classes. In Cambodia, results suggest that there are larger gaps in the high-performing schools compared to the low-performing schools. The mean difference between high- and low-performing classes in the high-performing schools is quite large, while the mean difference in low-performing schools is moderate. While the hypothetical intervention in Chapter 5 was meant for low-performing schools, this information is also useful for understanding high-performing schools. Schools in this class, while not participating in the school-level intervention, should be encouraged to recognize the gaps within their schools and to provide support for the low-performing students. Accounting for these differences in distributions allows the model to take into consideration the fact that schools differ in their context and environment. Additionally, the insights provided by estimating the different distributions allow education leaders to understand better the context and experiences of schools and students within these schools, which can lead to improved and targeted interventions.

6.2 Useful Extensions

This dissertation has demonstrated the usefulness of the MD-IR model in identifying low-performing schools on a specific curricular area of interest. The formulation of the model presented in Chapter 4.1 is well suited for this context, but some additional extensions to the model would make it more general and allow its application in a wider range of contexts. Some useful extensions for future research are briefly reviewed here.

1. Covariates can be included in different parts of the MD-IR model for different purposes (Li et al., 2016). First, covariates can be included to explain differences in latent class assignment. Second, covariates can be included in the measurement portion of the model to explain differences in the continuous latent trait or to control for contextual effects. Incorporating covariates in this way allows for the MD-IR model to consider both Type A and Type B effects as discussed in Chapter 2.

2. The current formulation of the MD-IR was proposed for two item groups. This is suitable in certain situations, as in the hypothetical intervention in Chapter 5. However, there may be situations where educators would like to differentiate latent classes based on additional attributes. For example, there may be interest in differentiating latent classes based on additional cognitive processes, allowing for targeted interventions at various stages of the mathematical problem solving process. In these situations, the MD-IR model could be extended to accommodate multiple attributes.

3. Many assessments are not limited to dichotomous responses. A different link function could be chosen (e.g., a multinomial or adjacent link function) in order to accommodate polytonomous items. Jeon (2018) has shown this extension in the single-level context. This extension could be useful for addressing the high level of difficulty found in the PISA-D data as allowing for partial credit may better differentiate among low-performing students.

4. Currently, the MD-IR uses the Rasch model, with only one item parameter, (i.e., the item location parameter). A two-parameter logistic model could be used instead, and an item discrimination parameter would then be included. Again, Jeon (2018) has shown this extension in the single-level context.

5. A learning method to identify an empirical Q-matrix (e.g., Liu, Xu, & Ying, 2012) could be applied to the MD-IR if attribute information is not available. While this is technically possible, the empirical Q-matrix must be validated by content and assessment experts.

The first two extensions allow for a more thorough understanding of the context of high- and low-performing classes, at both the student and the school level. The third and fourth extensions focus on extending the measurement model to accommodate more assessment contexts. Finally, the last extension furthers possible applications of the MD-IR model to assessments that do not currently have item attribute information. Together, these useful extensions allow for additional applications and opportunities, and will be the focus of future research.

6.3 Future Research

In addition to research on the enumerated extensions, future research will also incorporate simulation studies to understand better the behavior of the MD-IR model in various contexts. Some areas of focus will be on the sample size requirements within and between schools, as well as the ideal number of items needed for the item attribute of interest. Additionally, simulation studies will focus on the sensitivity of the MD-IR model to differentiate between classes based on different parameter sizes, as well as under alternative true population models. These simulation studies will further support future applications of the MD-IR model.

6.4 Conclusion

To conclude, the MD-IR model offers new opportunities for identifying schools for performance-targeted interventions with the overarching goal of improving education outcomes for all students. It also offers the opportunity to differentiate students and schools based on important curricular information that is often included in available assessment data. In addition to providing classifications based on these curricular areas, the MD-IR model also provides parameter estimates that can reveal interesting patterns in performance of schools through the school effect distributions. The parameter estimates can also reveal interesting patterns in the performance of students within schools through the student ability distributions. These opportunities make the MD-IR model a useful addition to the school effects literature, providing education researchers an additional tool that provides actionable data to leaders and practitioners working to make education equitable for all students in all countries.

APPENDIX A

Item Difficulty and Location Parameters

Table A.1: Item difficulty estimates from Rasch model and item location estimates from MD-IR model, for all 62 PISA-D math items, (continued on multiple pages).

	Multilevel Rasch		MD-IR	
Item	Item Difficulty	(SE)	Item Location	(SE)
1	1.877	(0.097)	2.235	(0.105)
2	3.126	(0.151)	3.476	(0.139)
3	-1.227	(0.107)	-0.812	(0.096)
4	0.992	(0.094)	1.361	(0.094)
5	1.997	(0.106)	2.352	(0.115)
6	-0.610	(0.112)	-0.209	(0.104)
7	1.374	(0.088)	1.743	(0.097)
8	-0.047	(0.098)	0.351	(0.094)
9	1.790	(0.092)	2.155	(0.106)
10	1.377	(0.098)	1.749	(0.102)
11	1.851	(0.115)	2.219	(0.107)
12	0.959	(0.101)	1.338	(0.095)
13	5.780	(0.383)	6.132	(0.377)
14	2.185	(0.116)	2.556	(0.128)
15	1.358	(0.127)	1.728	(0.121)
16	1.733	(0.105)	2.111	(0.084)
17	2.242	(0.132)	2.617	(0.126)

	Multilevel Rasch		MD-IR	
Item	Item Difficulty	(SE)	Item Location	(SE)
18	1.198	(0.102)	1.577	(0.115)
19	1.917	(0.118)	2.293	(0.105)
20	2.666	(0.128)	3.038	(0.116)
21	1.012	(0.103)	1.395	(0.103)
22	2.762	(0.150)	3.126	(0.163)
23	1.775	(0.126)	2.147	(0.104)
24	1.692	(0.123)	2.063	(0.100)
25	4.249	(0.235)	4.604	(0.232)
26	0.338	(0.103)	0.732	(0.092)
27	3.871	(0.179)	4.229	(0.170)
28	4.496	(0.247)	4.850	(0.238)
29	2.436	(0.135)	3.900	(0.328)
30	2.324	(0.123)	3.772	(0.312)
31	1.443	(0.113)	2.759	(0.268)
32	2.346	(0.132)	3.815	(0.292)
33	3.312	(0.170)	4.861	(0.285)
34	1.131	(0.103)	2.406	(0.277)
35	1.897	(0.122)	3.307	(0.320)
36	8.098	(0.849)	9.765	(0.931)
37	6.527	(0.660)	8.164	(0.682)
38	2.782	(0.125)	4.365	(0.294)
39	3.601	(0.198)	5.236	(0.338)
40	1.785	(0.122)	3.226	(0.287)
41	0.864	(0.105)	1.233	(0.100)
42	1.500	(0.112)	1.863	(0.100)
43	1.131	(0.102)	1.496	(0.094)

	Multilevel Rasch		MD-IR	
Item	Item Difficulty	(SE)	Item Location	(SE)
44	0.729	(0.106)	1.100	(0.117)
45	3.537	(0.219)	3.886	(0.217)
46	0.492	(0.094)	0.869	(0.091)
47	0.582	(0.089)	0.964	(0.095)
48	1.272	(0.104)	1.645	(0.094)
49	0.086	(0.103)	0.481	(0.103)
50	0.687	(0.100)	1.083	(0.090)
51	1.393	(0.109)	1.776	(0.112)
52	1.584	(0.115)	1.965	(0.117)
53	0.276	(0.109)	0.679	(0.089)
54	3.865	(0.189)	4.228	(0.174)
55	1.801	(0.107)	2.178	(0.104)
56	3.939	(0.195)	4.300	(0.194)
57	-1.344	(0.108)	-0.898	(0.108)
58	0.216	(0.109)	0.618	(0.090)
59	2.618	(0.143)	2.986	(0.138)
60	-0.209	(0.100)	0.201	(0.082)
61	1.296	(0.095)	1.675	(0.098)
62	1.401	(0.107)	1.776	(0.114)

BIBLIOGRAPHY

Al-Samarrai, S., Shrestha, U., Hasan, A., Nakajima, N., Santoso, S., & Wijoyo, W. H. A. (2017). *Introducing a performance-based school grant in Jakarta: what do we know about its impact after two years?* The World Bank.

Anderson, J. O., Milford, T., & Ross, S. P. (2009). Multilevel modeling with HLM: Taking a second look at PISA. In *Quality research in literacy and science education* (pp. 263–286). Springer.

Andrabi, T., Das, J., & Khwaja, A. I. (2015). *Report cards: The impact of providing school and child test scores on educational markets*. The World Bank.

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *Journal of classification*, *30*(2), 195–224.

Bacci, S., & Caviezel, V. (2011). Multilevel irt models for the university teaching evaluation. *Journal of Applied Statistics*, *38*(12), 2775–2791.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, *72*(2), 141.

Blum, W. (2015). Quality teaching of mathematical modelling: What do we know, what can we do? In *The proceedings of the 12th international congress on mathematical education* (pp. 73–96).

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

Bouhlila, D. S. (2015). The heyneman–loxley effect revisited in the middle east and north africa: analysis using timss 2007 database. *International Journal of Educational Development*, *42*, 85–95.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral*

Statistics, *39*(5), 333–367.

CCSSO. (2017). *Principles of effective school improvement systems* (Tech. Rep.). One Massachusetts Avenue, NW, Washington, DC 20001: Council of Chief State School Officers.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*(3), 336–370.

Chudgar, A., & Luschei, T. F. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Educational Research Journal*, *46*(3), 626–658.

Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education*, *40*(3), 291–309.

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*(4), 643–659.

Draney, K., & Jeon, M. (2011). Investigating the Saltus model as a tool for setting standards. *Psychological Test and Assessment Modeling*, *53*(4), 486.

DuFour, R. (2004). What is a "professional learning community"? *Educational Leadership*, *61*(8), 6–11.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *49*(2), 175–186.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*(6), 359–374.

Fox, J.-P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, *15*(3-4), 261–280.

Fox, J.-P. (2005). Multilevel irt using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 145–172.

Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2), 271–288.

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(1), 53–70.

Grilli, L., & Rampichini, C. (2009). Multilevel models for the evaluation of educational institutions: A review. In *Statistical methods for the evaluation of educational services and quality of products* (pp. 61–80). Springer.

Güler, H. K., & Arslan, Ç. (2019). Mathematical competencies required by mathematical literacy problems. *Malaysian Online Journal of Educational Sciences*, *7*(2), 57–70.

Hairon, S., Goh, J. W. P., Chua, C. S. K., & Wang, L.-y. (2017). A research agenda for professional learning communities: Moving forward. *Professional development in education*, *43*(1), 72–86.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* (Unpublished doctoral dissertation). ProQuest Information & Learning.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210.

Heyneman, S. P., & Loxley, W. A. (1983). The effect of primary-school quality on academic achievement across twenty-nine high-and low-income countries. *American Journal of sociology*, *88*(6), 1162–1194.

Höhler, J., Hartig, J., & Goldhammer, F. (2010). Modeling the multidimensional structure of students' foreign language competence within and between classrooms. *Psychological Test and Assessment Modeling*, *52*(3), 323.

Huang, F. L. (2010). The role of socioeconomic status and school quality in the philippines: Revisiting the heyneman–loxley effect. *International Journal of Educational Development*, *30*(3), 288–296.

Jeon, M. (2018). A constrained confirmatory mixture irt model: Extensions and estimation of the saltus model using mplus. *The Quantitative Methods for Psychology*, *14*, 120–136.

Jeon, M. (2019). A specialized confirmatory mixture irt modeling approach for multidimensional tests. *Psychological Test and Assessment Modeling*, *61*(1), 91–123.

Jeon, M., De Boeck, P., Li, X., & Lu, Z.-L. (2020). Trivariate theory of mind data analysis

with a conditional joint modeling approach. *Psychometrika*, *85*(2), 398–436.

Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics*, *42*(4), 467–490.

Jeon, M., Draney, K., & Wilson, M. (2015). A general Saltus LLTM-R for cognitive assessments. In *Quantitative psychology research* (pp. 73–90). Springer.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93.

Kamata, A., & Vaughn, B. K. (2011). Multilevel irt modeling. In *Handbook of advanced multilevel analysis* (pp. 49–66). Routledge.

Kohar, A. W., Zulkardi, Z., & Darmawijoyo, D. (2014). Investigating students? difficulties in completing mathematical literacy processes: A case of indonesian 15-year-old students on pisa-like math problems.

Koretz, D. (2008). A measured approach. *American Educator*, *32*(2), 18–39.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. University of Chicago Press.

Kyriakides, L., Creemers, B. P., & Charalambous, E. (2019). Searching for differential teacher and school effectiveness in terms of student socioeconomic status and gender: implications for promoting equity. *School Effectiveness and School Improvement*, *30*(3), 286–308.

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

Li, T., Jiao, H., & Macready, G. B. (2016). Different approaches to covariate inclusion in the mixture rasch model. *Educational and psychological measurement*, *76*(5), 848–872.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*(7), 548–564.

Lubke, G. H., & Muthén, B. (2005). Investigating Population Heterogeneity With Factor Mixture Models. *Psychological Methods*, *10*(1), 21–39. Retrieved 2017-07-26, from `http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.10.1.21` doi: 10.1037/1082-989X.10.1.21

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda. *Journal of Policy Analysis and Management*, *33*(4), 950–976.

Marcoulides, G. A., & Heck, R. H. (2013). Mixture models in education. In *Handbook of quantitative methods for educational research* (pp. 347–366). Brill Sense.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

Mehrans, W., & Cizek, G. (2012). Standard setting for decision making: Classifi cations, consequences, and the common good. In G. Cizek (Ed.), *Setting performance standards* (pp. 58–71). Routledge.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359–381.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika*, *61*(1), 41–71.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585.

Muthén, B. O., & Muthén, L. (2010). Technical appendices. *Los Angeles, CA: Authors*.

Muthén, L., & Muthén, B. (2019). Mplus. *The comprehensive modelling program for applied researchers: user's guide*, *5*.

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education*, *79*(4), 4–32.

OECD. (2010). *Education at a glance 2010: Oecd indicators*. OECD Paris.

OECD. (2013). *PISA 2012 results: What makes schools successful? Resources, policies and practices (Volume IV)*. OECD Publishing Paris, France.

OECD. (2018). *PISA for development assessment and analytic framework: Reading, mathematics and science*. Paris: OECD Publishing.

OECD. (2019). *PISA for development technical report* (Tech. Rep.). Organization for Economic

Co-operation and Development.

Opdenakker, M.-C., & Van Damme, J. (2006). Differences between secondary schools: A study about school context, group composition, school practice, and school effects with special attention to public and catholic schools and types of schools. *School effectiveness and School improvement*, *17*(1), 87–117.

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.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307–335.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219–262.

Sáenz, C. (2009). The role of contextual, conceptual and procedural knowledge in activating mathematical competencies (pisa). *Educational Studies in Mathematics*, *71*(2), 123–143.

Smit, J., Kelderman, H., Flier, H., et al. (2000). Collateral information and mixed Rasch models. *Methods of Psychological Research Online*, *5*(4), 31-43.

Stacey, K. (2015). The international assessment of mathematical literacy: Pisa 2012 framework and items. In *Selected regular lectures from the 12th international congress on mathematical education* (pp. 771–790).

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal*

of Educational Measurement, *44*(4), 313–324.

Strand, S. (2010). Do some schools narrow the gap? differential school effectiveness by ethnicity, gender, poverty, and prior achievement. *School Effectiveness and School Improvement*, *21*(3), 289–314.

Sulis, I., & Toland, M. D. (2017). Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *The Journal of Early Adolescence*, *37*(1), 85–128.

Turner, R., Blum, W., & Niss, M. (2015). Using competencies to explain mathematical item demand: A work in progress. In *Assessing mathematical literacy* (pp. 85–115). Springer.

UNESCO. (2015). *Education 2030 Incheon declaration and framework for action.* UNESCO Paris.

United Nations. (2019). *Sustainable development goals.* Retrieved from `https://sustainabledevelopment.un.org`

van Hek, M., Kraaykamp, G., & Pelzer, B. (2018). Do schools affect girls' and boys' reading performance differently? a multilevel study on the gendered effects of school resources and school practices. *School Effectiveness and School Improvement*, *29*(1), 1–21.

Vermunt, J. K. (2007). Multilevel mixture item response theory models: an application in education testing. *Proceedings of the 56th session of the International Statistical Institute. Lisbon, Portugal*, *2228*.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*(1), 33–51.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450–469.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and*

Assessment Modeling, *52*(1), 8.

von Davier, M. (2014). The log-linear cognitive diagnostic model (lcdm) as a special case of the general diagnostic model (gdm). *ETS Research Report Series*, *2014*(2), 1–13.

Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381–417.

Wang, W. C., & Qiu, X.-L. (2019). Multilevel modeling of cognitive diagnostic assessment: The multilevel DINA example. *Applied Psychological Measurement*, *43*(1), 34–50.

Willms, J. D. (2006). *Learning divides: Ten policy questions about the performance and equity of schools and schooling systems*. UNESCO Institute for Statistics Montreal.

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record*.

Willms, J. D., & Somer, M.-A. (2001). Family, classroom, and school effects on childrens educational outcomes in latin america. *School effectiveness and school improvement*, *12*(4), 409–445.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*(2), 276.