Title

Inequalities in Access to Educational Opportunities: An Investigation of the PISA 2009 Dataset using a Multilevel-IRT Framework

Author

Srinivasan, Jayashri

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Inequalities in Access to Educational Opportunities: An Investigation of the PISA 2009 Dataset using a Multilevel-IRT Framework

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy in Education

by

Jayashri Srinivasan

2021

ABSTRACT OF THE DISSERTATION

Inequalities in Access to Educational Opportunities:
An Investigation of the PISA 2009 Dataset Using a Multilevel-IRT Framework

by

Jayashri Srinivasan

Doctor of Philosophy in Education

University of California, Los Angeles, 2021

Professor Michael Seltzer, Chair

A fundamental goal in education is to provide access to quality education and educational opportunities for every student. Classroom processes, teaching, and students' learning experiences are at the heart of quality education and, as such, must be the key focus in investigating the issues of equity in access to education (O'Sullivan, 2006; Peske & Haycock, 2006; Raudenbush & Sadoff, 2008). In light of India's performance in PISA 2009, this dissertation study investigates the larger issues of access to high quality teaching practices, and other valuable school resources to get a better picture of India's poor performance. To this end, publicly available large-scale datasets such as the Programme of International Student Achievement (PISA; by OECD), and the Teaching and Learning International Survey (TALIS; by OECD) enable us to look beyond student's achievement scores or a country's ranking by providing us with a plethora of information on students, teachers, and schools. Moreover, even

though PISA assessments are low stakes tests, they often drive high stakes education policy decisions in multiple countries.

In this dissertation study, I make use of the PISA student and school questionnaires for India along with state-of the art multilevel IRT models implemented using MCMC. I describe and illustrate a methodology to examine students' exposure to key instructional practices based on students' responses to PISA survey items, and then use this measure as an outcome variable in a three-level IRT model to investigate differences in the amounts of exposure to key practices within schools and between schools. Measurement invariance was established across the rural and urban regions of Himachal Pradesh (HP) and Tamil Nadu (TN) before comparing the construct of interest across various sub-groups. This set of analyses indicates that the items in the student questionnaires capture the construct of interest, and are not an artifact of underlying translation errors, or cultural differences in the examinees understanding of these items.

A multilevel IRT approach, such as the one employed in this dissertation allows us to tease apart the variation in the extent to which students experience particular instructional practices into their within-school and between-school components. The analysis strategies developed in connection with my dissertation will hopefully be valuable to other researchers interested in investigating questions concerning inequality in the distribution of key instructional practices.

Lastly, in chapter 6, I depict the use of this approach to identify schools, whose students on an average, experience relatively high or low exposure to the instructional practices of interest. Futhermore, a key finding of this set of analyses indicated that the that a majority of the public or government-run schools were concentrated in the lowest end of the socio-economic scale; private schools were found to be more spread out, but still in low socio-economic areas.

The dissertation of Jayashri Srinivasan is approved.

Manisha Shah

Noreen Webb

José Felipe Martínez

Minjeong Jeon

Michael Seltzer, Committee Chair

University of California, Los Angeles

2021

To Praveen – my best friend and life partner,

my father, my mother, my brother

... for all their patience, love, and never-ending support. And Viaansh for teaching me immense patience

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my advisor and committee chair, Professor Michael Seltzer for his continued guidance, insights, and support throughout my time at UCLA. This dissertation work would not be possible without his assistance, encouragement, and patience. I would also like to thank, Professor José Felipe Martínez for his tremendous support, and discussions throughout my studies and my dissertation. I am also thankful to the other members of my committee – Professors Noreen Webb, Minjeong Jeon, and Manisha Shah for their helpful feedback and suggestions.

I would also like to thank my friends/colleagues I met at UCLA – Priscilla Liu, Seung Won Chung, Alana Kinarsky, Julie Liao, Joy Zimmerman, Meredith Langi, Melissa Goodnight for their emotional support and advice throughout my time at UCLA.

Finally, I would not have been able to get through my Ph.D years without the constant support from my parents, Sitalakshmi Srinivasan and Srinivasan Sambasivan, and my brother Kritarth. Their endless support, advice, and faith in me allowed me to pursue this journey. And Praveen, thank you for aupportting me throughout my Ph.D process and always being there for me.

VITA

Education

M.Ed., University of Illinois at Chicago (UIC) 2014

Education, Measurement, Evaluation, Statistics and Assessment (MESA)

M.S., University of Illinois at Chicago (UIC) 2014

Physics

Research Experience

UCLA School of Education & Information Studies 2014 – 2020

Graduate Researcher

WORLD Policy Analysis Center Jan – April 2020

Graduate Researcher

Publications

Kloser.M., Edelman.A., Floyd., C., Martinez.J.F., Stecher.B., **Srinivasan.J.,** & Lavin.E. (2020). Interrogating Practice or Show and Tell?: Using a Digital Portfolio to Anchor a Professional Learning Community of Science Teachers. *Journal of Science Teacher Education*.

DOI: 10.1080/1046560X.2020.1808267

Srinivasan.J., Jeon, M., & Seltzer, M. (Invited Revisions). A Cross-Classified Modeling Approach with Small Sample Raters in a Bayesian Framework.

Martinez.J.F., **Srinivasan.J.**, Kloser.M., Stecher.B., Edelman.A. (Invited Revisions). Measuring next generation science instruction using a tablet portfolio app: Reliability, Validity, and Feasibility.

Grants & Fellowships

Dissertation Year Fellowship (DYF) July'18 -June'19

Awarded by UCLA Graduate Division, $20,000

Aimée Dorr Fellowship Jan'18 –June'18

Awarded by UCLA Graduate School of Education & Information Studies, $10,000

Elwood H. Zillgitt & Mildred B. Finney Fellowship Oct'16 –June'17

Dean's Scholar Award, UCLA Graduate School of Education & Information Studies, $12,000

Graduate Summer Research Mentorship (GSRM) Grant June'16–Sep'16

Awarded by UCLA Graduate Division, $6000

Chapter 1

Introduction

"Quality education" has often been associated with students' scores on a variety of tests, since they strongly relate to improvements in skills and enable better social and economic outcomes (e.g., Hill & Chalaux, 2011, p.13). Classroom processes, teaching, and students' learning experiences are at the heart of quality education, and as such, are a key focus in investigating the issues of equity in access to quality education (O'Sullivan, 2006; Peske & Haycock, 2006; Raudenbush & Sadoff, 2008). Raudenbush & Sadoff, (2008) state that when much is known about the innovations and aspects of instruction that are crucial to student development, a fundamental issue that follows centers on "access" – the distribution of classroom quality and the extent to which this varies within schools, and across different demographic and geographic regions (p.140). The importance of access to educational opportunities and access to the same level of resources for all students has been a focus of educational research since the 1980s, and the main focus of research for scholars such as Burstein(1980) and Oakes (1989). In the Indian setting, researchers have found that access to quality education is a concern across several Indian states (e.g., Schleicher, 2013; Walker, 2011).

In 2000, as an initiative towards universal education, India started a program called the *Sarva Shiksha Abhiyan*, which means the "Education for All Program". The program focused on universal elementary education, bridging the gaps in gender and social context concerning education, and enhancement of the learning levels of children. In 2009 the Right to Education (RTE) Act was passed; as per this law, every child between the ages of 6 to 14 will receive "free and compulsory" education, a right that is aimed at reducing child labor and achieving universal education. Furthermore, the law also demands the need for "equitable quality in a formal school

which satisfies certain essential norms and standards" (RTE 2009 gazette). RTE aims at free and compulsory education in India, however, it is also essential to provide students with quality education along with an environment conducive to learning. Globally, many countries have adopted policies for universal education to ensure that every child receives free and compulsory education irrespective of their background.

1.1 Motivation to Study Access to Instructional Practices and School Resources

There has been a rise in student enrollment rates in the past few years across many states in India, however, access to education remains an open question for many subgroups (Lewin, 2011). In countries with highly competitive job markets, such as India, the equal distribution of educational resources is necessary (Shields et al., 2017) as there are usually many candidates for a particular job posting. Research indicates that the boom of the technology industry in India has created a "few" top performers, which tends to conceal the majority of underperformers, especially children from the rural areas (Das & Zajonc, 2010a). Therefore, it is essential that everyone has equal access to educational resources, and the distribution of these opportunities will largely impact students being placed into competitive jobs.

Numerous studies have indicated the need to improve the quality of education in India (e.g., Das & Zajonc, 2010b; Lewin, 2011; Singh & Sarkar, 2015). Two examples that point towards the need to examine student's performance in India come from the large-scale assessments conducted nationally as well as internationally. First, students' poor performance at a national scale is evident from the Annual Status of Education Report (ASER) study, which conducts household surveys across 600 rural districts in India and reports on children's basic learning skills in mathematics and English for children between the ages of 6 to 15. The ASER survey collects information on family composition, household items (e.g., presence of TV,

refrigerator, etc) available to the family at the time of the survey, parents' education levels, and gives children between the ages of 6 to 15 an in-home test. The survey also collects information on whether the children in the household go to school, if they attend the nearby government school, or if they have dropped out of school. In 2010, among 195 government primary schools sampled in the state of Himachal Pradesh in the North of India, only 53.9% of third-graders could do subtraction and only 25.3% could read text that was at the grade-two level. In Tamil Nadu, a state in the south of India, among the 395 government primary schools that participated in ASER, it was found that only 16.4% of third-graders could do subtraction and 7.2% could read a second-grade level text.

Another example indicating India's performance on a global scale comes from India's participation in the large-scale international assessment PISA, conducted by OECD every three years. In 2009, two states from India, Himachal Pradesh and Tamil Nadu participated in the PISA tests. These assessments are administered to 15-year-olds enrolled in school, and India ranked low in the 72nd and 74th positions among the 74 participating countries on reading literacy. While this ranking at the global scale points towards students' and schools' poor performance on overall reading literacy in the PISA test, the Government of India was far from accepting the results and blamed the results on socio-cultural issues in the PISA tests (Venkatachalam, 2017). This was the only time Indian states participated in the PISA tests.

These reports raise some fundamental questions: Why are students not performing well? What are the reasons or factors for some students to excel while others tend to be at the bottom? School assessments largely assess children's learning outcomes, which depend on several factors, for example, the quality of teaching, retention of quality teachers in schools, the availability of instructional materials, and other various school resources (Kingdon, 2007). Indicators such as

classroom instruction and school resources mediate and shape classroom learning while providing vital information about student performance (Oakes, 1989). Much of the emphasis in investigating educational inequality has primarily focused on student outcomes or test scores as the outcome variables, and teacher practices have often been the predictors in such studies. In light of the RTE act and student's performance on various large-scale assessments, investigating the distribution of instructional practices and school resources directly, instead of test scores and/ or performance data, can be pivotal in monitoring and accountability efforts of education systems, when making decisions about school's resources, policies, and processes. In particular, equitable access to instructional practices has been a challenge across schools in the U.S. (Cardichon et al., 2020), and in India, few studies have examined teacher quality and classroom practices across schools (e.g., Hill & Chalaux, 2011; Singh & Sarkar, 2015).

Every developed as well as a developing nation has at some point in time focused its efforts on ensuring "education for all". Even though governments and policymakers might intend to make primary and secondary education accessible for all, the distribution of the opportunities available to children varies considerably amongst various groups. For example, in a poor family, boys might be given preference to go to school over girls, or families with higher wealth might be able to afford better schooling and educational resources in comparison to their counterparts who might be less wealthy and/or hail from a rural village. These examples raise a fundamental question about educational opportunities and access to these opportunities. A variety of reasons can contribute to lack of access, such as the location of the school, unavailability of school programs or resources, funding challenges, etc. The ASER data indicates that there is a substantial lack of school resources (e.g., lack of qualified teachers, shortage of instructional materials) across many government schools in rural areas. Large scale assessments such as ASER and PISA

can be useful tools to assess the distribution of school resources (e.g., lack of qualified teachers, shortage of instructional materials), and school curriculum (e.g., activities offered by the school, differentiation of instruction) across countries or within countries (e.g., in rural and urban regions). This information can be useful to policymakers and/or government officials to make decisions regarding the allocation of resources to those who have fewer opportunities, thus reducing inequities across the nation.

1.2 Methodological Approach

Over the past few years, large-scale international assessments like PISA and TIMSS have been adopted by many countries as benchmark tests to make policy-related decisions. Moreover, even though the PISA tests are considered low stakes tests, the results often drive high stakes policy decisions with a strong emphasis on education policy. This level of attention demands the need for a more careful look into measurement and methodological issues, especially when these tests are adopted to guide policies in several countries or economies.

The PISA dataset is administered to 15-year-old students who are enrolled in schools. First administered in the year 2000, PISA tests are conducted every three years by the OECD. As a part of the PISA tests, a variety of information is gathered from students, including scores on test items in reading, mathematics, and science domains, and background questionnaires completed by students as well as school principals. These datasets often contain rich student and school variables in a wide range of areas. For example, the student questionnaires often ask students about their classroom activities, classroom climate, and specific items about their teachers (e.g., if the teacher asks questions in class; if the teacher assists students with their assignment). These PISA tests undergo a rigorous process for item development and field testing before they are administered to students. The PISA 2009 technical manual (OECD, 2012)

provides a detailed description of all aspects of the PISA tests – the development of the items in the various domains and background questionnaires, its administration, and data analysis across multiple countries in the study.

The PISA 2009 data for India have been utilized in few studies (e.g. Areepattamannil, 2014), and primarily study the effects of various student and school-level factors on mathematics and reading outcome scores. Although the use of datasets like PISA or other large-scale assessments for education policy decision processes has been controversial with a large group of supporters and critics, they could be beneficial for countries like India, where there is a dire need for education policy reforms – the gap between the number of skilled workers, that is, those who have attended a college or technical school and non-skilled workers is increasing at a rapid rate, and education is not affordable to many (Kingdon, 2007). In the past few years, India's primary focus has been on measures such as availability of school resources and enrollment numbers of students, but with the advent of large-scale data sources in India (e.g., National Assessment Survey (NAS), ASER), there has been a shift to focus on curriculum and classroom process.

Notably, the current dissertation work is unlike many other studies in India (e.g., Das & Zajonc, 2010; Kingdon, 2007) where the primary focus was on student's cognitive scores to assess equity; instead, I will directly look at student's access to key instructional practices, and other school resources to assess the equity of educational opportunities using the PISA 2009 dataset. The items employed in this dissertation study come from the background questionnaires completed by students and school principals. The use of these non-cognitive survey items allows us to look beyond the basic indicators (e.g., proportion of enrolled children) to gain insights into important questions about the distribution of teaching practices and school resources. Furthermore, another important impetus for doing this methodological work is that many

educational indicators are often based on, for example, principal's or district officials responses to survey questions about shortages of qualified teachers, or teacher absenteeism rates, and the like. Such factors are related to student learning, but can be poor substitutes for directly measuring the extent to which students experience key instructional practices. To this end, I make use of the construct labeled as teachers' stimulation of reading engagement (STIMREAD) from PISA 2009's student questionnaires as outcome variables as a way to understand inequalities in access to these instructional practices. These questionnaires were completed by students and the STIMREAD construct is measured using seven items (see Table 5.1 for item descriptions). Such indicators allow us to capture the amount of exposure students have to various key aspects of instruction, and how equitably exposure to these practices are distributed within schools and between schools.

Another key aspect when using large-scale assessments is the need to ensure that the test items are perceived similarly across various sub-groups (e.g., HP rural and HP urban). For example, one of the major criticism of PISA tests is that it fails to account for the differences across countries in terms of their living conditions and culture. In particular, India is a very diverse nation with 30 states and almost every state has its own spoken and/or written language. This cultural diversity could pose concerns for large-scale testing across the nation. Since a majority of the students attend government schools in rural areas, often English or Hindi, which are widely spoken languages across India may not be the preferred language or test language. Therefore, when an assessment or test is administered across multiple states, it becomes crucial to ensure that the differences between various groups (e.g., rural/ urban regions in a country) are trustworthy and not an artifact of underlying translation errors or cultural differences in the examinees understanding of the items. To this end, measurement invariance checks allow us to

make sure that the various groups in the study similarly perceive the items and the underlying construct is comparable across the sub-populations, thereby reducing the chances of biases and unfairness. Such an analysis, allows us to ensure that the estimates we obtained from fitting a model to multiple groups are comparable across the sub-groups and the construct of interest or items across the different groups (e.g., rural and urban regions or private and public schools) are compared on a common measurement scale.

Hierarchical data structures are common in education research due to the hierarchical structure of schooling (e.g., students nested within different classrooms, which are nested in different schools). Scholars like Oakes (1989), and Burstein (1980) have encouraged researchers to make use of large-scale assessments and multilevel models to assess students' access to resources, as not all students in a school will have access to the same level of resources. Using a multilevel model, to model the student-level and school-level data of PISA tests allows us to capture both the within-school and the between-school variations. We can examine the factors associated with differences in students' ratings of teachers with respect to their instructional practices, and the factors that might be related to between-school differences, such as school mean ESCS, and teacher shortage. In particular, we can study the distribution of learning opportunities and exposure to key instructional strategies in a given school with respect to a student's socioeconomic status – that is, whether these opportunities are equitably distributed for students. A multilevel Item Response Theory (IRT) framework, in particular, allows us to investigate the substantive issues related to equity of access to teachers' practices by using the student's responses to surveys to better understand the school conditions, for example, within and across rural and urban regions of India.

To this end, another concern with the PISA 2009, and the motivation to make use of the test items directly in this study is PISA's use of scaled scores for composite indices obtained via an IRT model (e.g., Partial Credit Model (PCM)) while ignoring the nesting of students within different schools. These composite indices are often created using a set of items (e.g., 5 to 10 items) and these composite indices are often employed by secondary analysts for research. In such an approach, most times the estimates for latent variables (e.g., person abilities) are shrunk towards a grand mean, thus resulting in a bias in the estimates and a homogenization of the estimates. This motivates working directly with the actual item responses, and specifying and fitting a 3-level multilevel IRT model with level-1 modeling the item responses. A multilevel IRT framework allows us to easily handle Likert scale items as the outcome variables (common among education research) and captures the within-school and between-school differences with respect to the various student- and school-level characteristics. Note also, that if we work with estimates of the student latent variables of interest created in a separate analysis, we would essentially be back in a two-level model situation in which the ability parameters for the students are treated as outcomes in the student-level model. If standard errors of measurement are not available for these estimates, then the estimate we obtain for the within-school variance component would reflect both measurement error connected with the estimates and the actual differences across students within schools in the extent to which they experience certain practices of interest. The latent scale of interest is formed using the responses to the items, hence the estimate we obtain for the within-school component may be fairly large, but a substantial portion of that estimate may reflect the error. A latent variable modeling approach provides a comprehensive framework for handling measurement error, in which the measurement models link the observed indicators (e.g., teacher survey responses, student characteristics) to the latent variable or the construct of interest.

To estimate these complex multilevel IRT models one can employ a Maximum Likelihood (ML) or a Bayesian approach using MCMC. In this study, I make use of MCMC estimation to obtain the various parameters of interest. A Bayesian approach, although in large–scale data sets may produce estimates similar to those obtained from ML estimation, has multiple advantages. Firstly, it provides flexibility in fitting a variety of models with ease. Secondly, Bayesian inference and the interpretations of various parameters along with the credible intervals is different from that of ML estimation and can be considered to be more intuitive. Credible intervals, in a Bayesian framework, captures the uncertainty in our parameters, and the 95% credible interval is the central portion of the posterior distribution that contains 95% of the values. The 95% credible intervals can be interpreted as a probabilistic value, that is "given the observed data, there is a 95% probability that the true value of the parameter falls within the interval" while the 95% confidence intervals obtained for, say, the ML estimates would be "that we are 95% confident that the true value of the parameter is contained within the lower bound and the upper bound if we repeated the study many times. Thirdly, the plots of posterior distributions of parameters of interest, the posterior means, and medians and other quantiles can help us get a much better sense of what the data are telling us in comparison to just point estimates. Such an approach requires specifying priors for the parameters in the model. In many cases, we may have little prior knowledge about the magnitude of various parameters in our models. In some cases, however, findings in previous relevant studies or assessments may provide the basis for specifying priors that contribute valuable information in our analyses. An improved, and refined interpretation of the results can assist us while making decisions that impact many students and families.

1.3 Research Goal

In light of India's poor performance on the PISA tests and the current situation of Indian education, primarily in rural areas, this study seeks to investigate inequalities in access to instructional practices, and differences across schools with respect to a number of school resources (e.g., availability of qualified teachers) for students enrolled in schools in the rural and urban regions. The PISA data allows us to investigate how equitably various key instructional practices, and other school resources are distributed across students and schools, and the extent to which they differ within and across rural and urban regions. The central Research Questions (RQ) for this dissertation study are:

RQ1. To what extent do school resources, such as lack of qualified teachers, and teacher absenteeism differ between schools within the rural and urban regions of HP and TN? What are the limitations and problems of working with indicators of such resources based on principals' responses? What critical information about the quality of students' educational experiences is missing, especially in the area of language instruction?

RQ2a. Does measurement invariance hold for the items that form the teachers' stimulation of reading engagement (STIMREAD) construct across rural and urban regions?

RQ2b. What possibilities arise when we employ student STIMREAD values as outcomes in multilevel models in which students are nested within different schools? Are there appreciable differences in teachers' instructional practices across the rural and urban regions? To what extent does accounting for different student and school-level characteristics (e.g., student SES, teacher shortage at a school) explain differences in exposure to these practices?

RQ3: What do the school-specific estimates of students' exposure to STIMREAD practices look like across public and private schools in TN's rural region?

1.4 Significance of the Study

In recent years, there have been numerous initiatives by the Government of India (e.g., NAS by NCERT, NITI Ayog (a policy think tank by GoI)) to improve the Indian education system aimed towards education for all. This study provides another lens to assess equity and access in education. In particular, it adds to the existing work in three important ways:

(a) To our knowledge, no other study in the Indian context has treated instructional practices and other school resources as outcome variables and assessed the distribution of these resources within and between schools, and between different regions, in contrast to focusing on student test scores or enrollment numbers.

(b) Few studies have made use of large-scale international assessments to examine education equality and access to educational opportunities, particularly while employing non-cognitive background questionnaires. Even though this study focuses on India, the strategies developed in connection with my dissertation will be valuable to other researchers interested in investigating questions concerning inequity in the distribution of key instructional practices. In particular, in countries like the United States, where education equity is a pressing concern, international assessments can be complemented by other large-scale assessments such as NAEP,

(c) Large-scale international assessments are used across multiple countries for a variety of educational and non-educational policy decisions. To this end, the current work adds to the sparse existing literature (e.g., Asil & Brown, 2016; Rutkowski & Svetina, 2014) as few authors have investigated the issue of measurement invariance in large-scale assessments within countries. A multilevel modeling framework allows us to discern the differences in access to instructional

practices as a function of student and school-level demographic factors, such as school funding, and whether a school is public or private or school funding. We can tease apart the variation in the extent to which students experience a particular instructional practice into “within” and “between” school components. It will allow us to identify outlying schools as well, for example, a school in a high poverty region in which students have access to high-quality educational opportunities.

(d) Lastly, multiple nations have employed the Sustainable Development Goals (SDG) 2030 agenda to further their education initiatives. Specifically, the SDG 4 goal aims to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. With the Covid-19 pandemic raging across every nation in the world, it is of utmost importance to pay close attention to the inequalities in education, predominantly in the low-income nations where students are dropping out of schools and digital learning is out of reach for millions. Availability of quality instruction and resources for all is crucial now and in the future.

Chapter 2

Inequalities and Access to Educational Opportunities

Access to educational opportunities is a fundamental right for every student. Most nations in the world have instituted laws and policies to provide equal access to educational opportunities for all students. To that end, the goal of this chapter is: (a) to examine the importance of access to educational opportunities, globally, as well as in India; (b) Next, since the focus of this dissertation work is on teachers' classroom practices and various school resources, I will discuss the importance of these opportunities in light of students' education, and its impact on educational outcomes and academic growth; (c) Since the focus is on India, I will briefly examine the education system in India, education equity and access to opportunities, and discuss schooling in rural versus urban settings in India; and (d) Lastly, I discuss the use of large-scale international assessments, such as PISA to examine access to educational opportunities and how large-scale datasets allow us to gain better insights into classroom teaching practices.

2.1 Equality and Access to educational opportunities

Educational equity and access to educational opportunities (e.g., school resources, availability of qualified teachers, or availability of instructional materials) have been of concern since the 1970s (see, e.g., National Academies of Sciences, 2019; National Center for Education Statistics (DHEW) et al., 1978; Oakes, 1989; Shavelson et al., 1989). Equality of opportunity in education is noted in Article 1 of the Convention against Discrimination in Education (UNESCO, 1960), which notes that no person or group of persons should be deprived of access to education of any type or at any level or limit any person or group of persons to education of an inferior standard. Additionally, Coleman, (1968) and the National Center for Education Statistics (DHEW) et al., (1978) note that equality of education implies that students are provided with free

education with a common curriculum for all children regardless of their background, and opportunities for students to rise to their full potential. Children from diverse backgrounds should be able to attend the same school irrespective of the locality they reside in. For example, in the US high-quality secondary education is often accessible to families who can afford housing in middle-class neighborhoods, or to those who can afford private school tuitions (Shields et al., 2017). Such glaring examples exhibit the unequal distribution of educational opportunities in the US and similar examples are seen in various countries around the globe. To address these concerns in the U.S., the National Academies of Sciences, (2019) has developed a set of educational equity indicators as a starting point for addressing the longstanding disparities in key educational opportunities and outcomes (see pg. 5-6 of this report for a summary of the indicators).

In South Asian countries, from 1990 to 2000 significant improvements were made in access to, and participation in, primary education, but at the same time the provision of these educational opportunities has been accompanied by diminution of learning quality (Maclean & Vine, 2003, p.24). For example, the authors note that even though, in India the enrollment rate for children in primary school might be 75%, only 50% complete grade 5. Moreover, among these 50% students, only 50% of those students further attain the expected competencies in language and mathematics – most children barely achieve a sustainable level of literacy and numeracy (p.158). Maclean, (2003) notes that in the Asia-Pacific region when it comes to access and equity, and equality of opportunity in education, the situation still needs considerable attention, as there are large populations that are completely excluded from quality and effective education and schooling. He states,

> "As a general rule, those living in cities have a better chance of receiving a good education than do those in remote rural or isolated parts of a country. Moreover, boys have a better chance of achieving an effective education than do girls; the rich have greater opportunities for a good education than do those living under conditions of poverty; and those who belong to the mainstream of society have access to better educational opportunities than do ethnic, racial and religious minorities (p.144)".

This is particularly true in India where access to education is a challenge with variations across and within states in the country. The divide is large among various lower caste and minority groups (e.g., girls, and some religious groups like Muslims) (Lewin, 2011). Due to the enactment of policies such as the Right to Education (RTE) and *Rashtriya Madhyamic Shiksha Abhiyan* (RMSA)[1] access to education and various opportunities has gained a lot of attention and is slowly improving. These policies have also helped improve education equality. More recently, motivated by the SDG 2030 goals, India has been working towards improving and enforcing various laws and policies aimed at educational equality for all. To this end, it is vital to make sure that the basic school resources are available to all students across all schools. In recent years, multiple researchers (e.g., Govinda & Bandyopadhyay, (2008); Lewin, (2011)) have examined access to education and educational equity in India.

2.1.1 Looking beyond enrollment numbers

"An ideal education system would make sure that all kids are enrolled and would ensure achievement is uncorrelated to pupils' socio-economic background**"** write Gamboa &

[1]Launched in March, 2009, this scheme's objective was to enhance access to secondary education and to improve its quality. One of the main goals of this scheme was to provide a secondary school within reach for residents. Other objectives included improving quality of education imparted at the secondary level, reducing gender inequalities to name few.

Waltenberg, (2015, p. 321)**.** Access to educational opportunities for students is impacted by factors such as enrollment in a private or public school (Ammermueller, 2005), institutional factors such as the amount of instructional time (Oakes, 1989), the location a student resides in, and the wealth of a student's family (Govinda & Bandyopadhyay, 2008; Lewin, 2011). Most research studies have primarily focused on student enrollment numbers as an indicator to assess education quality and access to schooling. Even though primary education is free for many students in India, to achieve universal education it is essential to have a strong monitoring and assessment system (Govinda & Bandyopadhyay, 2008; Hill & Chalaux, 2011), which looks beyond enrollment as the primary outcome.

Enrollment numbers help us better understand the age at which children enter schools and how these students flow through the education system. For example, high enrollment numbers give us an indication about the number of students in school, however, this does not provide us information about students' achievement scores, classroom resources in a given school, teaching activities, or other pertinent student, school, or parental factors. In addition, in India, household income was found to be a strong factor impacting enrollment into secondary schooling and was found to be one of the main reasons for gaps in access to education both at the primary and the secondary level, as schooling costs are substantial (Lewin, 2011, p.388). Four factors that have impacted student's access to schooling and influenced exclusion from schools have been gender, disadvantaged social groups (e.g., scheduled caste or tribe caste groups), location of schools, and poverty (Govinda & Bandyopadhyay, 2008, p.72).

2.1.2 Impact of School Location on Access

In diverse nations like India, the location of a school plays a vital role in a child's schooling. For example, a large population of Indians live in rural India, and students often have

limited options in selecting a school to study in. To better understand the impact of being in a rural area with low accessibility to higher secondary schools, we must understand the degree of urbanization in a country (Ammermueller, 2005, p.22). The rural/urban divide is visible in India (Das & Zajonc, 2010a; Govinda & Bandyopadhyay, 2008) and this particularly impacts students who live in remote areas of India with a lack of access to schooling. Special measures and programs by the local and state governments of India help students enroll in schools, however, it is seen that these students tend not to progress through grades (Govinda & Bandyopadhyay, 2008, p.72). The situation is often worse for girls from low-income families residing in rural areas of India, as educating boys is usually preferred over girls. As a result, girls tend to be excluded from school when the schools are in remote locations, as families might be concerned to send girls to far locations (e.g., for safety concerns).

2.2 Access to Instructional Practices

Student test scores or achievement scores have been commonly used as the outcome variable when measuring students' classroom performance. Test scores provide us with information about student learning and proficiencies in various subject areas and are often used to determine whether a child is promoted to the next grade or not. Also, they are used as a proxy for measuring school quality, school climate, or other contextual characteristics. Although, students' performance data and data on prior ability can be useful to distinguish learning within a given setting and judge how students benefit at different levels (e.g., schools, classrooms), various student characteristics (e.g., gender, socio-economic status) are essential for monitoring equity among students (Burstein, 1989). Regarding instructional experiences, Burstein (1989, p.4) states importantly:

> "Information about actual topic coverage and instructional methods are of even greater value [when interpreting student performance data in addition to classroom and school-level information]".

The growing inequalities in access to educational opportunities are a pressing concern for nations like India, where there are wide differences in demographics across geographical regions. While discussing the relationship between access, and knowledge and skills with student outcomes, Oakes, (1989) states,

> "access is a matter over which schools have considerable influence and control. This clear connection between access and policy makes access an appropriate focus of monitoring (p.46)"

This idea can easily be extended and developed to better our understanding of what access looks like for students. Teaching is a social (or group) activity, not an isolated one; therefore, the exposure to various instructional practices students receive in classrooms cannot be matched by other activities outside a classroom or school environment. Teachers' classroom practices can be valuable indicators in comparison to enrollment numbers to monitor school quality, as they are vital for a student's academic growth and achievement. The distribution of instructional practice indicators among various sub-groups (Raudenbush & Sadoff, 2008; Shields et al., 2017) can give us insights into the amount of exposure a student receives to a particular classroom instructional practice and the extent and kinds of learning opportunities a child receives.

Teachers' instructional practices have been found to have a strong impact on students' achievement scores (e.g., Kane and Staiger, 2012; Muijs, 2006) and even though classroom instruction is key for quality and effective education, policy efforts have often focused on educational outcomes, and classroom instruction has received less attention in comparison to

outcomes (Correnti & Martínez, 2012). A variety of measures of instruction and classroom processes have been developed for multiple research and policy needs; Correnti & Martínez, (2012, p.52) categorize these into six broad categories. For example, a few that stand out as key to the premise of this study are "assessing the extent to which different learning opportunities are equitably distributed across classrooms and schools", and "understanding and comparing instructional practices and classroom processes across localities, states, and countries". Across schools in India, few studies have examined the importance of teacher quality (e.g., Hill & Chalaux, 2011; Singh & Sarkar, 2015). Singh & Sarkar, (2015) studied the impact of teacher quality in India on student outcomes across public and private schools. They make use of a multitude of approaches (e.g., teacher and child questionnaires, school observations) to study various teacher quality factors such as teacher characteristics (e.g., subject knowledge) and qualifications (e.g., experience). Their results suggest that across both public and private schools, teacher characteristics and practices (e.g., teachers checking homework regularly) have a significant impact on student's learning outcomes. Studies such as these reflect the importance of teacher quality and the impact they have on students.

Student surveys to measure classroom practices

Classroom practices can be measured using multiple methods, namely classroom observations, teacher logs, and questionnaires (e.g., teacher or student surveys). To that end, questionnaires have been commonly adopted for capturing teachers' classroom behavior or pedagogical knowledge as they are affordable to administer and less time-consuming (Muijs, 2006). Student surveys have been widely employed in higher education, and are often used for teacher evaluations in universities. Although student questionnaires have gained popularity in recent years, there is still some apprehension about their validity, as student responses may not

be very accurate in the K-12 school setting. But, as noted by Muijs (2006) students in grade eight and higher could provide us with useful information about their teachers as well as their classroom activities. Recently, student surveys are being used widely across various school districts in the US and were employed in the large-scale MET study by making use of the Tripod survey (Kane & Staiger, 2012).

In addition to the student cognitive scores, PISA also gathers student and school questionnaires. In the current dissertation study, I make use of the PISA student questionnaire which is completed by students who give their responses to a wide range of items about their classroom experiences, in addition to their background characteristics (e.g., their family, home, reading activities, etc). The questionnaires will be discussed in more detail in chapters 4 and 5.

2.3 Access to School Resources

School resources comprise another set of factors that require attention. The goal of "education for all" aims for every student to have access to high-quality schooling irrespective of their financial background. This places a large amount of burden on the school systems to provide all students with access to quality school resources. Schools need to be well-equipped to provide all students with a high-quality education. Measures of schools' resources are excellent diagnostics of what schools are capable of providing to their students. As Oakes, (1989, p. 187) notes, "these resources define the outer limits of what is possible [by schools]". For example, students' opportunities to learn and perform well in academics are dependent upon schools' having enough resources to retain well-qualified teachers and to be able to fund small class sizes.

2.4 The case of India

The focus of this dissertation work is to make use of the PISA large-scale international assessment data for India to examine students' exposure to instructional practices and school

resources. Before we discuss the statistical models that we will employ in this study, it is essential to get a clear picture of the education system in India, the research carried out on access and equity in India, and schooling in rural versus urban regions in India.

2.4.1 A Brief Description of the Education System in India

The RTE act came into effect in 2010 in India. As a result of this legislation, every child has a right to full-time elementary education, and it was the responsibility of the government of India to ensure that every child between the ages of 6 to 14 is enrolled in a school and completes elementary education. Education across most states in India follows a 10+2+3 system, which is 8 years in primary school and 2 years in secondary school, followed by two years of senior secondary schooling, and three years of education leading to a bachelor's degree (Kharpade, 2002). In most cases, children start grade one at the age of six. Secondary schooling in India is divided into two cycles – one is secondary school comprising grades 9-10 or grades 8-10 in some states, and the second one is upper secondary comprising grades 11 and 12 (Lewin, 2012). There are primarily four types of schools in India, namely, government schools, private aided schools, private unaided schools, and unrecognized private schools. Private schools can be run by an individual or a private organization, and schools that receive part or full support from the government are considered to be private aided schools. Private unaided schools do not receive any support or funds from the government for maintenance. (For specific details about various types of schools, please refer to Lewin, 2012, p. 384). In general, schooling in India is the responsibility of the states under the government of India and most government schools are funded by state budgets.

Schools in India, both government-run and private, follow either a central or state board of education – that is, a governing body that sets the curriculum and provides guidance to schools.

Two boards, the Central Board of Secondary Education (CBSE) and the Council of Indian School Certificate Examination (CISCE) have a uniform curriculum across the entire country, hence fall under the category of "central boards". Furthermore, these boards conduct examinations for 10^{th} and 12^{th} graders, often termed "board" exams. The scores of these examinations play a very important role in a child's education and career, in general. Apart from the central boards, each state in India has a "state government board", wherein the curriculum and other related decisions regarding education are the responsibility of the state. In most cases, the first language or medium of instruction in the schools that follow a state board will be the official language of the state. In Himachal Pradesh (HP) and Tamil Nadu (TN), the official languages are Hindi and Tamil, respectively. In HP many schools teach in English as well.

Himachal Pradesh in the fifties and sixties was considered one of the most backward states in the hilly Himalayan regions of India – an underdeveloped and extremely poor state, an argument used against it to become an independent state (Drèze & Sen, 2002, p.102). However, in the last few years Himachal Pradesh has seen tremendous growth, particularly in the field of basic education, and today it is at the forefront of providing every child primary and secondary education.

2.4.2 Access to Opportunities and Education Equity in India

A fundamental question to think about is why should we care about inequality in access to opportunities in India? And what does that look like? In 1989, the National Council of Educational Research and Training (NCERT) conducted a baseline study with a sample size of 65,842 students to assess their basic language and arithmetic skills (Maclean & Vine, 2003). In a larger study in 1997, the NCERT baseline study revealed a wide range for the achievement levels for fourth and fifth graders. For language, the mean scores ranged from 16 % in Madhya

Pradesh to 74% in Kerala. Furthermore, in some states children in grades 4 or 5 had only mastered less than half the curriculum taught the previous year (World Bank, 1997, p. 84). Particularly, in rural India, many students in grades 4 or 5 had not mastered the skills taught at grade 2 (Banerji et al., 2013; Govinda et al., 1993).

More recently the ASER 2009 report (*Annual Status of Education Report (Rural) 2009*, 2010) found that in Himachal Pradesh, 80.4% of the children in the ages range 11-14 years were enrolled in government schools, and 83.4 % of children in the age range 15-16 years were enrolled in government schools. The data found that nearly 4% of the children in the age range of 15-16 years were not in schools, that is these children either dropped out of schools or never enrolled in schools. Furthermore, 73.2 % of the students enrolled in grade 5 could read grade 2 level text and the learning levels for students enrolled in government schools were found to be lower than those enrolled in private schools. In Tamil Nadu, 73.9 % of the students in the age range of 15-16 years were enrolled in government schools and 10% were not enrolled in schools at all. Only 35.3% of the students enrolled in grade 5 could read grade 2 level text. Although the enrollment numbers and the percentages of students with reading skills required at grade level were found to be improving, there is still a lot of room for improvement. For example, in the ASER 2018 report, it was found that in the age range of 11 to 14 years, 65% of the students were enrolled in government schools and 30.6% of students were enrolled in private schools. In the age range of 15-16 years, 57.4 % were enrolled in government schools and only 50% of the students enrolled in grade 5 could read grade 2 level text. In 2016 this number was 47.9%; therefore, we see only a slight improvement in the numbers.

To judge the quality of education in India and to evaluate the implementation of the *Sarva Shiksha Abhiyan* (SSA), students' enrolled in third, fifth, and eighth grades participated in a large-

scale assessment known as the National Assessment Survey (NAS, 2014). The assessment has been conducted periodically since 2001 to assess the health of the Indian education system as a whole and monitor childrens' learning levels. The goal of NAS is to understand the key indicators of quality education to improve childrens' learning equitably over time. These assessments aim to get a better picture of students' knowledge in specific grades while identifying gaps in areas that need improvement. The results from these assessments are reported on the national and state level, which can help guide policies and interventions for improving children's learning under the SSA program.

In the most recent NAS report, significant differences were found among the average achievement levels of students between the Indian states (NAS, 2014, p.7), and large inequities were seen in science, mathematics, and reading outcomes for students in the rural and urban areas of India. In Himachal Pradesh, the NAS study included 92% of schools from rural areas, and in Tamil Nadu, the sample consisted of 65% of schools from rural areas. Students' performed better in reading comprehension and mathematics in HP, with a large number of students in the 90^{th} percentile compared to their counterparts in TN, where the average scores in both reading and mathematics were below the national average of 250 points. On a scale of 0 to 500, students in TN scored 241 on reading and 229 on mathematics. In HP on average students scored 259 on reading and 248 on mathematics.

In India, even though the government of India provides the country with guidelines, the individual states must ensure the availability of various facilities and resources to students and are responsible to provide their citizens with public amenities. The differences in policies across states give rise to inequality of educational opportunities (Asadullah & Yalonetzky, 2012). The increased gender gaps and particularly, the low participation of women across Indian states, the

differences in access to public infrastructure by various social groups such as Muslims or lower caste men and women, and the impact of education policies due to caste- discrimination can influence the inequality in educational opportunities. Knowledge of these various factors/ reasons helps us better understand why there can be inequality in access to educational opportunities.

2.4.3 Schooling in the Rural versus Urban setting

A large portion of the population of India lives in rural areas and almost 80% of the schools in India are government-run (NUEPA, 2011). Most of the students in rural areas often go to government schools, which lack basic educational resources (e.g., blackboards, tables, and chairs for students) and have high teacher absenteeism rates (Kingdon, 2007), hence making students disinterested in schooling. It is due to these reasons that we are seeing a rise in private and low-fee private schools (Woodhead et.al, 2013; Chudgar & Quin, 2012) and parents prefer to send their children to these schools even though they might be expensive (Kingdon, 2007). Private schools are often expensive and less prominent in rural areas. Students attending schools in rural areas of India often have inadequate access to schools, experience poor quality of teaching, and are forced to travel large distances for schooling (Agrawal, 2014). Maclean & Vine, (2003) note that there are large variations in the learning achievement between rural and urban areas, and the mean levels of performance between the districts within states. The significant disparity in educational attainment between the rural and urban regions in India increases the educational inequality in India, thereby contributing to the rural-urban divide. Only a few studies in India have looked into the quality of schools, particularly government schools (e.g., Kingdon, 2007; ASER, 2014) and access to quality education (e.g., Schleicher, 2013; Walker, 2011), which is a major concern in both rural and urban areas. These points motivate the examination of differences in exposure to access to educational opportunities across rural and urban regions.

Disparities are noticed among childrens' participation in primary and upper primary levels in rural and urban areas. Children in rural areas seem to start schooling late in comparison to their counterparts in urban areas, one possible reason being that urban children tend to live nearer to schools than rural children (Govinda & Bandyopadhyay, (2008)). At higher grades, Govinda & Bandyopadhyay, (2008) report that the rural and urban gap persists with about 68% of children between ages 6-14 years attending school in rural areas, while on the other hand, 81% of children in this age range are attending school in urban areas. These numbers are significantly worse for girls. Additionally, the rural areas of India have been found to have the poorest schools in the country with a majority of untrained or under-trained teachers (Govinda & Bandyopadhyay, 2008, p.27). Teachers often lack the necessary conditions to ensure teaching and student learning.

Private schools are growing at a fast pace in urban areas, and since rural areas primarily consist of government schools, parents often prefer to send their kids to private schools even if they are farther away from home. In particular, low-fee private schools, schools that are run privately but at lower costs are becoming popular in certain parts of urban India and are outnumbering public schools (Tooley and Dixon, 2003).

2.5 Use of Large-Scale Assessments to Examine Access and Equity

Large-scale international assessments have been an excellent source of data for many countries where testing is scarce (e.g., India) or expensive to administer, while these are employed as an additional source of information for policy decisions in many other countries (e.g., Germany, United States). These large-scale data sources can provide us with rich information about various student and school characteristics, such as students' background characteristics, parental information, students' classroom activities, information on various teaching activities, and the like. In addition to the student test scores, the additional contextual information provided

by these background questionnaires can assist in teasing apart the factors that impact students' schooling. Recently, large-scale assessments such as ECLS-K and the MET project have been used to investigate the impact of teacher practices on student learning. Datasets such as PISA allow us to measure access to instructional practices directly using student responses to teachers' classroom activities, and the test items undergo multiple rounds of revisions to validate the items in the assessments. Therefore, large-scale assessments such as PISA, TIMSS, and more recently TALIS and the MET study, provide us with some insights into the teachers' classroom activities by making use of a combination of questionnaires, video recordings, and/ or observations.

In India, large-scale assessments such as NAS (administered by the Government of India), and ASER (administered by a non-profit called Pratham) have been employed to study students' academic growth and the condition of schooling over the years. In particular, datasets such as ASER can provide us with information about children who are in school or out of school, and who dropped out of school. This in turn can guide policy reforms to improve education systems that strive for equal opportunities for all. International assessments can be valuable to make comparisons across different countries based on a common international assessment framework. However, they must be used with caution while making high-stakes policy inferences.

Chapter 3

Multilevel Measurement Models

3.1 Measurement Invariance across Regions

International assessments (e.g., PISA, TIMSS) are often administered across multiple countries, various regions within a country, and in a variety of different languages. A fundamental question that one needs to think about is: "Are these assessments behaving similarly across these various populations?" In other words, are the test scores for students comparable across various sub-populations on the same construct of interest?

While employing scaled scores of a latent construct it is essential to make sure that the construct of interest or indicators across the various groups are compared on a common measurement scale. In particular, while making inferences pertaining to the differences in scaled scores we would like the differences between various groups (e.g., between countries, within multiple regions in a country, or across gender) to be trustworthy and not an artifact of underlying translation errors, or cultural differences in understanding the items or underlying construct of interest. Measurement invariance allows us to make sure that the various groups in the study perceive the items in a similar manner and the underlying construct is comparable across the sub-populations, thereby reducing the chances of biases and unfairness. When the estimates of the variables of interest across the sub-groups are used for high stakes education policy decisions we must ensure that all populations and sub-populations are treated fairly.

In international assessments such as PISA, the items used to construct the latent variable of interest often undergoes many rounds of discussions and testing. While assessing measurement invariance or lack of invariance it is vital to be clear about the latent variable of interest and the items that are measuring this construct (Millsap, 2012, p.47). Furthermore, the presence or absence of measurement bias (or DIF) in the items is dependent on the conceptualization of the

latent variable, and the violation of measurement invariance can be considered as measurement bias (or DIF) (Millsap, 2011, p.47). Multiple studies in the past have focussed on the examination of measurement invariance and DIF of items across achievement scale scores in large-scale assessments. However, few studies (e.g., Braeken & Blömeke, 2016; Rutkowski & Svetina, 2014) have investigated invariance or equivalency for non-achievement scale scores across subgroups using the items in the background questionnaires in PISA tests.

One of the primary criticisms of PISA tests is that they fail to account for the differences across countries in terms of their living conditions and culture, which warrants the need for measurement invariance checks among the different groups of students. For example, as pointed by Pokropek, Borgonovi, & McCormick, (2017) having a car in Japan does not have the same impact on socioeconomic status as in the United States. In the context of India, one example is the possession of a dishwasher. It is uncommon to find a dishwasher even in wealthy homes in India. Hence, while creating scale indices, not all the items can be comparable or given the same importance across different countries. A similar issue can be encountered with student responses to items on a questionnaire. Students from different countries may interpret the items differently, thereby posing concerns about the indices created from student questionnaires. Therefore, comparisons among different countries, or within a particular country, such as within rural and urban regions of a country, and across cultures is warranted.

Assessing measurement invariance is particularly essential in the Indian context because of India's poor performance in PISA, which was primarily attributed to the construction of the test. The Indian government blamed the socio-cultural differences between Indian students and those for whom the PISA tests were originally designed, and felt this made the test items unfair towards Indian students. Measurement invariance (MI) checks will help in understanding these

underlying differences if any. Furthermore, in India, we see a wide range of diversity among the rural and urban regions. In addition to this rural-urban divide, there are a vast number of languages spoken in India, since every state has its own spoken and/or written language. This cultural diversity could pose concerns for large-scale testing across the nation. The majority of students attend government schools in rural areas, often English or Hindi, which is a widely spoken language across India may not be the preferred language or language for the test. Therefore, when the same assessment is administered across multiple states, it becomes crucial to ensure that the meaning of the items are interpreted in the same manner across various sub-groups of students and there is no bias in the items across these sub-groups.

In addition, by assessing invariance among the item responses across the different languages we want to make sure that items do not favor a particular group. PISA follows a rigorous procedure for translation of tests (OECD, 2012), however, these procedures do not guarantee the absence of biases in the test, and it is vital to be able to capture the inherent bias that might be present in tests.

3.1.1 An IRT Approach to Assess Measurement Invariance

One of the most popular approaches for assessing measurement invariance across various sub-groups has been the factor analytic approach. Such an approach, typically, involves conducting nested tests that range from being least to most restrictive, which can be briefly described as follows. The first test is the configural invariance check across the sub-groups, which indicates that the items constituting the underlying latent variable and the factor structure are similar across the different groups in the study. Next is the metric invariance (or weak factorial invariance) (Meredith, 1993), which is established when the strength of the relationship between the items and the underlying latent factor(s) are equivalent across the same populations. The third

test is the scalar invariance, which is the most restrictive test, suggesting that for a particular score on the latent variable, people in the two sub-groups cannot have different intercepts on the observed variables.

Even though this approach is often used by researchers across a variety of applications (e.g., to assess invariance across gender on survey instruments), assessing measurement invariance across various sub-groups (e.g., across different countries) has not been the first step in the analysis of large-scale assessments. The TALIS study is one of the few large-scale assessments that has employed such a factor analytic approach to investigate measurement invariance of non-cognitive scales (OECD, 2010). For example, to capture the cross-cultural validity for indices such as teachers' beliefs about teaching and teachers' teaching practices in TALIS, the configural and metric invariance was established but scalar invariance was not. Thus, for these measures, the means across countries are not directly comparable and the focus was primarily on the patterns of cross-cultural differences than specific country-by-country comparisons (OECD, 2010).

In this dissertation work, I will be making use of the IRT approach to assess measurement invariance, which is a model-based approach that allows us to capture the relationship between an examinee's ability level on a particular latent trait and the probability of a particular item response (Lord, 1980). Often while administering a test or survey the responses received from the respondents are the observed variables. The underlying latent trait or unobserved latent variables are influenced by the test scores or responses to the survey items and are measured by observing the behavior on relevant tasks or items (Embretson & Reise, 2000). In the IRT context, often item bias or DIF analysis is used to assess the presence or absence of invariance. It must be noted that although the terminology used to define these invariance checks varies for the factor analysis and

IRT approach, they essentially make use of the same principle by assessing the levels of strictness for the measurement scale and enable inferences when comparing various sub-groups (Millsap, 2012). An IRT approach makes use of a nonlinear monotonic function to model the examinees' item responses, that is the item response function (IRF) (Reise et al., 1993). Multiple software programs, such as MPlus, flexMIRT, and mirt package in R allows multiple-group analyses. Multiple group analysis can help assess measurement invariance across different populations by simultaneously estimating the key parameters, such as the item parameters, and values for the proficiency or ability levels, instead of conducting separate analyses for each population group.

Before the items are examined closely for item-level biases it is important to establish measurement invariance among the various populations. For this, as a first step, we fit an IRT model to the data, the model fit is assessed, and the item discrimination (α_k) and item difficulty or threshold (κ_k) parameters are expected to be equivalent across the different groups being examined (Reise et al., 1993). Secondly, using multiple groups analysis, the items are simultaneously calibrated in the two different populations, and the mean and standard deviation of a population of one group is estimated relative to the N (0, 1) distribution of another group (the reference population). For example, while assessing invariance in outcomes for females and males, we could set females to be the reference group. This calibration allows the items for the two sub-groups to be on the same scale. Next, we place an equality constraint across the groups and assess the model fit for this constrained model. Once measurement invariance is established, we take a closer look at the IRF's across the groups to make sure that the resulting theta estimates are on a common scale, and are not biased thereby exhibiting Differential Item Functioning (DIF).

3.1.2 Measurement Invariance for Multilevel Models

Hierarchical data structures are common in education research. While analyses with these data structures have been popular for a while, it is only recently that multilevel structures have been accounted for when assessing measurement invariance across various sub-groups. When encountered with nested data structures the traditional factor analysis approach may not properly account for the within-individuals and between-individual variances, therefore it is essential to make use of multilevel factor analysis. This is particularly important when the inferences are used for policy-related decisions, for example, Schweig, (2014) demonstrates the importance of correctly modeling between-classroom and between-school variables using exploratory multilevel factor analysis and the policy implications of these modeling choices. Assessing cross-level invariance helps to ensure that the factors at the within-level and between-level components are interpreted as a part of the same latent variable (Schweig, 2014). To this end, to judge the amount of variance explained at the between level, we calculate the intra-class correlation (ICC) for each of the items. The ICCs help us evaluate the need for multilevel factor analysis (Muthén, 1994). The ICC will tell us how much variance in the ability levels is due to the between-school differences in comparison to the total variance. A value of an ICC for a given item that is greater than zero and closer to one indicates that the between-school differences are significant.

More recently, a few studies (e.g., Fox & Glas, 2016; Muthén & Asparouhov, 2018; Verhagen & Fox, 2013) have examined the use of multilevel IRT models in a Bayesian framework to assess measurement invariance. While computationally intensive, a Bayesian approach implemented in MCMC can be advantageous in comparison to maximum likelihood approaches as it allows us to obtain the marginal posterior distributions of parameters of interest (e.g., item parameters). Such distributions can be a great tool for comparing multiple groups on various item

parameters. Wang et al., (2008) note we can plot the posterior distribution of the difficulty parameters for the relevant items across the various sub-groups – we can assess the difference between the sub-groups, particularly for items that show DIF. For example, in the current study, the amount of overlap between the posterior distributions for the rural regions in TN and HP will indicate the extent of differences in how the items of interest are behaving across the two groups of examinees.

3.2 Multilevel - IRT Models

Increasingly, surveys are being used in multiple educational settings, and these survey items are often on a Likert scale. Such items can be modeled using Item Response Theory (IRT) models, such as the Graded Response Model (GRM) or Graded Partial Credit Model (GPCM). Furthermore, to capture the hierarchical structure of students nested within schools, we make use of multilevel models. The multilevel analysis allows us to capture the differences across outcomes while controlling for background characteristics. In particular, students attending different schools may not have the same level of access to resources. Moreover, students within a particular school may not have the same level of access to resources and instructional practices. Hence, a multilevel model allows us to tease apart these differences into their within-school and between-school components. Most studies in the access to opportunities literature have primarily made use of descriptive analysis with few studies employing statistical modeling (e.g., Borman & Dowling, 2010), and studies have often made use of a variety of indices to measure inequalities in educational opportunities (e.g., Asadullah & Yalonetzky, 2012). In Asadullah & Yalonetzky, (2012) the authors make use of four indices – a Pearson-Cramer index, an Overlap index, and two versions of Reardon's index -- to measure inequality of educational opportunity across Indian states (p.1152).

A multilevel IRT model allows us to capture the relationship between observed and latent variables, where the IRT model relates the responses on survey items or test performances to the latent variables of interest (Fox, 2005). In particular, the relationship between the latent variables and the observed background characteristics for students or group level characteristics can be analyzed in one modeling framework while taking into account the errors of measurement in the survey item responses. In these models, student responses to a set of items are modeled using an IRT model at level-1, and the differences in the magnitude of the construct of interest within-schools and between-schools will be modeled in levels 2 and 3 as a function of student-level and school-level factors, respectively. In this section, I will discuss the statistical framework for multilevel IRT models, the strengths of these models, and the estimation of these models using MCMC.

3.2.1 Statistical Framework for Multilevel IRT models

Item response theory (IRT) models can handle both dichotomous variables as well as variables with more than two categories (i.e., polytomous variables), and allows us to characterize the interaction between respondents and items through probabilistic models conditional on underlying latent traits. In this study, I model the polytomous items using the graded response model (GRM; Samejima, 1969) as the level-one measurement model.

A graded response model can be advantageous in comparison to a Partial Credit Model (PCM) as described below. The underlying assumption in PCM is that the item loadings are equal to each other. That is, in our example, seven items are equally related to the underlying latent construct of exposure to reading strategies. Such, a simplified assumption may not hold well in practice. In reality, such a model may not be suitable for empirical investigations (e.g., Kreiner & Christensen, 2014; Rutkowski & Rutkowski, 2016). The assumptions made in PCM may not hold across

countries as the item loadings (slopes) may be different across countries; there might be differences within a country as well, in terms of how students are interpreting some of the items. However, in models such as the Graded Response Model (GRM) (as described in Equation 1), we see that the item loadings or slope parameter, α_k, will indicate how strongly the items differ in relation to the underlying latent variable of the construct. Therefore items such as, "the teacher explains beforehand what is expected of the students" and "the teacher gives students the chance to ask questions about the reading assignment" can have different slopes in GRM modeling settings. This impacts the measurement of the latent construct where some items likely provide more information regarding the construct than others. GRM is a generalization of the Two - Parameter Logistic (2PL) model and can be described as a "difference" or "indirect" model (Embretson & Reise, 2000). It models the probability of any given response category or higher, so for any given category it will be similar to the 2PL model. Such a model forces the categories' to be ordered which is not the case in PCM or GPCM.

I will discuss these models in the context of an example focusing on teachers' use of stimulation of reading engagement (STIMREAD) in their language lessons as the outcome variable. This measure is constructed using seven items (see Table 5.1), and each item has four response categories varying from "1 = never or hardly ever", "2 = in some lessons", "3 = in most lessons", to "4 = in all lessons". In the level-1 (within-student) model, we model via an IRT model the student's responses to a set of items pertaining to the students' perceived exposure to teachers' use of these reading strategies. At level-2 (within-school) we model the differences in student exposure to such strategies as a function of various student characteristics (e.g., family SES, student attitudes toward school); and in level-3 (a between-school) model, we model differences

across schools in the STIMREAD construct as a function of various school compositional characteristics (e.g., school-mean SES, school size).

The level-one measurement model below links the observed responses to the construct of interest, in this case, students' perceived exposure to teachers' stimulation of reading engagement.

Level 1: Measurement model

For a polytomous ordered data responses, in a GRM, the probability of a student i (i = 1, 2, …, n_j) in school j (j = 1, 2, …, J) with an underlying latent trait of θ_{ij} giving a response falling into category c (c = 1, 2, …, C_k) and above on item k is defined by $P(Y_{ijk} = c|\theta_{ij})$. The conditional cumulative probabilities for an item k with four ordered response categories (C_k = 4) and the student response c (c = 1, 2, 3) are as follows:

$$P(Y_{ijk} \geq 0 \,|\theta_{ij}) = 1$$

$$P(Y_{ijk} \geq 1 \,|\theta_{ij}) = \frac{1}{1 + \exp(\kappa_{kc} - \alpha_k\theta_{ij})}$$

$$\vdots$$

$$P(Y_{ijk} \geq C_k - 1 \,|\theta_{ij}) = \frac{1}{1 + \exp(\kappa_{kc-1} - \alpha_k\theta_{ij})} \quad (1)$$

where α_k is the item slope and $\kappa_k = (\kappa_{k1}, \ldots \kappa_{k(C_k-1)})$ is a vector of ordered category intercepts for an item k. The category response probability is defined as the differences between two adjacent cumulative probabilities as shown in equation 2.

$$P(Y_{ijk} = c|\theta_{ij}) = \frac{1}{1 + \exp(\kappa_{kc} - \alpha_k\theta_{ij})} - \frac{1}{1 + \exp(\kappa_{kc-1} - \alpha_k\theta_{ij})} \quad (2)$$

$$= P(Y_{ijk} \geq c|\theta_{ij}) - P(Y_{ijk} \geq c - 1|\theta_{ij})$$

The slope parameter describes the strength of the relation between item k and the latent variables θ_{ij}. Furthermore, the GRM model treats the items as $C_k - 1$ dichotomies. In our example, this can be viewed as 0 vs 1, 2, 3; 0, 1 vs 2, 3; 0, 1, 2 vs 3.

Level - 2 (within-school)

The level-2 (within-school) model allows us to model the within-school relationships between the θ_{ij}'s and various student-level predictors of interest. This model captures the nesting of the N students in the sample, in the set of J schools in the sample. Thus within a given school (j = 1, 2, …, *J*) we have a sample of n_j students (i = 1, 2, …, n_j). The Q predictors in the within-school model are denoted **X**:

$$\theta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{Qj}X_{Qij} + e_{ij} \quad (3)$$

This model helps us capture within-school relationships between the student-level predictors of interest (e.g., family SES, student attitudes toward school) and differences between students in their perceived amount of exposure to reading strategies, for example. The residuals are assumed normally distributed with variance σ^2 (i.e., $e_{ij} \sim N(0, \sigma^2)$); σ^2 captures the amount of variation that remains in the θ_{ij}'s within schools after accounting for the predictors in the level-2 model. The predictors in this level will be group-mean centered (centered around the school mean values for a given school), hence the intercept parameter β_{0j} will represent the mean perceived exposure to reading strategies for the students in school *j*.

Level - 3 (between-school)

In a Level-3 (between-school) model, we model differences across schools in the β_{0j}'s as a function of differences in various school-level characteristics (e.g., school-mean SES):

$$\beta_{oj} = \gamma_{00} + \gamma_{01}W_{1j} + \cdots + \gamma_{0S}W_{Sj} + u_{oj} \quad (4)$$

$$\beta_{1j} = \gamma_{10}$$

$$\vdots$$

$$\beta_{Qj} = \gamma_{Q0}$$

where u_{oj} is a random effect assumed normally distributed with mean 0 and variance τ_{00}. Thus τ_{00} captures the variance in the β_{0j}'s that remains after taking into account the school-level predictors in the model. Note that we can also treat various regression coefficients in the within-school model as varying across schools, or we can treat them as fixed.

3.2.2 Strengths of Multilevel IRT Models

In addition to helping us investigate how various student-level and school-level characteristics relate to differences in their perceived exposure to potentially important instructional practices, it is also of interest to obtain sound estimates of the β_{0j}'s. These estimates can help identify schools in which students' perceived exposure to STIMREAD, for example, tends to be unusually high or low compared to other schools in the sample. In this regard, it would be valuable to plot such estimates versus various school-level compositional characteristics (e.g., school-mean SES). It would also be valuable to look more closely at such schools using other available data in the PISA survey or other sources of data.

A key feature of multilevel models is that they can provide us with potentially more precise group-specific estimates of outcomes of interest (e.g., the β_{0j}'s). This is especially valuable in cases where the number of students in a given school is small. Furthermore, we can obtain more precise estimates of the θ_{ij}'s for the students in a given school, which we will discuss further below. This can be valuable when our measures of students' perceptions of teacher practices are based on a relatively small number of items.

To help illustrate this, consider first an example based on the High School Beyond (HSB) data, in which we have approximately 7,000 high school seniors nested within 160 schools and the outcome of interest is a measure of student's 12th-grade math achievement. The level-1 (within-school) model is as follows:

$$Y_{ij} = \beta_{0j} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2) \tag{5}$$

where Y_{ij} is the observed math achievement score for student *i* in school *j*, β_{0j} represents the true mean achievement score for school *j*, and the e_{ij} are residuals assumed normally distributed with mean zero and variance σ^2; σ^2 represents the variance in observed student achievement scores (Y_{ij}'s) within schools. (Note this is in contrast to the three-level models described above where σ^2 represents the within-school variance in the θ_{ij}'s – the parameters capturing students' perceived exposure to instructional practices of interest.)

In the level-2 (between-school) model, the true mean achievement scores are viewed as varying around the grand mean for the population of schools:

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00}) \tag{6}$$

where γ_{00} represents the grand mean and u_{oj} is a level-2 residual or random effect capturing the deviation of the true mean achievement score for school *j* from the grand mean. The random effects are assumed normally distributed with a mean of zero and variance τ_{00}.

The mean of the outcome scores for the students based on the sample of n_j students nested in school *j* (i.e., $\overline{Y_j}$) provides us with an estimate of the mean achievement score for school *j*, with an error variance $V_j = \sigma^2/n_j$. This is an estimate based strictly on the data for school *j*. In addition, the estimate of the population mean based on all of the schools in the sample (i.e., γ_{00}) might be viewed as supplying potentially helpful information concerning the magnitude of β_{oj},

especially if n_j is small, in which case the error variance connected with $\overline{Y_{.J}}$ would be relatively large. The empirical Bayes (EB) estimate of β_{oj} (i.e., β_{0j}^*) would be a compromise between (or composite estimate of) $\overline{Y_{.J}}$ and γ_{00} :

$$\beta_{0j}^* = \lambda_j \overline{Y_{.J}} + (1 - \lambda_j)\hat{\gamma}_{00} \quad (7)$$

where $\lambda_j = \frac{\tau_{00}}{\tau_{00} + V_j}$. Note that when V_j is very small relative to τ_{00}, λ_j will be close to a value of 1, and nearly all of the weight will be placed on $\overline{Y_{.J}}$ and very little weight will be placed on $\hat{\gamma}_{00}$. But, on the other hand when V_j is very large relative to τ_{00}, λ_j will be close to a value of 0, and nearly all of the weight will be placed on $\hat{\gamma}_{00}$. Also, the error variance of β_{0j}^* , $var(\beta_{0j}^*) = \lambda_j * V_j$. When V_j is very small relative to τ_{00} and most of the weight is placed on $\overline{Y_{.J}}$, the vari\ance of β_{0j}^* is approximately equal to V_j. But, when V_j is large relative to τ_{00} and a lot of information is borrowed from the estimate of the grand mean and as a result of this additional information the error variance of β_{0j}^* will be substantially smaller than V_j. The basic idea is that we can attempt to "borrow strength" in estimating β_{oj} via the shrinkage or composite estimator for β_{oj}. As V_j increases in magnitude relative to τ_{00}, the amount of weight placed on $\hat{\gamma}_{00}$ increases.

However, often overlooked in discussions of shrinkage estimators is the need to borrow strength from information based on *other similar groups* (e.g., other schools that are similar to school *j* in key ways). Rather than shrinking toward a grand mean, a more sensible strategy would be to shrink a given school toward a conditional mean based on various school characteristics (e.g., school mean SES, whether a school is a private school or public school) that are available. For example, when we add school mean SES, we have

$$\beta_{0j} = \gamma_{00} + \gamma_{01} SchMn\,SES + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00}) \quad (8)$$

and the fitted value will be given by

$$\text{FV}(\hat{\beta}_{0j}) \;\; = \hat{\gamma}_{00} + \hat{\gamma}_{01} SchMnSES_j \tag{9}$$

where $SchMnSES_j$ is the school mean SES value for school j.

The EB estimate, in this case, will be a composite, of $\overline{Y_{.j}}$ and FV($\hat{\beta}_{0j}$)

$$\beta_{0j}^{*} = \lambda_j \; \overline{Y_{.j}} \;\; + (1 - \lambda_j)(\hat{\gamma}_{00} + \hat{\gamma}_{01}\text{SchMnSES}_\text{j}) \tag{10}$$

where, $\lambda_j = \frac{\tau_{00}}{\tau_{00} + (\frac{\sigma^2}{n_j})}$ and $\overline{Y_{.j}}$ is shrunk towards a predicted value based on school mean SES rather than towards the grand mean (see Raudenbush & Bryk, 2002, p.48).

This has important implications for estimating and drawing inferences concerning the parameters in the multilevel IRT models, for example, school-mean perceived exposure to STIMREAD for the schools in the sample, that is, β_{oj}. Estimating the multilevel IRT model using MCMC will yield a posterior distribution of the school-mean perceived exposure parameter for each school. The mean of each of these posterior distributions can be viewed as shrinkage or composite estimate that combines information concerning β_{oj} based solely on the information provided by the sample of students in school *j* and an expected value based on key predictors in the between-school model. If there are no predictors in the between-school model, we would end up shrinking toward a grand mean for school-mean perceived exposure, which as we saw in the HSB example can be problematic. For a given school, we must shrink toward an expected value based on key predictors of school-mean exposure values (e.g., key school compositional characteristics); we want to borrow strength from schools similar in important ways.

Also, we can obtain more precise estimates of student perceptions of exposure to instructional practices of interest. As can be seen in Equation 3 above, the θ_{ij}'s are modeled as a function of various student-level predictors. Thus the estimation of the three-level models described above using MCMC will yield a posterior distribution for each θ_{ij} that is a combination

of (1) an estimate of θ_{ij} for a given student based on the students' responses to the items that define the construct of interest, and (2) a predicted value based on Equation 3. If there are no predictors in Equation 3, the estimate of θ_{ij} for that person will be shrunk toward an estimate of the mean perception for the students in her school. When there are predictors in Equation 3 (e.g., a measure of a student's attitude toward school), the estimate of θ_{ij} will be shrunk toward an expected value based on the grand mean for the student's school (β_{0j}), plus a certain amount based on the magnitude of the coefficient for *student attitude toward school* times the student's value on that predictor (e.g., $\beta_{1j}X_{1ij}$).

Another point regarding shrinkage is that many software programs for IRT analysis often shrink estimates of latent variables (e.g., person abilities) toward a grand mean, thus causing problems analogous to those discussed in the context of the HSB example. Another example is the calibration of the student item parameters for the various scale indices in the questionnaires in PISA 2009. These are based on fitting the partial credit model to a sample of 15,500 students, which is 500 students are randomly selected from each of the 31 OECD countries (OECD, 2012, p.284). The international item parameters from the calibration step were used in a weighted maximum likelihood estimation (WLE) approach to obtain the individual student scores. These WLE estimates were then transformed using a linear transformation to an international metric with an OECD mean of zero and standard deviation one. Such a modeling framework and scaling approach adopted in PISA 2009, (1) ignores the dependence between the individuals nested within schools, and (2) shrinks the estimates to a common mean of zero, which can result in bias in the estimates and homogenization of the estimates. This is further motivation for working directly with the actual item responses and specifying and fitting multilevel IRT models.

3.3 Estimation of Multilevel IRT Models: A Bayesian Framework

An IRT model can be fit to a set of dichotomous and/ or polytomous test items. For each item, the IRT model describes an item response curve, and these response curves are multiplied together to result in the likelihood density function for ability. To estimate the ability scores we can make use of three approaches, Maximum Likelihood (ML), Expected A Posteriori (EAP), or Maximum a posteriori (MAP). The ML approach may not be ideal when we have perfect responses for a student. For example, when a student has all correct or incorrect responses, the ML estimate of θ for the student will tend toward positive or negative infinity. In some scenarios, we see that researchers may drop those individuals that have extreme scores. In such scenarios, the EAP or MAP approach or a fully Bayesian approach might be better suited. In this study, we make use of a fully Bayesian approach involving the use of MCMC for estimation. Few advantages of this approach are discussed below.

First, multilevel IRT models can become quite complex very quickly, especially with the addition of predictors in the level two and three models in the three-level models used in this dissertation. Bayesian techniques are flexible enough to handle these complex models and allow for useful presentations of results. With improved computational power and widely available software such as MPlus, STATA there is a growing interest to make use of Bayesian methods among researchers'. The shape of, and range of values spanned by, the marginal posterior distribution of a parameter of interest; 95% percent intervals based on the .025 and .975 quantiles of the posterior distribution; and the mean and median of the posterior distribution, can provide us with a more complete picture of the magnitude of the parameter of interest, and the amount of uncertainty of the magnitude of the parameter. For example, this dissertation will be looking at the distribution of learning opportunities (e.g., teachers' instructional practices). The marginal

posterior distribution for a given school can inform us about the distribution of teaching practices for that school.

Second, Markov chain Monte Carlo (MCMC) also opens up an array of useful possibilities, such as incorporating prior information based on results from other relevant studies. In Bayesian inference, both the observed data (*Y)* and the parameters (η) are considered to be random, and the joint distribution of these models (equation 11) is a function of the conditional probability distribution of the data given the parameters and the prior probability distribution of the parameters.

$$P\ (\eta\ |Y) \propto\ P\ (Y\ |\eta)\ P(\eta) \tag{11}$$

$P\ (\eta\ |Y)$ is referred to as the posterior distribution of the parameters (η) given the observed data (Y), and $P\ (\eta)$ is the prior distribution. Equation 11 can also be written as follows

$$P\ (\eta\ |Y) \propto\ L\ (\eta\ |Y)\ P(\eta) \tag{12}$$

where $L\ (\eta\ |Y)$ is the likelihood function based on the data. The actual sample data will be summarized by the likelihood function and the posterior summaries are a result of the updated information after the prior distribution is added. Our inferences for each of the parameters of interest are based on the marginal posterior distributions of the parameters of interest (e.g., $(\beta_{0j}|Y)$). To obtain the marginal posterior distribution of a particular parameter we integrate over all other unknowns in the model using MCMC, i.e., using MCMC we can simulate the marginal posterior distribution for each parameter in our model. Moreover, priors can be particularly useful as they allow us to update our knowledge of the parameters using "prior information". Particularly, in large-scale assessments, this property can be capitalized by incorporating our summaries from the posterior distribution from the previous year's assessment (e.g., PISA 2006) into our current study (e.g., PISA 2009). Such information is provided using the prior

distribution, $P(\eta)$. The addition of a real prior that provides substantial information can improve our inferences compared to using the likelihood alone (Gelman et.al, 2014, p.93).

Thirdly, in classical methods, Confidence Intervals (CI) are interpreted based on the notion of a repeated draw of samples from the population which contains a fixed and unknown mean μ. In the case of a 95% CI, when our sample size is sufficiently large, we are confident that our interval captures or contains the value of the parameter of interest. On the other hand, in a Bayesian approach, one forms a 95% *posterior interval for a parameter of interest* based on the 2.5% lower quantile of the posterior distribution and the 97.5% upper quantile. The interpretation would be that there is a 95 % probability that the true value of the parameter lies between the value of the 2.5% quantile and the value of the 97.5% quantile of the posterior distribution. A posterior distribution is essentially a probability distribution for a parameter of interest. The unknown parameters are estimated using Markov chain Monte Carlo (MCMC). In this approach, the model is fit to the data and allowed to run for a sufficiently large number of iterations, usually 5,000 or more, so that our MCMC algorithm converges or reaches equilibrium. After reaching equilibrium, we let the algorithm run for a large number of iterations (e.g., 20,000); the values of the parameters in the model generated in these subsequent iterations provide us with accurate approximations of the shape of the marginal posterior distribution for each parameter, and enables us to calculate 95% Bayesian intervals, posterior means and medians, and the like. As stated by Wang et al., (2008):

> "the MCMC methods essentially turns inference into simply adding, counting, and sorting. This is convenient in that it allows different users of the methods to choose whatever inferences they want once they have the posterior samples (p. 368)".

The estimates obtained via this fully Bayesian approach can be used to construct caterpillar plots. These plots allow us to plot the posterior mean and the 95% posterior interval of β_{oj} for each school – that is, we can plot the means of the marginal posterior distributions for the school-specific instructional practice estimates. Such a plot will allow us to see how different the schools are from one another in terms of the practices and what the spread for different schools looks like in each of the two states. For example, it will allow us to pinpoint the specific schools in which the perceived use of the practices of interest seem to be particularly low, moderate, or high or schools where there is a lot of variability.

Chapter 4

The distribution of school resources based on PISA's School questionnaire

RQ1. To what extent do school resources, such as lack of qualified teachers, and teacher absenteeism differ between schools within the rural and urban regions of HP and TN? What are the limitations and problems of working with indicators of such resources based on principals' responses? What critical information about the quality of students' educational experiences is missing, especially in the area of language instruction?

For this dissertation work, I make use of the publicly available PISA dataset. PISA was first administered to 15-year-olds in the year 2000 and is conducted every three years by OECD. The PISA 2009 study was conducted in 64 countries. An additional ten countries that were unable to participate in the PISA 2009 study were administered the test in 2010, and are known as the 2009+ countries or economies. Two states from India, namely, Himachal Pradesh (HP) and Tamil Nadu (TN) participated in the 2009+ study. India or any of these states did not participate in the PISA study ever since.

For each participating country, PISA ensures that the items are thoroughly reviewed, and often revised based on any feedback from personnel from individual countries. Due to the large cultural diversity among countries in the PISA study, before the administration of the items, OECD screens the items for any cultural issues in the different national contexts, and sometimes items are discarded if they do not meet the necessary standards (OECD (2012), p.36). International assessments require a high amount of scrutiny, as they consist of a large number of countries from around the globe with an extremely diverse population. The PISA data collection, sample, and measures are described in detail in the "PISA 2009 technical report" (OECD,2012) and "PISA 2009 Plus Results: Performance of 15-year-olds in reading, mathematics and science for 10

additional participants" (Walker, 2011). In this chapter, I first discuss the data and measures used to assess the (un) availability of school resources. Then, I discuss and present the descriptive analysis and various histograms to assess the extent to which teacher shortages, student absenteeism, and teacher absenteeism differ between schools within the rural and urban regions of HP and TN.

4.1 Data and Measures

PISA uses a two-stage stratified sampling strategy. This means that for each randomly selected school, a random sample of eligible students are selected, who must be enrolled in 7th grade or above, and fall between the ages of 15 years and 3 months to 16 years and 2 months at the time of the assessment. There was a strong recommendation that the school samples be selected by the PISA Consortium rather than the participating countries (OECD, 2012, p.71). For every school that participates in the PISA study, a school coordinator is appointed, who compiles a list of all 15-year-olds in the school and sends this list to the PISA National Centre in the country, which randomly selects 35 students to participate. To paraticipate in the study, students must have their parents' consent. Schools were randomly selected proportional to their size. Of the schools randomly selected in HP and in TN, over 90% of the schools in each of these states agreed to participate. The participation rates of the randomly selected schools in HP and TN were over 90%, i.e., 91% in the case of TN, and 94% in the case of HP (Walker, 2011, p. 103).

For the set of analyses in this chapter, I make use of the school questionnaire completed by the school principals in each of the states in the study – Tamil Nadu (TN) and Himachal Pradesh (HP). The PISA school questionnaire, which was completed by principals in the TN and HP sample of schools, consisted of items on a wide range of topics encompassing the structure and organization of the school, the teachers and the student body, the school's resources, and the

school's learning environment. The items on the school questionnaire were on a four-point Likert scale from "not at all" to "a lot". In terms of missing responses on the part of the principals, all of the principals of the schools in the HP sample completed the survey, and in the rural region of TN six principals did not complete the survey, and in TN's urban region two principals did not complete the survey. In all the number of principals in TN who did not respond to the survey is relatively small.

I focus on the following variables from the school questionnaire:

(1) School type (SCHTYPE), i.e., public or private school, which is an indicator variable coded 1 for public schools and 0 for private schools.

(2) The student-teacher ratio (STRATIO), which is obtained by dividing the school size by the total number of teachers. The number of part-time teachers is weighted by 0.5 and the number of full-time teachers is weighted by 1.0.

(3) The Index of Teacher Shortage (TCSHORT), which is derived from four items measuring the school principal's perceptions of the potential factors hindering instruction at school. This included (a) a lack of qualified science teachers, (b) a lack of qualified mathematics teachers, (c) a lack of qualified <test language> teachers, and (d) a lack of qualified teachers of other subjects. A Weighted Least squares Estimate (WLE) of TCSHORT is provided by PISA 2009. A higher WLE implies that fewer teachers are available at a school.

(4) Lastly, measures of student absenteeism and teacher absenteeism.

4.2 Analysis and Results

The goal of this descriptive analysis is to make use of the school principals' responses to the school questionnaire to examine how different/ similar the distribution of various school

resource factors (e.g., index of teacher shortage) are across the rural and urban regions of HP and TN.

To assess how different or similar the school resource factors discussed above are within the rural and urban regions of HP and TN, I conducted a series of detailed descriptive analyses. In Table 1, I first present descriptions of the samples of participating schools for HP and TN, i.e., the total number of schools, and the means and SDs for the number of participating students in these schools across the four regions – HP rural, HP urban, TN rural, and TN Urban. Recall that these results are for 10^{th} and 11^{th} graders and describes the sample of schools in the current study that were randomly selected and agreed to participate in the study. In the case of the HP rural schools, we see that on average there is a sample size of 16 students in a school, with a minimum of 3 students and a maximum of 35. In HP urban, there is on average a sample size of 17 students in a school, and the minimum and maximum sample sizes of students are 10 and 27, respectively. In TN's rural region a total of 76 schools participated, and on average there was a sample size of 18 students in a school, with a minimum sample size of two students and a maximum of 25. In TN's urban region a total of 51 schools participated and there was on average a student sample size of about 19 students, with a minimum and maximum sample size of 3 students and 25 students, respectively. We see that the mean numbers of students per school for HP-rural and HP-urban are fairly close, and the same applies to TN-rural and TN-urban.

Table 4.1: Total number of schools in each region (e.g., HP rural, HP urban), and the descriptives for the number of students in a school. The mean represents the average number of students in a particular school, who were randomly selected to participate in the study.

		Descriptives of the number of students in a school			
Region	No. of schools in each region (n)	Mean	SD	Min	Max
HP-rural	53	16.45	8.35	3	35
HP-urban	11	17.18	4.81	10	27
TN-rural	76	18.44	5.70	2	25
TN-urban	51	19.25	4.15	3	25

Next, in Table 4.2, we see a breakdown of the number of public and private schools in rural and urban regions across the two states, HP and TN. Overall, we see that there are large numbers of public schools, which are mostly located in rural areas. For HP, 46 of the 64 HP schools are rural public schools, and we find that there are relatively small numbers of rural private schools (6), urban public schools (6), and urban private schools (5). For TN, there are almost twice as many schools, i.e., 121 schools. While there are more rural public schools (i.e., 49) than in the other categories, there are relatively large numbers of rural private schools (23) and urban public schools (30). There is also an appreciable number of urban private schools (19). Figure 4.1 contains histograms based on principals' perceptions of the shortage of qualified teachers in the test language, and histograms based on their perceptions concerning shortages of qualified teachers. These are a few of the variables that form the composite variable, TCSHORT.

Table 4.2: The number of schools, and descriptive statistics of the number of students across public and private schools in the rural and urban regions across the two states, HP and TN

State	Regions	Type of school	Number of schools (n)	Descriptives of students in the schools			
				Mean	SD	min	max
HP	Rural	Public School	46	16.78	8.13	3	35
		Private	6	12.33	8.24	3	23
	Urban	Public School	6	18.3	2.8	15	23
		Private	5	15.8	6.61	10	27
TN	Rural	Public School	49	19.27	4.7	2	25
		Private	23	17.17	7.07	3	25
	Urban	Public School	30	19.57	3.77	8	25
		Private	19	18.74	4.9	3	25

I now take a closer look at these variables to see what the differences look like across the various regions. Recall that these questionnaires are completed by the school principals of the participating sampled schools, and the response category for these items was on a four-point scale, where 1=not at all, 2=very little, 3=to some extent, and 4= a lot. Responding to category 1 (not at all) implies that the principal thought there was no shortage of teachers. As we see in the top plots in Figure 4.1, for HP rural and HP urban nearly 80% of the principals responded that there was no shortage of teachers in the test language. Thus a large percentage of HP schools did not have shortages of teachers, according to the principals' responses. Note that the majority of the schools in HP's rural region are public schools, and for HP-urban there is a small sample of 11 schools. In the TN rural region we see that almost 50% of the principals responded "not at all", and 25% responded "very little". Nearly 12% indicated "a lot" of shortage of teachers in the test language. On the other hand, in TN urban we see that nearly 78% of the principals responded not having any shortage of teachers, and a small number (approx. 5%) responded having "a lot" of shortage of teachers.

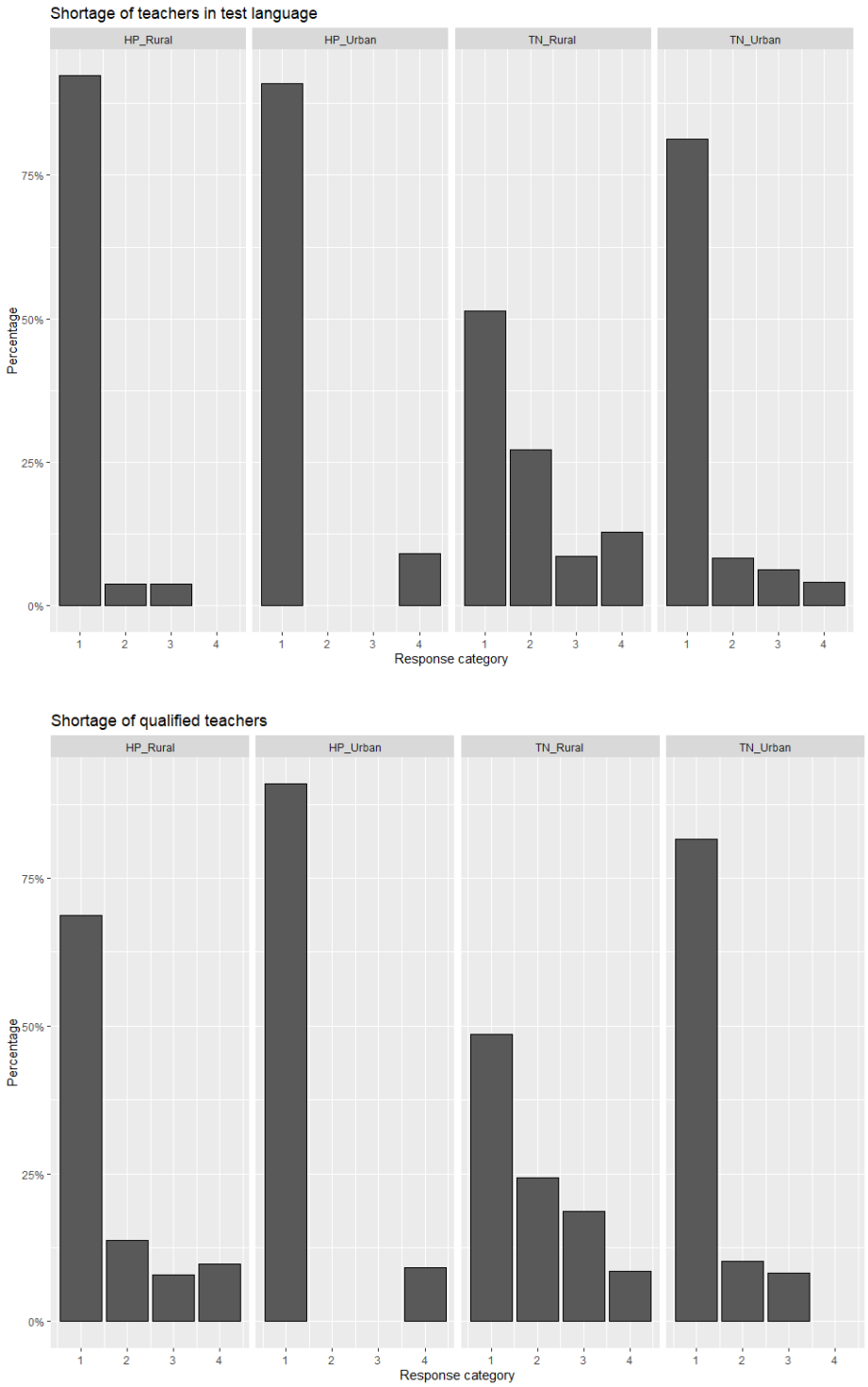


Figure 4.1: Shortage of teachers across rural and urban regions of HP and TN. The response categories for these items were on a four-point scale, where 1=not at all, 2=very little, 3=to some extent, and 4= a lot

We see similar patterns for the shortage of qualified teachers variable in the set of plots on the bottom of Figure 4.1. In TN rural 48% of principals responded that there was no shortage of qualified teachers and in TN urban more than 75% of principals reported having no shortage of qualified teachers. In comparison to TN urban, 9% of TN rural principals reported having "a lot" of shortage of qualified teachers. For both HP rural and HP urban, we see that nearly 10% of principals reported that there was "a lot" of shortage of qualified teachers. In HP rural, approximately 65% indicated that there was no shortage of qualified teachers at all.

Figure 4.2 shows the histograms for the teacher shortage (TCSHORT) scale for rural and urban regions in HP and TN. We see that for the HP and TN urban regions, and the HP rural region, a very large percentage of the scores on the TCSHORT scale for the schools in those regions are negative, implying that for large percentages of schools in those regions, there were no shortages of teachers according to the principals' responses. (Recall that a low score on the scale for TCSHORT implies that there is less shortage of teachers at a school.) In HP rural, we notice that there is a small percentage of schools that had some shortage of teachers – about four schools had a value above 1.0 on the TCSHORT scale indicating a shortage of teachers (see the bottom plot of Figure 4.1). For the rural region of TN (bottom left plot in Figure 4.2), some principals reported that there is a shortage of teachers, as can be seen from the higher positive scores in the TN rural regions histogram. This is also evident from Figure 4.1, where principals reported having "to some extent" and "a lot" of the shortage of teachers in test language as well as of qualified teachers.

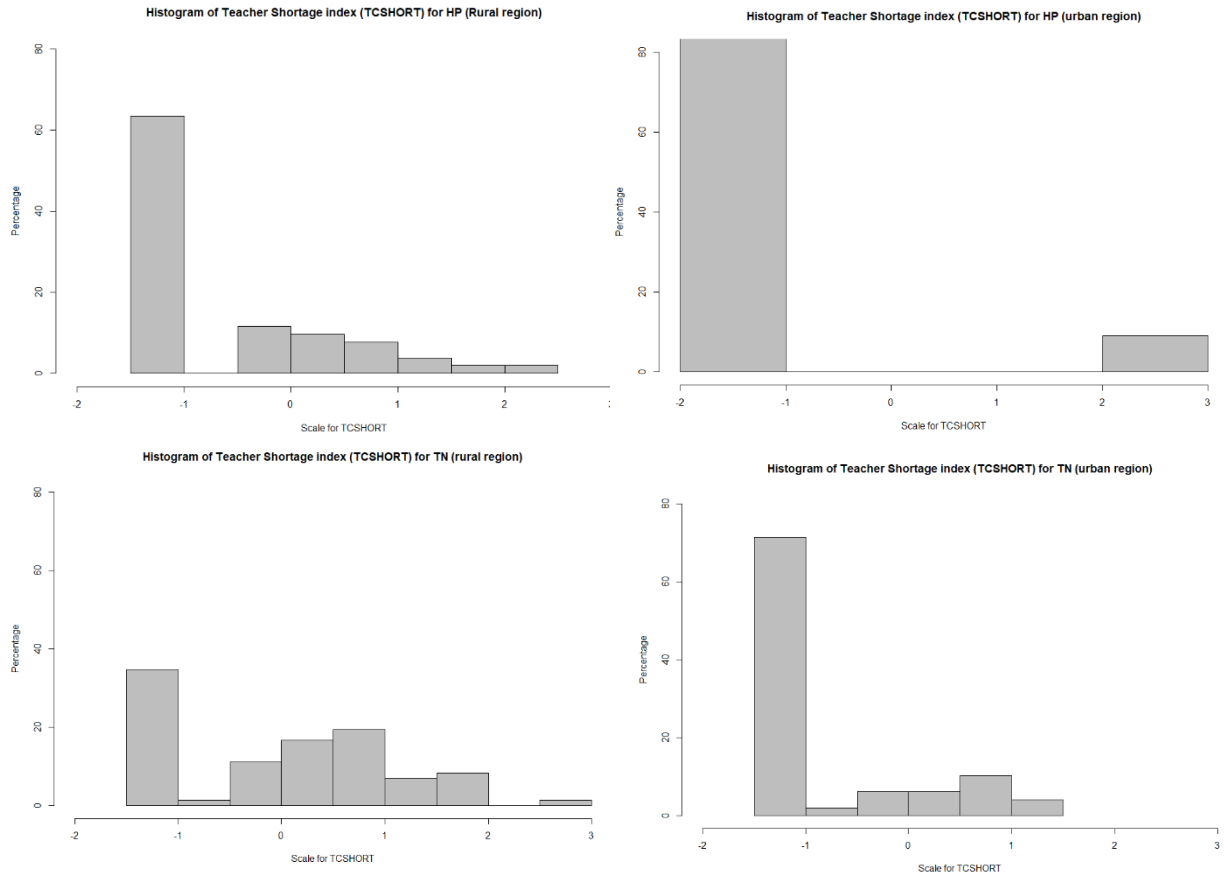


Figure 4.2: Percentage of teacher shortage (TCSHORT) in schools across rural and urban regions in HP and TN states. For the TCSHORT variable, a higher score implies that there are fewer teachers available at a particular school.

Figure 4.3 presents the percentages of student absenteeism and teacher absenteeism. Regarding student absenteeism, we notice that across most of the regions it tended to fall into the "very little" (category 2) or "only to some extent" (category 3) categories, with few principals agreeing that it was "not at all" a concern. For example, In HP's rural region 15% of the principals responded to category 1, that is student absenteeism was "not at all" a concern, while in HP's urban region we see that nearly 65% of the principals responded that student absenteeism was not a problem at all. In HP's rural region 30% of the principals stated that there was student absenteeism "to some extent", and 12% of the principals stated there was "a lot".

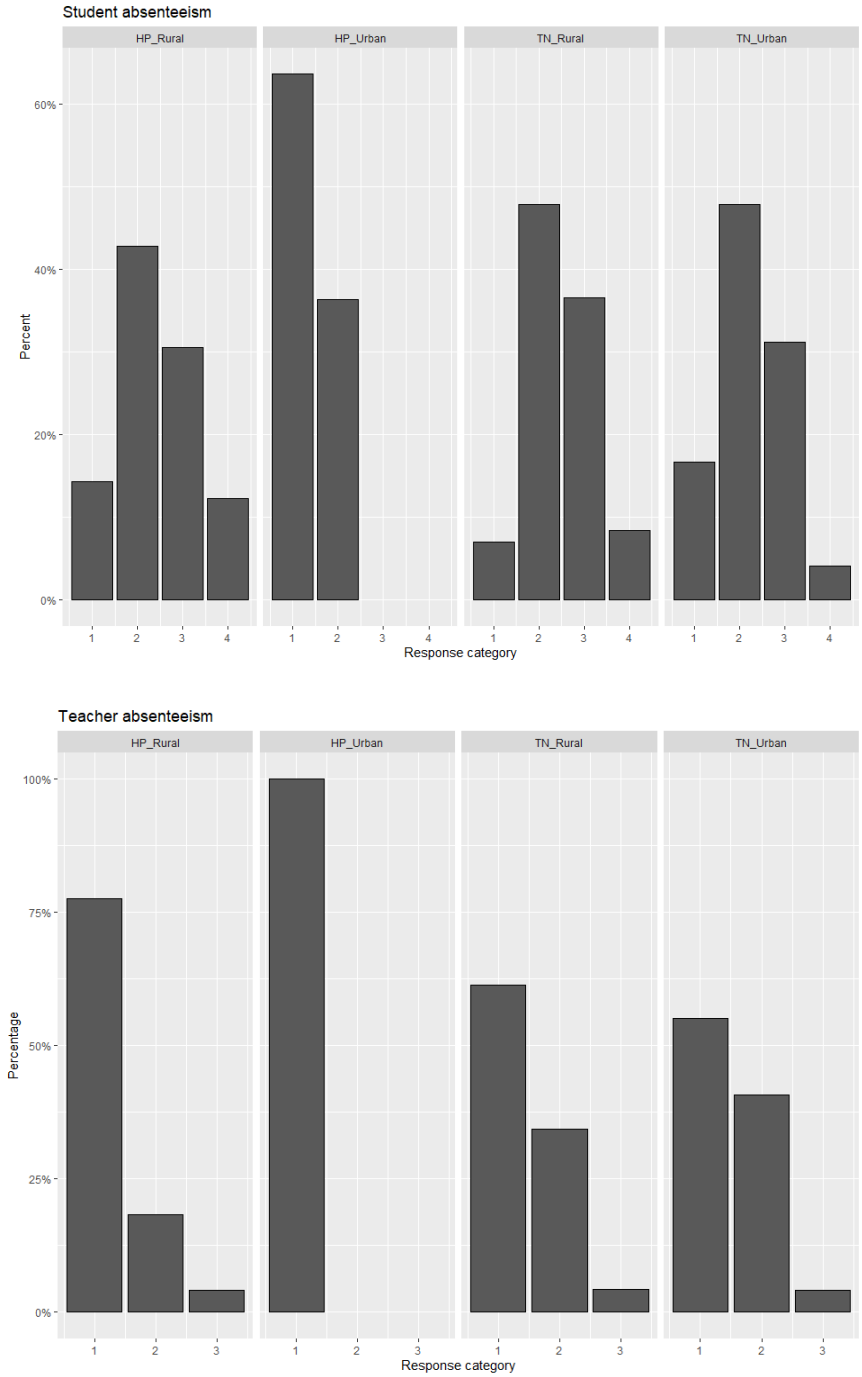


Figure 4.3: Percentage of student absenteeism (top plot) and teacher absenteeism (bottom plot) across rural and urban regions of HP and TN. The response categories for these items were on a four-point scale, where 1=not at all, 2=very little, 3=to some extent, and 4= a lot

In TN's rural and urban region 48% of the principals said that there was "very little" (category 2) student absenteeism.

Focusing our attention on teacher absenteeism, in the HP region, we see that most of the principals responded that it was not a concern -- nearly 75% of principals responded that teacher absenteeism was not a problem at all in HP's rural region. In HP Urban 100% of the principals responded that teacher absenteeism was "not at all" a concern. In TN rural and urban approximately 5% of principals responded that "to some extent (category 3)" teacher absenteeism was a concern.

4.3 Summary of Findings

This research question made use of the school questionnaire completed by school principals.

First, in the Himachal Pradesh (HP) sample, 80% of schools were located in the rural region and were public (or government-run) schools. In the Tamil Nadu (TN) sample, 68% of schools were public schools located in rural regions. An important finding from the descriptive analysis was that few principals reported that there were teacher shortages or teacher absenteeism concerns. The teacher shortage variable (TCSHORT) helps us get a good sense of teacher shortages across the two states— HP and TN. In HP, we see that across both rural and urban regions 60% of the schools reported having no shortage of teachers. In HP urban among the 11 schools in the study, only one school – a public school in a small town reported having a teacher shortage. In TN's urban region we see that nearly 70% of the school principals reported not having any teacher shortage. In TN's rural region less than 40% of the school principals reported having no teacher shortage, that is, 30 school principals reported some degree of teacher shortage concerns.

Secondly, across HP's rural region, and TN's rural and urban region most school principals responeded "very little" to "some extent" with respect to student absenteeism. Furthermore, in HP's rural region little over 10% of principals reported that they see "a lot" of

student absenteeism. Previous literature (Lewin, 2011) suggests that student absenteeism is a concern in the rural regions of India where low-income families often prioritize children going to work instead of their schooling. Related to teacher absenteeism, 75% of the school principals from HP's rural schools reported that teacher absenteeism was "not at all" a concern, and all of the principals of HP's urban region reported that teacher absenteeism was not a problem. In TN's rural and urban regions only 5% of principals responded to category 3 ("to some extent") indicating that most school principals' did not face any challenges with teacher absenteeism.

Previous research (e.g., Kingdon, 2007; Lewin, 2011) suggests high rates of teacher shortage and teacher absenteeism, especially across public schools in the rural regions of India, therefore the above results are slightly surprising. We do need to be cautious about the trustworthiness of the PISA data, since the responses are self-reported by school principals. This might be a cause for concern, because it is possible that the responses might be influenced by social desirability; for example, a number of principals may have wanted their schools to "look good", and therefore underreported various school problems. But it seems unlikely that such a large percentage of principals would be motivated to do so. It would be good, however, if there was another dataset that contains these variables for all of the HP and TN schools to corroborate the results.

Educational indicators regarding school resources are often based on principal responses to survey questions about shortages of qualified teachers, the extent to which teachers are absent, and the like. Such factors are likely related to some degree to student learning, but they are poor substitutes for directly measuring the extent to which students experience key instructional practices. Moreover, school principals may not be in a position to provide us with accurate information about the quality of classroom instruction and students' classroom experiences.

In the next chapter of this dissertation, I focus on the importance of assessing key instructional practices directly using students' responses to PISA survey items. In particular, I construct a latent variable measuring students' perceived exposure to key reading practices (i.e., STIMREAD), and using student latent scores as an outcome variable, show how using multilevel models we can investigate differences in exposure to STIMREAD practices within schools, and differences in exposure between schools, what student predictors are related to differences within schools, and what factors are related to differences between schools. Then in chapter 6, I show how this approach can be used to identify schools whose students on average experience relatively high exposure to STIMREAD practices, and schools whose students tend to experience relatively low exposure.

Chapter 5

Examining student's exposure to reading strategies across the rural and urban regions

RQ2a: Does measurement invariance hold for the items that form the teachers' stimulation of reading engagement (STIMREAD) construct across rural and urban regions?

RQ2b: What possibilities arise when we employ student STIMREAD values as outcomes in multilevel models in which students are nested within different schools? Are there appreciable differences in teachers' instructional practices across the rural and urban regions? To what extent does accounting for different student and school-level characteristics (e.g., student SES, teacher shortage at a school) explain the differences in the extent of these practices?

To examine RQ2b, I construct a latent variable measure of students' exposure to teachers' stimulation of reading engagement practices (STIMREAD), based on student responses to key items in the student questionnaire. In addition, I use various student-level and school-level predictor variables obtained through the student questionnaire and principal questionnaire, respectively, to investigage factors related to differences in student STIMREAD scores within-schools, and factors related differences across schools in their mean STIMREAD scores. Before examining RQ2b, I assess measurement invariance across the rural and urban regions of HP and TN for STIMREAD (RQ2a), and then this construct is used as the outcome variable in a series of analyses involving various student and school predictors variables.

In this chapter, I first describe the sample and participants used for these analyses. Next, I discuss the measures employed in the analyses and the handling of missing data. I then assess the measurement invariance for STIMREAD across the various HP and TN sub-groups, and then fit three-level models (e.g., item responses nested within students, who in turn are nested in

different schools) to key subgroups (i.e., the rural and urban regions of HP and TN), and finally present the results and a summary of the findings.

5.1 Sample and Participants

Before I discuss the specific sample size for this dissertation study, I discuss the sample size requirements set forth by the OECD for PISA data collection. The sample for PISA data includes students attending both educational institutions and vocational training programs, and typically, PISA requires a minimum sample size of 4,500 students from at least 150 schools in each country (OECD, 2012). The PISA assessment excludes few groups of children from its data collection, such as 15-year-old children who are not enrolled in schools, students who were intellectually disabled, students with functional disabilities, and students with limited proficiency in the language of the PISA assessment. Furthermore, students who were not provided with test materials in the language in which they were taught were excluded from the test.

For India, the PISA sample consists of students and schools from two states – Himachal Pradesh (HP) and Tamil Nadu (TN). PISA collected the data independently for these two states. Going forward for this dissertation study, the analysis is conducted separately for HP and TN, and due to the inherent geographic and cultural differences among the students from these states, it is sensible to avoid direct comparisons among these states.

The sample for this analysis was restricted to students who were enrolled in 10^{th} and 11^{th} grades, as these grades had the largest sample size. The total sample size consists of 3,445 students enrolled in 191 schools across the two states. The number of students within these schools varied from a minimum of two students in a school to a maximum of 35 students. In HP, the sample consists of 1,061 students nested in 64 schools, and in TN, the sample is comprised of 2,384 students nested in 127 schools. Furthermore, the data were disaggregated by rural and urban

regions for the two states. In HP, 53 schools (i.e., 82% of the HP schools) were public schools, and in TN, 75 schools (59% of TN schools) were public schools located in rural areas. The testing language for PISA included English, and the two native languages of the two states, i.e., Hindi and Tamil, which are the native languages of HP and TN, respectively. A total of 876 students took the PISA assessment in Hindi in HP, and 1842 students took the test in Tamil. 185 students took the test in English in HP and 541 students took the test in English in TN.

One concern reported by the PISA coordinators was that they were not provided with some of the data on the HP and TN student populations that were needed for the student sampling process. In particular, "it was established after the testing that these [states] sampled from student lists that were often incomplete: not all 15-year-olds within [a given] school were listed, [and as a result] it was not possible to determine whether any bias existed in the obtained sample" (see Walker (2011), p. 104). As a result, while HP and TN met the PISA standard for the sampling of schools, they did not meet PISA standards for student sampling within schools, thus caution must be exercised regarding the results based on the HP and TN samples of students. For example, if samples of participating students within the TN schools tended to be higher performing students – that is if high performing students were over-represented in a given sample -- this in turn might result in positively biased school-mean exposure scores to STIMREAD practices. This could also result in biased estimates of the amount of variation in student exposure scores within schools. Due to the incomplete lists of students in the sampled schools, it was not possible for PISA personnel to determine if any bias was operating in a given school's sample of participating students, and to what extent[2].

[2] Information provided by researchers familiar with the student sampling processes in HP and TN schools – for example, information concerning the kinds of students who may tend to be

In spite of these concerns, these data still provide a valuable opportunity to illustrate the methodological approach that is a key focus of this dissertation. An approach in which a latent variable construct (e.g., STIMREAD) is utilized to measure the perceived amounts of exposure students have to key instructional practices, and to investigate the amount of variation that we see in STIMREAD within-schools and between-schools, and to identify, for example, those schools among the low SES schools in a given region (e.g., TN rural), that have the highest or lowest exposure to key practices.

5.2 Measures

The PISA assessments consist of two parts. First, students complete a two-hour test on the cognitive items, which assesses their knowledge and skills in the various domains (e.g., reading, science, and mathematics). In the second session, students complete a background questionnaire and have about 30 minutes to complete it. Note that the development of the PISA tests and the choice of variables in the PISA framework are driven by theory and evidence using the literature on educational effectiveness and related research areas. Test development is a complex aspect of the PISA tests. A detailed description of the test framework, item development, and field trial are presented in the PISA technical report (OECD, 2009, p.32-40).

For the analyses that address RQ2b, I make use of the student questionnaire which consists of a variety of items pertaining to the student's family and home, their experiences in their test language classrooms, their attitudes towards reading, their perceptions of their test-language classroom's climate and their school climate, and their engagement and motivation. I make use of the teachers' stimulation of reading engagement (STIMREAD) construct as the outcome that

over-represented in the within-school samples – would be extremely valuable in interpreting results.

is constructed using items capturing students' perceptions of their test-language teachers' instructional practices. Table 5.1 presents the specific items for the construct of interest; the construct is comprised of seven items. All of these items have four response categories ranging from "never or hardly ever", "in some lessons", "in most lessons", and "all lessons". The psychometric properties of these items are documented in the PISA 2009 technical report. Since reading performance was the major focus of the PISA 2009 assessment, the responses to these items were collected with respect to the students' test language lessons. In India, there were three test languages – Hindi, English, and Tamil. Motivated by previous research, the following student-level variables are used as predictors in the analyses:

(1) A student-level scale index for economic, social, and cultural status (ESCS). This index is derived using three other indicators: The highest occupational status of parents (HISEI), the highest educational level of parents in years of education according to ISCED (PARED), and home possessions (HOMEPOS) (please refer to OECD, 2012, p.170 for more details). The values on the ESCS index have an OECD mean of zero and a standard deviation of one[3].

(2) Gender, female is coded as 1 and male as 0.

(3) Attitude towards school (ATSCHL) is constructed using four items (see Table 5.2 for specific items) on a four-point scale varying from "strongly disagree", "disagree", "agree", to "strongly agree". For the ATSCHL variable, a positive score implies that the student has a better attitude towards school.

[3] Using the international item parameters obtained from the calibration sample, weighted likelihood estimation (WLE) is used to obtain the individual student scores.These WLEs were transformed to an international metric with an OECD average of 0 and an OECD standard deviation of 1. Sample size for this analysis is noted in pg. 47 of this dissertation. For more details, please see OECD technical report (OECD, 2012, p.285).

(4) The teacher-student relations (STUDREL) variable is comprised of five items (see Table 5.2 for specific items) that pertain to the student's perceptions of whether their teacher is interested in their well-being, listens to what the student has to say, and treats them fairly. A positive score for STUDREL indicates a positive student-teacher relationship in the classroom.

Both variables ATSCHL and STUDREL have an OECD average of 0 and a standard deviation of 1. Variables such as STUDREL and ATSCHL help us better understand what the student-teacher relationship looks like and what students' perceptions are about their school. As will be seen later, this is particularly useful in understanding the within-school variation that we see between students regarding their responses concerning their teachers' instructional practices.

Table 5.1: Item descripotion for the STIMREAD construct of interest. Students respond to the overarching question that asks, "In your <test language lessons>, how often does the following occur?"

Construct of interest (or Scale Index)	Item description	Response Category
Teachers' stimulation of reading and teaching strategies (STIMREAD)	1. The teacher asks students to explain the meaning of a text 2.The teacher asks questions that challenge students to get a better understanding of a text 3. The teacher gives students enough time to think about their answers 4. The teacher recommends a book or author to read 5.The teacher encourages students to express their opinion about a text 6. The teacher helps students relate the stories they read to their lives 7. The teacher shows students how the information in texts builds on what they already know	1 = Never or hardly ever 2 = In some lessons 3 = In most lessons 4 = In all lessons

Lastly, two school-level predictors are included in the three-level model, (a) student-teacher ratio (STRATIO), and (b) the Index on teacher shortage (TCSHORT). Student-level variables

such as ESCS, STUDREL, and ATSCHL were aggregated to the school level and included in analyses of differences across schools in students' responses concerning their teachers' instructional practices. However, these aggregated predictors were found to have no relationship to the outcome and hence were not included in the analyses.

Table 5.2: Item descriptions and response categories for the student-level variables – attitude towards school (ATSCHL), and teacher-student relations (STUDREL) scales.

Scale	Items	Response category
Attitude towards school (ATSCHL)	Thinking about what you have learned in school: To what extent do you agree or disagree with the following statements? a. School has done little to prepare me for adult life when I leave school b. School has been a waste of time c. School helped give me the confidence to make decisions d. School has taught me things that could be useful in a job	1 = Strongly disagree 2 = Disagree 3 = Agree 4 = Strongly agree
Teacher-student relations (STUDREL)	How much do you disagree or agree with each of the following statements about teachers at your school? a. I get along well with most of my teachers b. Most of my teachers are interested in my well-being c. Most of my teachers really listen to what I have to say d. If I need extra help, I will receive it from my teachers e. Most of my teachers treat me fairly	1 = Strongly disagree 2 = Disagree 3 = Agree 4 = Strongly agree

5.3 Analysis

In this section, I first focus on issues of measurement invariance concerning the STIMREAD outcome variable. I then present a series of analyses conducted to examine if there

are appreciable differences in teachers' instructional practices across the rural and urban regions of HP and TN, and how the student and school predictors might be related to those differences.

5.3.1 Measurement Invariance using IRT Models

Before fitting a three-level model with an IRT model at level-1, a Graded Response Model (GRM) is fit to students' responses to the seven items that form the STIMREAD construct, and measurement invariance is assessed across the rural and urban regions of each state. Specifically, a GRM was fit to HP and TN separately, and then to each of the four different sub-groups – HP-rural, HP-urban, TN-rural, and TN-urban. For this analysis, the model was fit using the mirt package (Chalmers, 2012) in R, and the resulting overall model fit and item fit statistics were assessed. It is important to assess the fit of the IRT model (GRM in this case) first to ensure that the measurement model is appropriate before building the three-level models to examine RQ2 (ii).

Next, using the multiple group function in the mirt package, I estimated the item parameters across the rural and urban regions of HP and TN. This model will be the common baseline measurement model required to assess measurement invariance. This set of analyses follows a similar approach as described in Reise et.al (2003) and Millsap (2012). For measurement invariance, we would expect the item parameters to be similar across the sub-groups (i.e., similar between HP urban and rural, and similar between TN urban and TN rural). For each item we have four response categories, therefore we obtain estimates for three threshold parameters (κ_k), and one slope parameter (a_k). Wald statistics were calculated to test the parameter hypotheses and fit of items across the groups. The Wald test compares multiple groups simultaneously using a contrast matrix (Woods, Cai & Wang, 2012). Also, since we are conducting multiple group comparisons, we need to control for type I error. At the α-level of 0.05,

the Benjamini – Hochburg procedure is employed (Thissen, Steinberg & Kuang, 2002). Lastly, an equality constraint is placed across the sub-groups, in this case the rural and urban regions of each state, to obtain a common set of item parameters. This common set of item parameters are used in the three-level model analyses for *RQ2b*. The model fit for the constrained model with the equality constraint and the baseline model are compared.

5.3.2. Three-level Multilevel IRT Model: A Fully Bayesian Approach

Once measurement invariance is established, using MCMC a three-level multilevel IRT model is fit separately to the HP and TN data using JAGS (Plummer, 2011) in R, and then to the rural and urban regions of both HP and TN. Next, for each region, (e.g., HP rural, HP urban) the full three-level model was specified with a GRM measurement model at level-1, and then with the student- and school-level predictors specified at levels 2 and 3, respectively (The JAGS code is presented in Appendix A). The item parameters in the three-level analyses in JAGS were set to the ML estimates of the item parameters obtained from fitting a GRM to the student responses to the items (see Appendix B, Table B.1 for the item parameter estimates). Such a setup helps us avoid some convergence issues that we may encounter when we are estimating the 40 or so item parameters in our measurement model, plus all of the other parameters in our three-level HMs (e.g., fixed effects and variance components) simultaneously.

Prior specification

In a Bayesian framework, priors are specified for all the unknowns in one's model. Priors on the ability parameters (i.e., the θ_{ij}'s) will be assumed normally distributed. In essence, the level-2 (within-school) model can be viewed as the prior for the θ_{ij}'s, and the level-3 (between-school) model can be viewed as the prior for the β_{oj}'s. I will make use of diffuse proper priors for the fixed effects and variance components in the multilevel models, which allow the data to

dominate our inferences (Gelman et al., 2014). A normal distribution with a mean equal to zero and a large variance is specified for the fixed effect, γ_{00} and a uniform prior (Unif (0.1, 10)) is placed on the level-2 and level-3 variance components, σ^2 and τ_{00}, respectively. The starting values for the study were obtained using the maximum likelihood estimates of key parameters obtained via the mixedmirt package in R.

Monitoring convergence

Once the model, including the priors, was specified, two chains of the Gibbs sampler were run for 10,000 iterations with a burn-in of 1000, and for efficiency, the parameter values generated every 5 iterations were retained. The convergence diagnostics and estimates of the parameters were saved for every parameter of interest including the item parameters, fixed effects, and the variance components. Convergence was examined using times series plots and autocorrelation plots of the sampled values, and using Gelman-Rubin diagnostic statistics ($\hat{R}$).

5.4 Results

5.4.1 Descriptive Statistics

In Table 5.3, I present the means and standard deviations of the student responses to the items measuring the two constructs for each of the following four regions – HP rural, HP urban, TN rural, and TN urban. (Recall that the response scale to items that define STIMREAD ranges from 1 to 4: "never or hardly ever", "in some lessons", "in most lessons", "all lessons".) In the case of HP, in Table 5.3, we see that for the STIMREAD construct, item 2 (*better understanding*) has the largest means across both rural and urban regions, with values of 3.16 and 3.19, respectively. Item 1 (*explain text*) has the lowest mean of 2.67 in the HP urban region and a value of 2.88 in the HP rural region. Similarly, in the case of TN, item 2 (*better understanding*) has the largest means, which takes on values of 2.99 and 3.0 in the rural and urban regions, respectively.

We see that most of the students responded in the higher categories (e.g., category 3 or 4; see Figure 5.1) to these items – that is, students perceived that their teachers providing them with opportunities to "ask questions that challenge them to get better understanding" or "encourages them to express their opinion about a text" in most or all lessons.

Table 5.3: Descriptive statistics for the specific items for the construct, teachers' stimulation of reading engagement (STIMREAD) for the rural and urban regions in HP and TN.

Item (Item description)	**HP Rural**			**HP Urban**			**TN Rural**			**TN Urban**		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Item1 (*Explain text)*	836	2.88	1.00	187	2.67	0.94	1375	2.62	0.98	969	2.69	0.98
Item2 (*Better understanding*)	830	3.16	0.87	183	3.19	0.83	1367	2.99	0.92	966	3.00	0.91
Item3 (*Time to think*)	825	3.14	0.96	183	3.08	0.96	1362	2.86	0.99	968	2.88	1.01
Item4 (*Recommend books*)	821	3.10	1.01	182	3.14	0.91	1357	2.75	1.01	960	2.76	1.01
Item5 (*Express opinion*)	817	3.09	0.98	179	3.16	0.86	1359	2.97	0.94	963	2.99	0.96
Item6 (*Relate to lives*)	828	2.98	0.97	180	3.00	0.91	1362	2.92	0.99	967	2.92	0.99
Item7 (*Build on knowledge*)	826	2.98	0.98	177	2.98	0.91	1369	3.04	0.96	963	2.98	0.98

Across the seven items in Table 5.4, we see that in HP rural, on average approximately 8% of the students answered "never or hardly ever"; approximately 22% answered "in some lessons"; approximately 28% answered "in most lessons"; and approximately 42% answered, "in all lessons". In HP urban, the pattern in the 4 response categories was roughly similar: approximately 6%, 22%, 34%, and 37%. For TN rural and TN urban, across the seven items, we see on average a small percentage of students in the "never or hardly" category – approximately 8.5 - 9% -- which is similar to the pattern in HP rural; for category 2 we see a jump to approximately 28-29%, which is a little higher than what we see in the HP rural/urban data, and

we see a similar percentage in category 3; and finally the percentages in category 4 are approximately 34%, which is slightly lower than what we see in the HP rural/urban data.

Table 5.4: Percentage of student responses in each of the four categories for all seven items in the rural and urban regions of HP and TN.

Items	**Percentage of student responses in each of the four categories**							
	HP rural				HP urban			
	Cat 1	Cat 2	Cat 3	Cat 4	Cat 1	Cat 2	Cat 3	Cat 4
Item 1	8.85	30.26	25.12	35.77	10.16	35.29	32.09	22.46
Item 2	3.98	19.16	33.49	43.37	3.28	16.94	37.71	42.08
Item 3	7.27	18.42	27.03	47.27	8.74	15.85	34.43	40.98
Item 4	8.53	21.07	21.80	48.60	5.50	18.13	32.97	43.41
Item 5	9.55	15.42	31.09	43.94	3.35	19.55	34.64	42.46
Item 6	8.33	23.07	31.04	37.56	6.11	22.78	36.11	35.00
Item 7	8.23	24.58	28.33	38.86	5.09	27.12	32.20	35.59
Average of the seven items	7.82	21.71	28.27	42.20	6.03	22.24	34.31	37.43
Items	**Percentage of student responses in each of the four categories**							
	TN rural				TN urban			
	Cat 1	Cat 2	Cat 3	Cat 4	Cat 1	Cat 2	Cat 3	Cat 4
Item 1	10.11	43.49	20.58	25.82	9.39	39.53	23.43	27.66
Item 2	6.22	24.29	33.65	35.85	5.90	24.02	34.47	35.61
Item 3	9.40	28.34	28.86	33.41	9.92	28.31	25.93	35.85
Item 4	11.57	32.35	25.94	30.14	11.04	32.29	26.15	30.52
Item 5	6.84	25.09	32.67	35.39	7.58	24.09	30.43	37.90
Item 6	8.66	27.09	27.97	36.27	8.79	26.58	28.65	35.99
Item 7	7.23	22.43	29.51	40.83	9.14	21.81	31.26	37.80
Average of the seven items	8.58	29.01	28.45	33.96	8.82	28.09	28.61	34.48

In Table 5.4, we see that across all items for HP rural, category 4 (*all lessons*) had the highest percentage of responses. In HP urban, we see that for item 1, category 2 (*in some lessons*) and category 3 (*in most lessons*) had the highest percentage of student responses (see Figure 5.1). For other items, we see that categories 3 and 4 had higher student response rates in HP urban. In TN's rural and urban region for item 1 (*explain text*) most students responded to category 2 (see

Figure 5.1 top plot). For other items in TN categories, 3 and 4 had the highest percentage of student responses.

In Figure 5.1 I graphically depict a few of the items' percentage of responses. For item 1 in the case of TN, both in rural and urban regions, most responses were in category 2 ("in some lessons"). For HP, in the rural region, we see that the highest percentage of responses were in category 4 ("in all lessons") (i.e., approximately 35%), and for urban areas, the response percentages were approximately 35% and 32% in categories 2 and 3, respectively. For item 7, the distributions look fairly similar. In TN rural nearly 40% of the students responded to category 4 ("in all lessons"), and in HP rural a little less than 40% of students responded to category 4.

Turning our attention to the student-level predictors, we first look at the distribution of the index for economic, social, and cultural status (ESCS) for students. Recall that the values for the ESCS index have an OECD mean of zero and a standard deviation of one. Figure 5.2 presents the distribution of ESCS scores for the students in the rural and urban regions in HP and TN. In HP rural 90% of the students in the sample have values that are below zero, and in HP urban 76.5% of students in the sample have values that are below a value of zero. In TN rural 95.6% of the students are found to have values below a value of zero, and in TN urban 90.6% of the students are found to have values below a value of zero. The means and SD's for the ESCS scores are presented in Table 5.5. These numbers indicate that the students from all four regions tend to have low socioeconomic status relative to the OECD average (Walker, 2011).

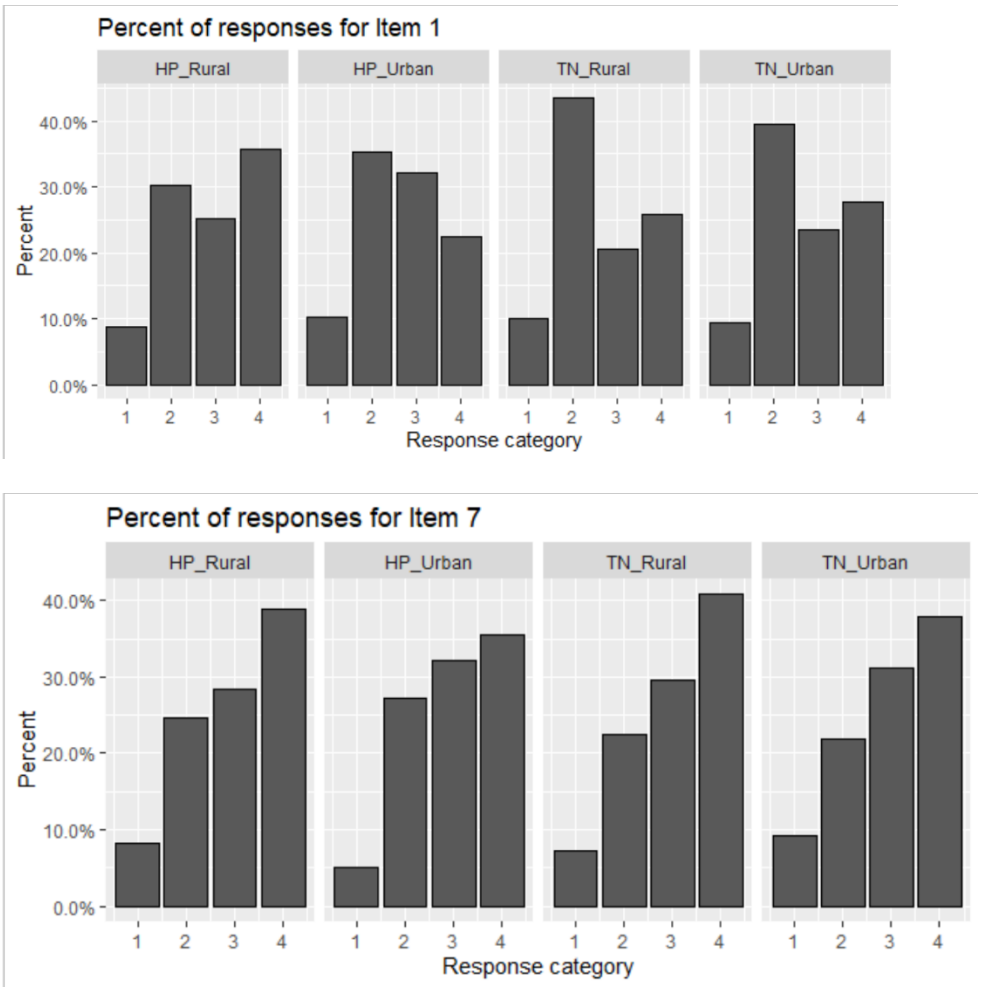


Figure 5.1: Percent of response cases in each category for items 1 and 7 for the STIMREAD construct across rural and urban regions of HP and TN.

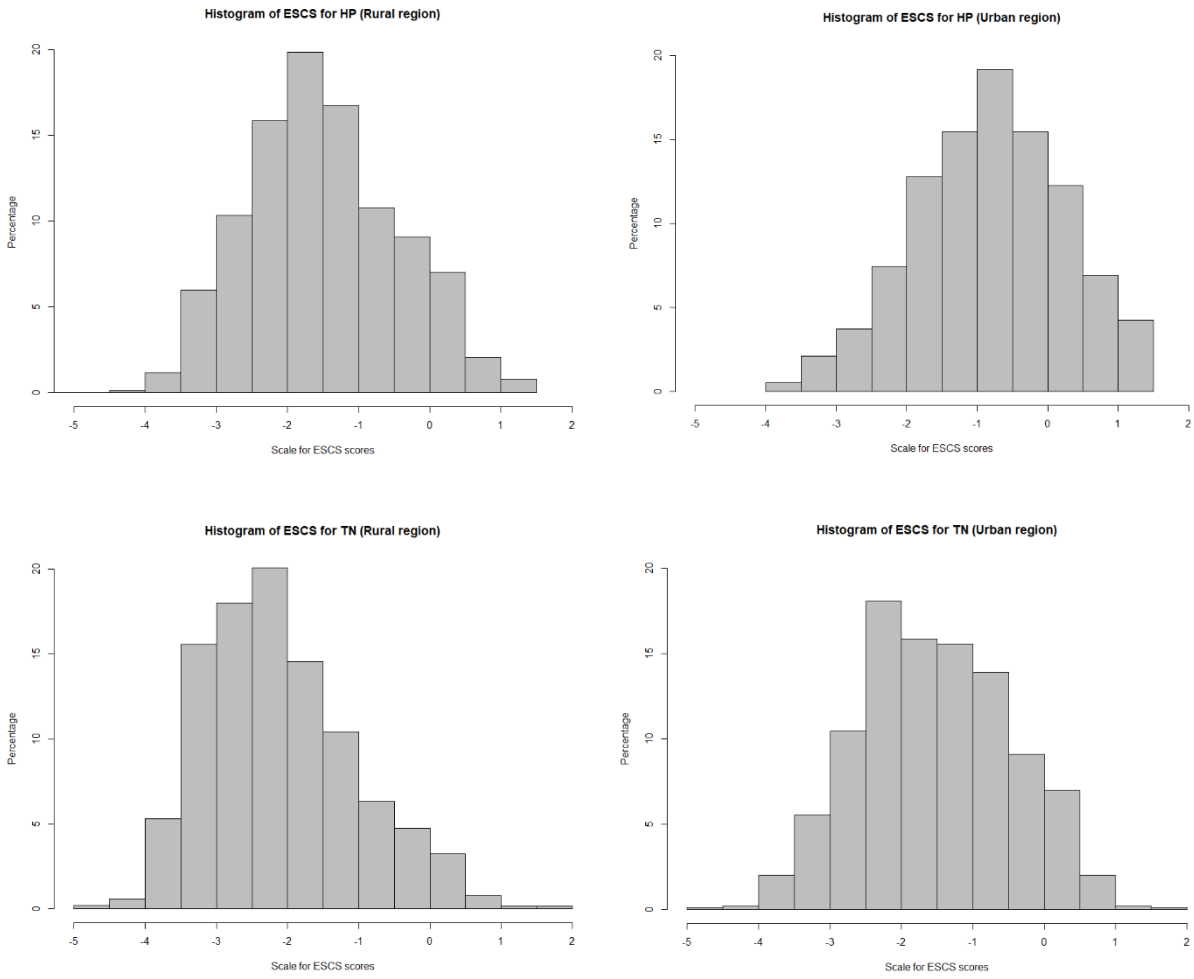


Figure 5.2: Histogram for the index for the economic, social, and cultural status (ESCS) scores for students in rural and urban regions of HP and TN. The ESCS is an OECD scale that has a grand mean of 0 and an SD of 1.

In Table 5.5 I have tabulated the total number of students who completed the student questionnaire, and the total number of schools that participated in the study across the four regions. HP Urban has the smallest sample of schools – 11 schools, with a total of 189 students. Next, Table 5.6 presents the descriptive statistics for the student and school-level predictors for HP and TN across both rural and urban regions. From Table 5.6 we see that 50% of the students were female, except in HP urban where 65% were female students. For the student-teacher relations (STUDREL) variable the minimum score was found to be -2.9 and the maximum was

2.45. A positive score for STUDREL indicates a student feels their teacher is interested in their well-being, listens to what the student has to say, and treats them fairly. (The OECD mean and SD for this variable are equal to 0 and 1, respectively, and the same applies to ATSCHL). The specific items for STUDREL and ATSCHL are tabulated in Table 5.2.

Table 5.5: Total number of students who completed the student questionnaires, and the total number of schools in the sample for the current study.

Region	Number of students who completed the survey	Number of schools in each region
HP-rural	872	53
HP-urban	189	11
TN-rural	1402	76
TN-urban	982	51

Across the four regions, it is seen that the means for STUDREL are positive, indicating that the students tend to perceive themselves as having a good relationship with their teachers. For TN we see that the mean for the rural region (0.47) is slightly higher than in the urban region (0.38). For HP the means for the students in the rural and urban regions are very similar (i.e., .59 and .62, respectively). The mean for HP urban is approximately a quarter of the standard deviation larger than the mean for TN urban. The variable student's attitude towards school (ATSCHL) has a minimum value of -2.99 and a maximum value of 2.45. A score above on this variable implies that the student has a more positive attitude towards school. Given the scale of the ATSCHL variable, we can say that for the ATSCHL variable there is about a third of an SD difference between the mean for HP rural (0.10) and that of the mean for TN rural (-.20). This is a meaningful difference indicating that students' perceptions toward their schools in rural and urban regions of TN were less positive in comparison to students from HP.

Table 5.6: Descriptive statistics for the student and school-level predictors across rural and urban regions of HP and TN

Variables	**HP Rural**			**HP Urban**			**TN Rural**			**TN Urban**		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Student-level *ESCS*	870	-1.51	1.05	188	-0.81	1.08	1395	-2.11	1.05	981	-1.52	1.06
Gender (female)	872	0.5	0.5	189	0.65	0.48	1402	0.49	0.5	982	0.54	0.5
Student-teacher relations (STUDREL)	852	0.59	0.92	184	0.62	0.95	1394	0.47	1.17	975	0.38	1.11
Attitude towards school (ATSCHL)	817	0.10	1.01	174	0.10	0.91	1331	-0.20	0.83	934	-0.17	0.82
School-level *Student-teacher Ratio (STRATIO)*	51	20.64	7.5	11	22.76	9.21	71	34.7	13.04	45	36.62	27.25
Teacher Shortage (TCSHORT)	52	-0.46	0.88	11	-0.69	1.11	72	0.06	0.98	49	-0.61	0.73

Next, for the school-level predictors recall that the student-teacher ratio (STRATIO) is obtained by dividing the school size by the total number of teachers. Schools in TN rural had an average of 99 students, and schools in the TN urban region had an average of 167. Some schools in TN urban are large with approximately 700 students. To that end, we see that the mean student-teacher ratio for TN's urban region is 36.6 with a standard deviation of 27.25. That is, there is one teacher for approximately 36 students in a classroom, which is much higher than HP's urban region where on average teachers have 22 students in a classroom. Next, for the TCSHORT variable, we see that across the three regions (except TN rural) the scores are on the lower end of the scale, which ranges from -2 to +3, indicating that teacher shortage was not a concern. This is also graphically depicted in Figure 4.2 in Chapter 4. Lastly, the correlations between the various student- and school-level variables were found to be small. For example, the correlation between STUDREL and ATSCHL was found to be 0.19, and the correlation between ESCS and gender was found to be 0.05.

5.4.2 RQ2a: Assessing Measurement Invariance using IRT Models Across Different Regions

In this section, I present the results from the measurement invariance analysis for the seven items that are used to construct the *teachers' stimulation of reading engagement* (STIMREAD) construct for the two states – HP and TN.

First, a graded response model was fit to the student responses to the items used to measure the construct. The estimates of the item parameters were obtained along with the model fit statistics. For each item, the slope parameter (α) indicates how well a particular item can distinguish between students with high and low latent scores. An item with high discrimination provides a relatively large amount of information about differences in the magnitude of a latent variable of interest across students. An item with low discrimination does not provide much information or is considered less relevant in differentiating across students. Estimates for the discrimination parameters for the items ranged from 1.02 to 1.61 on the STIMREAD construct.

The threshold parameters, κ_{kc} , indicate the latent trait score that is required to respond above the threshold κ_k with a 50% probability for item *k.* For example, in Table 5.7 for HP we see that the value of the first threshold (κ_1) for item 1 on construct 1 (STIMREAD) is -2.65, that is those with a latent trait score of -2.65 have a 50% chance of selecting category 1 or above for that item. The value for the second threshold was found to be -0.51 indicating that those with a latent trait score of -0.51 have a 50% chance of selecting category 2 or above. Those students with a latent score of a value of 0.77 for the third threshold have a 50% chance of selecting category 3 or above. A low threshold value (e.g., -2.65 for item 1 vs -2.06 for item 6 on STIMREAD) indicates that fewer people endorsed the first item response category for item 1 compared to item 6. Furthermore, we can obtain the latent scores for students via the Maximum

a Posteriori (MAP) approach. This approach is usually adopted when some students have the same responses to all items (e.g., all 1's or 4's).

Table 5.7: Item parameters (slope (α) and thresholds (κ)) for all items in the STIMREAD construct for Himachal Pradesh (HP) and Tamil Nadu (TN) states. A graded response model was fit to the items.

Items	HP				TN			
	α	κ_1	κ_2	κ_3	a	κ_1	κ_2	κ_3
1.Explain text	1.02	-2.65	-0.51	0.77	1.57	-2.70	-1.07	0.24
2.Better understanding	1.57	-2.70	-1.07	0.24	1.41	-2.45	-0.81	0.53
3.Time to think	1.38	-2.30	-1.04	0.13	1.28	-2.17	-0.51	0.63
4.Recommend books	1.42	-2.21	-0.87	0.07	1.05	-2.29	-0.33	0.92
5.Express opinion	1.39	-2.23	-1.11	0.22	1.50	-2.21	-0.70	0.49
6. Relate to lives	1.61	-2.06	-0.73	0.44	1.43	-2.11	-0.58	0.51
7.Build on knowledge	1.57	-2.11	-0.67	0.40	1.45	-2.15	-0.77	0.39

Next, a graded response model was fit to the seven items for the rural and urban groups of HP, and then to the seven items for the rural and urban groups for TN, and the item parameter estimates from these separate analyses were examined. Since the main focus of the analyses is to examine the differences between the rural and urban regions of HP, and the differences between the rural and urban regions of TN with respect to STIMREAD, the overall model fit statistics such as the M2, RMSEA, SRMR, and the Tucker Lewis Index (TLI) were examined in each of these analyses. RMSEA and SRMR were found to be less than 0.05 and TLI was found to be close to 0.95 for all four regions, which are within the recommended limits. Once the overall model fit was assessed, the multiple group function in mirt was employed, which performs a full-information maximum-likelihood multiple group analysis for both dichotomous and polytomous data using either the EM algorithm or the MHRM algorithm (Cai, 2010). The item parameter estimates were found to be similar for the rural and urban regions of HP, and in the analysis for TN, the item parameter estimates for the the rural and urban regions were found to be similar.

The adjusted p-values (using the B-H approach) for the Wald statistics were found to be above the 0.05 level for all items. This implies that the baseline model is functioning similarly across the rural and urban regions for HP for the various items, and in the analysis for TN, the baseline model is functioning similarly across the rural and urban regions.

Next, an equality constraint was placed across the slopes using the *invariance* argument in mirt allowing us to obtain a common set of item parameters for the rural and urban regions of HP, and a common set of item parameters for the rural and urban regions of TN (see Appendix B, Table B.1). This second model where the slopes are constrained is slightly more restrictive than the previous baseline model. Since these models are nested, the Akaike Information Criterion (AIC) statistics were used. The difference in AIC between the two models for the urban and rural regions in HP, and between the urban and rural regions in TN, was found to be more than ten points (the constrained model had a lower AIC value in both states). The likelihood ratio test also indicated a change in moving from the baseline to the constrained model, with a χ^2 value of 6.5. With regard to the model fit statistics, the M2 model fit statistics for the two models were found to be very similar. The SRMSR for the two models across the regions were very similar and found to be less than 0.05, and TLI and CFI were found to be approximately 0.97. It can be concluded that the two models are found to be similar in terms of practical fit indices. Moreover, for our substantive interest, the common set of item parameters for HP urban and HP rural, and the common set of item parameters for TN urban and TN rural, allows us to address the next research question (RQ2b) regarding the group differences in means and variances using the latent variable – STIMREAD construct. Lastly, this study did not examine DIF for the specific items, since the primary goal of the current study is to examine students' exposure to teachers' use of instructional practices.

To sum up, as seen from the above results we have established measurement invariance of the items across the rural and urban regions of HP, and across the rural and urban regions of TN, which allows us to compare the students' exposure to STIMREAD construct across rural and urban regions in HP, and across the rural and urban regions in TN. The item parameters are found to be equivalent between regions for each state and none of the items showed DIF.

5.4.3 RQ2b: Three-level Multilevel IRT Models: A Fully Bayesian Approach

Once measurement invariance has been established across the rural and urban regions of HP and across the rural and urban regions of TN, to compare the rural and urban regions of HP, three-level models were fit to the items of the STIMREAD construct using the HP data, and to compare the rural and urban regions of TN, three-level models were fit to the items of the STIMREAD construct using the TN data. Note that the measurement model at level-1 assumes that the construct is unidimensional, i.e., the $\theta_{ij}{}'s$ are capturing a single latent variable. The JAGS (Plummer, 2011) package in R was used to fit a three-level model to the rural and urban regions in HP, and to the rural and urban regions of TN. As noted previously, in this set of analyses rural and urban regions of HP and TN were analyzed separately. The goal of the analyses is not to compare the two states, but to compare the urban and rural regions within HP, and the urban and rural regions within TN.

In this section, I present the results from fitting (i) an unconditional model, and (ii) a model that includes student- and school-level predictors of the STIMREAD construct, to the data for the rural and urban regions of TN , and to the data for the rural and urban regions of HP. The three-level model fit to the data for this analysis makes use of a hybrid approach, where we treat the ML estimates for the item parameters (that is, the slope and threshold parameters) as fixed

values in the fully Bayesian analysis, and the fixed effects (e.g., the grand mean or regression coefficients of the predictors) and variance components are all estimated using MCMC.

Assessing convergence

Before drawing inferences regarding the various parameters of interest based on their marginal posterior distributions, the convergence of the models was assessed by examining the trace plots and autocorrelation plots.

In addition, Gelman-Rubin diagnostic statistics ($\hat{R}$) were used to assess convergence. This approach focuses on two or more chains of values for key parameters in one's model generated by one's MCMC algorithm; the chains are typically based on different starting values. The Gelman-Rubin diagnostics compare between-chain and within-chain variation. An $\hat{R}$ factor of 1 indicates good convergence, whereas a value larger than 1.1 could imply that there is still a notable difference between chains (for more details refer to Gelman et al., 2014, p.284; Gelman & Rubin, 1992). $\hat{R}$ values are available in JAGS output summary files along with the posterior means and SD for each parameter.

In the analyses presented below, the $\hat{R}$ values were monitored for the variance components and fixed effects. The values weres found to be around 1.00 in most cases. In Figure B.1 in Appendix B, I present the trace plots for the grand mean, and for the within-school and between-school variance components in HP's rural region. We see that the trace plot is evenly distributed, thereby telling us that there were no convergence issues. Once the model has converged, we save the simulation matrix, which is a matrix of values generated for all of the parameters (the columns of the matrix) in the model over a large number of iterations (the rows of the matrix). This matrix can be used to construct a plot of the marginal posterior distribution for each parameter of interest, and calculate various summaries for each posterior distribution, such as the posterior mean and

median, a 95% Bayesian interval, and the proportion of mass above values of substantive interest for key parameters.

Results for the unconditional model (Model 1) for rural and urban regions of HP and TN

The latent trait for this analysis is student's perceptions towards teachers' stimulation of reading engagement (STIMREAD). Since the items are completed by students, this latent variable can be labeled as "students' perception of the extent to which teachers use various techniques and practices for stimulating engagement in reading. First, an unconditional model (i.e., a model with no predictors) is fit to the data to obtain an estimate of the grand mean (γ_{00}), the within-school variance component (σ^2), and the between-school variance component (τ_{00}). The level-2 and level-3 model specifications are below (please see equation 13). (The level-1 model (i.e., the measurement model) is the same graded-response model (GRM) as depicted in Equation 2)

$$\theta_{ij} = \beta_{0j} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2) \tag{13}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00})$$

where, γ_{00} is the grand mean of students' perceptions of the extent to which teachers are using stimulation of reading engagement practices. The within-school variance component (σ^2) captures the variance in the outcome variable across the students within schools, and τ_{00} represents the between-school variance.

A three-level model as shown in Equation 13 is fit to the samples of students and schools in the rural and urban regions of HP and in the rural and urban regions of TN. Let's first take a look at Figures 5.3 and 5.4, which present the marginal posterior density plots for the grand mean, within-school variance component, and between-school variance component of the STIMREAD

construct for HP's rural and urban regions. The marginal posterior distribution for a parameter of interest is a probability distribution providing information about how probable it is that the true value of the parameter lies above or below a particular value of substantive interest, and how probable it is that it lies within a certain range of particular values. An important property of marginal posterior distributions is that they take into account uncertainty in all other unknowns in the model. The shape and spread of a marginal posterior distribution depend on whether the parameter of interest is, for example, a fixed effect or variance component. Likelihood functions and marginal posterior distributions of variance components are typically positively skewed since variances can not take on values smaller than 0.

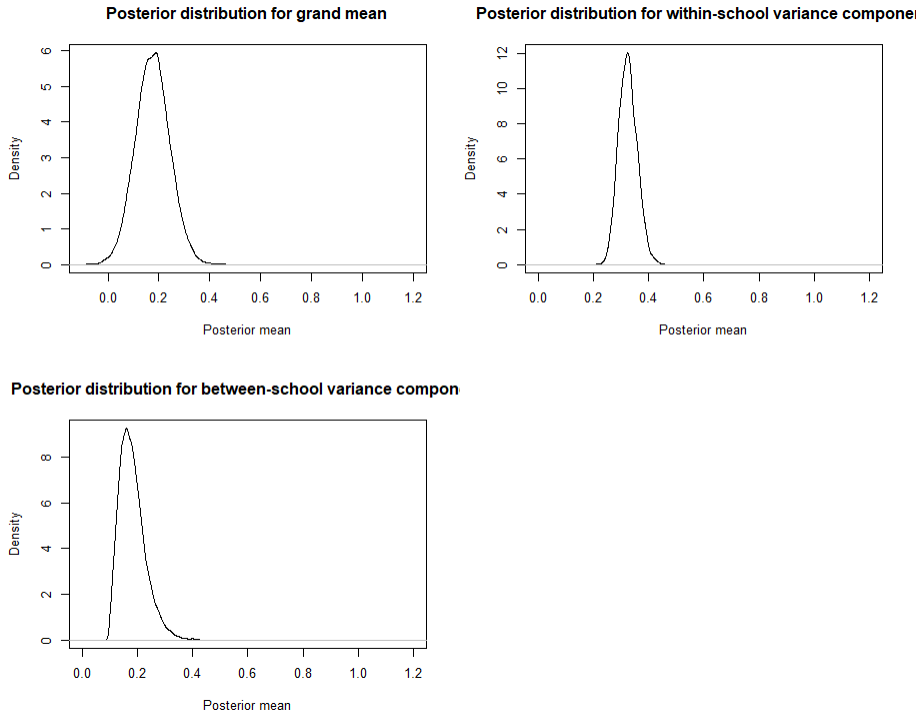


Figure 5.3: Posterior density plots for the grand mean, within-school variance, and between-school variance for HP rural region without any predictors for STIMREAD.

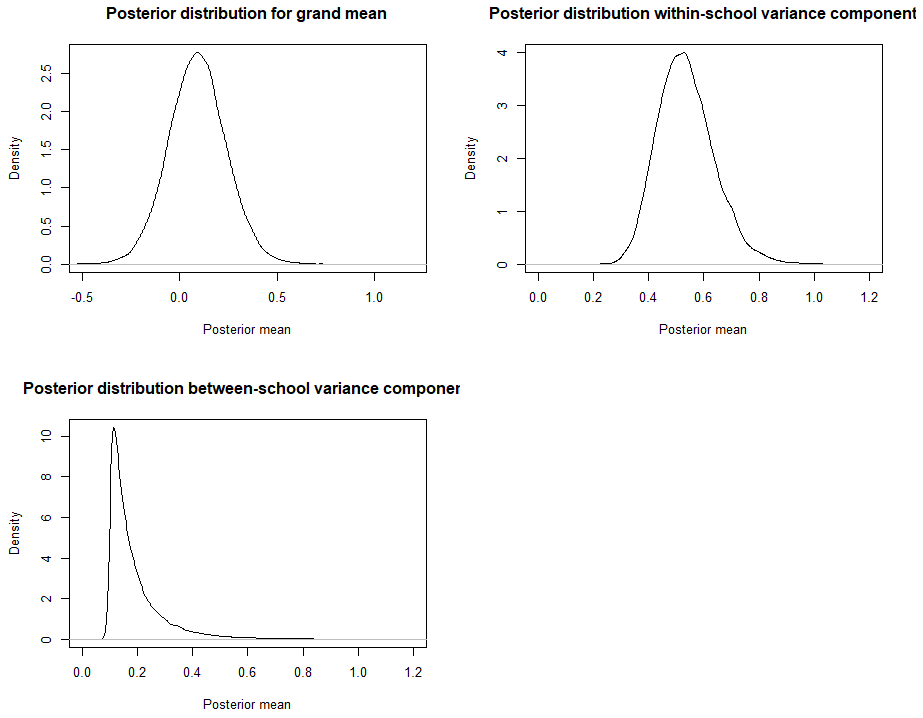


Figure 5.4: Posterior density plots for the grand mean, within-school variance, and between-school variance for HP urban region without any predictors for STIMREAD.

In the case of within-school variance components, as the number of students in one's sample increases, the marginal posterior becomes more tightly distributed around the mode of the distribution, and less skewed. In the case of the marginal posterior distributions of between-school variances, as the number of schools increases, the marginal posterior becomes less skewed and more narrow in shape. For fixed effects in models in which the distributional assumptions are normal at the student- and school levels, the marginal posterior distributions will tend to be symmetric (but maybe skewed if there is an outlying school, for example); the number of students in a sample will drive how wide or narrow the marginal posterior distribution of a coefficient for a student-level predictor will be, and the number of schools will drive how wide or narrow the marginal posterior distribution of a coefficient for a school-level predictor will be. For example, in Figures 5.3 and 5.4 we notice that the marginal posterior for the grand mean and the within-school variance component is much more symmetric in comparison to the between-school

variance component. Note, for example, there are 53 schools in the HP rural region and only 11 in HP urban.

Results for the grand mean for Model 1

In Tables 5.8 and 5.9, I present results from an unconditional model for the rural and urban regions of HP and TN, respectively. For HP rural, the posterior mean of the grand mean (γ_{00}) was found to be 0.17, and for HP urban, the posterior mean of the grand mean was found to 0.04. These are estimates of the overall means of the latent scores for students' exposure to STIMREAD in HP rural and HP urban. For HP rural the 95% interval was found to be 0.04 and 0.30, and for HP's urban region we see that the 95% interval ranged from -0.20 to 0.31. The lower boundary of the 95% credible interval (e.g., a value of 0.04 for HP rural) for a given marginal posterior corresponds to the lower 2.5th percentile of the marginal posterior, and the upper boundary corresponds to the 97.5th percentile. In the Bayesian framework, we can say that there is a 95% probability that the true value of the grand mean lies between 0.04 and 0.30 for HP rural. HP urban has a small sample size with 11 schools and a total of 189 students. This small sample size can result in wide 95% intervals. As can be seen, the 95% interval for the grand mean for HP urban is appreciably wider than the interval for the grand mean for HP rural. For the TN rural and urban regions, the posterior mean of the grand mean for both regions was found to be -0.1. For TN rural the 95% interval was found to be (-0.19, -0.016), and for TN's urban region we see that the 95% interval ranged from -0.21 to 0.00.

An advantage of the multilevel IRT model is that it links the polytomous responses to an item *k* of a student *i* in a particular school *j* to the overall mean γ_{00} , and the variance components (σ^2 and τ_{00}) (see Equations 2-4). In other words, the IRT model provides us with the thresholds κ_{kc} and the slope parameters α_k , and the level-2 and 3 models described above in Equation 13

capture the multilevel structure of the students nested in different schools. This multilevel model provides us with estimates for γ_{00}, σ^2, and τ_{00}.

To further understand the differences in the grand mean estimates across the rural and urban regions let's look at the relationship between the trait level and items responses. This can be described by the (i) expected probabilities of responding to a particular category on an item at a given theta estimate and (ii) expected score on an item at a particular trait level. Recall the GRM formulation from chapter 3 (p.41-42): The probability of a student selecting a score/category at or above each item score category (c) conditional on the trait level (θ_{ij}) is given by $P\left(Y_{ijk} \geq c|\theta_{ij}\right) = \frac{1}{1+\exp(\kappa_{kc} - \alpha_k\theta_{ij})}$. In IRT terminology these are referred to as the "operating characteristic curves". For a graded response item with four response categories, we have three probability curves which are further used to compute the probability of responding in each of the four categories as shown below,

$$P\left(Y_{ijk} = c|\theta_{ij}\right) = P\left(Y_{ijk} \geq c|\theta_{ij}\right) - P\left(Y_{ijk} \geq c-1|\theta_{ij}\right)$$

These probability curves are referred to as the "Category Response Curves (CRCs)", and they allow us to describe the probability of a student responding in a particular category conditional on their trait level. Therefore, to further understand the differences across the rural and urban regions we compute the expected probabilities of a student with a particular value for theta (e.g., a theta value equal to the grand mean of that student's region) responding in a particular category (e.g., lowest or highest categories) on a given item.

Table 5.8: Posterior means, SD's, and the 95% Bayesian intervals for the grand mean and the variance component parameters across HP's rural and urban regions for STIMREAD construct 1. The three-level model was fit separately to each region.

Parameter of Interest	HP-Rural	HP-Urban
	Posterior Mean (SD)	Posterior Mean (SD)
Fixed effects	0.17(0.06)	0.04 (0.13)
Grand mean (γ_{00})	[0.04, 0.30]	[-.20, 0.31]
Variance components		
Within-school variance (σ^2)	0.32 (0.03) [0.26, .39]	0.29(0.06) [0.18 ,0.43]
Between-school variance (τ_{00})	0.17(0.04) [.10, 0.27]	0.17 (.11) [.10, 0.41]

Note: Common IRT item parameters were used for rural and urban regions of HP.

Table 5.9: Posterior means, SD's, and the 95% Bayesian intervals for the grand mean and the variance component parameters across TN's rural and urban regions for STIMREAD construct

Parameter of Interest	TN-Rural	TN-Urban
	Posterior Mean (SD)	Posterior Mean (SD)
Fixed effects	-0.10 (0.044)	-0.10 (0.055)
Grand mean (γ_{00})	[-0.19, -0.016]	[-0.21, 0.00]
Variance components		
Within-school variance (σ^2)	0.25 (0.02) [0.21, 0.30]	0.28 (0.027) [0.23, 0.33]
Between-school variance (τ_{00})	0.11 (0.014) [0.10, 0.15]	0.12 (0.01) [0.10, 0.14]

For example, for HP across rural and urban regions item 6 (*relate to lives*) was found to have the steepest slope (1.94). Recall, that the steeper the slope, the narrower and peaked the CRCs are, which indicates that the response categories differentiate among trait levels fairly well. In HP's rural region, for a student with a latent score of 0.17 -- the grand mean for HP rural -- the probability of responding to category 4 was found to be 48%. Next, the expected probability of

responding to category 4 for a student who is in a school whose mean outcome score is 2 $SD_{between}$[4] above the grand mean (i.e., a latent score of 0.99) was found to be 82%, and the expected probability for a student who is in a school 2 $SD_{between}$ below the grand mean (a latent score of -0.65) was found to be 16%. In HP's urban region, for a student in a school whose, mean outcome score is equal to the grand mean (i.e., a latent score of 0.04), the probability of responding to category 4 was found to be 42%. The expected probability for a student who is in a school whose mean outcome score is 2 $SD_{between}$ above the grand mean in HP urban (i.e., a latent score of 0.86) was found to be 78%, and the expected probability for a student who is in a school whose mean is 2 $SD_{between}$ below the grand mean (a latent score of -0.78) was found to be 13%.

For TN across rural and urban regions for item 6 (*relate to lives*) the slope was found to be 1.46. In TN's rural and urban region, the grand mean was found to be -0.1; for a student with a latent score of -0.1 the expected probability of responding to category 4 on item 6 was found to be 34%. The posterior means for the between-school variance component for TN rural and TN urban was found to be 0.11 and 0.12, respectively. Next, for TN rural, for example, the expected probability of responding to category 4 on item 6 for a student who is in a school whose mean outcome score is 2 $SD_{between}$ bove the grand mean (i.e., a latent score of 0.56) was found to be 57%, and the expected probability for a student who is in a school 2 $SD_{between}$ below the grand mean (a latent score of -0.76) is 16%. In TN's urban region the values only change slightly; The the expected probability for a student who is in a school whose mean outcome score is 2 $SD_{between}$ above the grand mean in TN urban (i.e., a latent score of 0.59) was found to be 58%, and the

[4] $SD_{between}$ represents the standard deviation of the between-school variance component for a region. It is the square root of the posterior mean of the between-school variance component. Therefore, for HP rural, the grand mean +/- 2*($SD_{between}$) is found to be 0.99 and -0.65, respectively. This provides information about the extent to which HP rural schools vary in their school-mean STIMREAD values

expected probability for a student who is in a school whose mean is 2 $SD_{between}$ below the grand mean (a latent score of -0.79) is 15%.

Another approach to describing the relationship between the item responses and the student's trait level is to graph the expected score on a scale of 1 to 4 versusa range of values for the theta estimates, i.e., latent trait values. Figure 7 shows how the item responses change as a function of the trait level. Since each of the seven items has different slopes, an average expected score was obtained across the individual expected item scores for the seven items. This is of substantive use to us since the value of each of the categories has a particular meaning (i.e., 1 = "never or hardly ever", 2 = "in some lessons", and so on). In Figure 7 we see that students with a latent trait value around zero have an expected score of around 2; for students with a latent trait value of 2, the expected score is slightly below a value of 4, and for students with a latent trait value of -2, the expected score is approximately equal to an expected score of 1

To sum up, we see that the expected probabilities for each item based on a range of latent scores, as well as the expected item score plots, can be useful ways of investigating and interpreting the relationship between different latent trait scores and students' expected item responses.

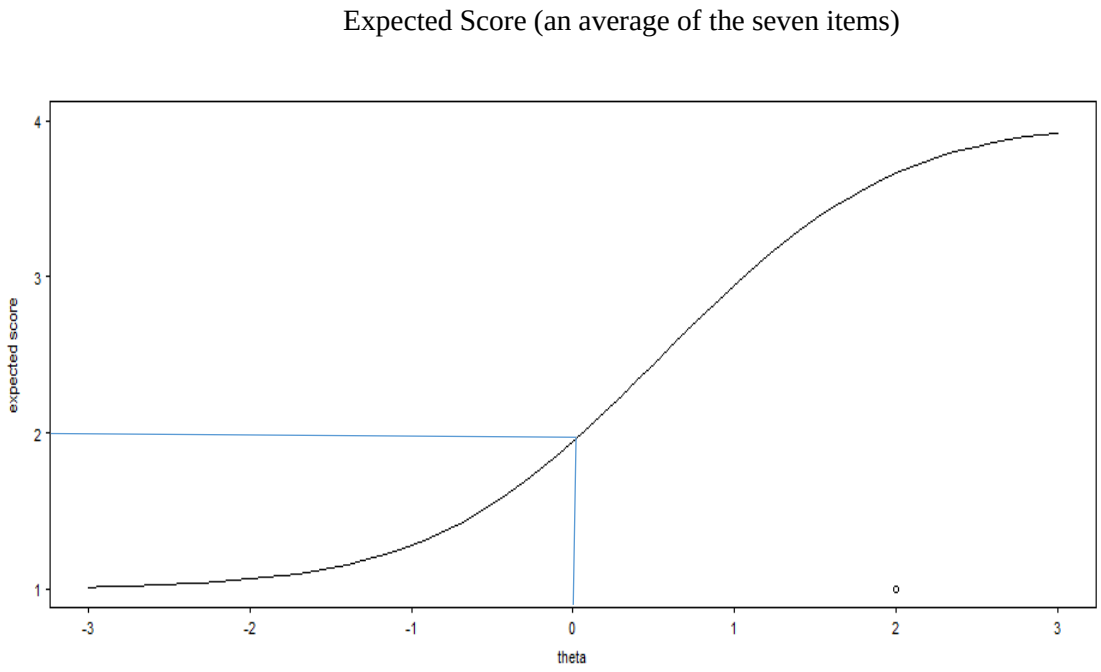


Figure 5.5: Average of the expected score across the seven items under the graded response model.

Results for the variance components for Model 1

The schools in HP rural and urban look similar in terms of their grand means and their variance components, and the TN rural and urban schools look similar in their grand means and variance components as well. As described in the previous section, we see that the between-school variance components indicate that there is appreciable variation across schools, and provides us information about the extent to which TN rural and urban schools vary in their school-mean STIMREAD values. The expected probabilities of a student in a school responding to category 4, for example, on an item helps us in interpreting these variance components.

From Tables 5.8 and 5.9, we see that the posterior means for the within-school variance (σ^2) for HP rural and HP urban is found to be 0.32 and 0.29, respectively. As can be seen, for both HP rural and urban, there appears to be substantially more variation among students within

schools, than variation in school-mean outcome scores across schools. The square root of the estimate of the within-school variance component for HP rural is 0.56 and for HP urban is 0.53 which provides us with an estimate of the standard deviation of outcome scores within schools in HP. Let's consider a student in a school in HP rural whose mean outcome score is equal to the grand mean (i.e., 0.17). A student in this school whose STIMREAD score is 1 SD above a value of 0.17 would have a latent outcome score of 0.17 + .56 = .73; for a student 2 SDs above a value of .17, their outcome score would be 1.29. Similarly, a student who is 1 SD below the mean would have an outcome score of -0.39, and a student 2 SDs below would have a score of -0.95. For HP urban, consider students in a school whose mean is equal to the grand mean estimate of .04. The square root of the within-school variance for HP urban is .53. For students 2 SDs below, 1 SD below, 1 SD above, and 2 SDs above the mean of 0.04, the corresponding outcome scores would be -1.02, -0.49, 0.57, and 1.1, respectively. This helps us see that there is substantially more variation within schools than between schools in terms of STIMREAD outcome scores in HP rural as well as urban regions.

Next, the posterior means for the within-school variance (σ^2) for TN rural and TN urban is found to be 0.25 and 0.28, respectively. The square root of the estimate of the within-school variance component for TN rural is 0.50 and for TN urban is 0.53 which provides us with an estimate of the standard deviation of outcome scores within schools in TN. Let's consider a student in a school in TN rural whose mean outcome score is equal to the grand mean (i.e., -0.1). A student in this school whose STIMREAD score is 1 SD above a value of -0.1 would have a latent outcome score of -0.1 + .50 = .40; for a student 2 SDs above a value of -0.1, their outcome score would be 0.9. Similarly, a student who is 1 SD below the mean would have an outcome score of -0.60, and a student 2 SDs below would have a score of -1.1. For TN urban, the grand

mean is found to be similar as TN rural, a value of -0.1. The square root of the within-school variance for HP urban is .53. For students 2 SDs below, 1 SD below, 1 SD above, and 2 SDs above the mean of -0.1, the corresponding outcome scores would be -1.16, -0.63, 0.43, and 0.96, respectively. This helps us see that there is substantially more variation within schools than between schools in terms of STIMREAD outcome scores in TN rural as well as urban regions. In the next section, we explore student factors that might be related to differences in student outcome scores within schools and investigate various school-level factors that might be related to differences in STIMREAD scores across schools.

Results from incorporating predictors (Model 2)

The student-level predictors such as ESCS, gender, teacher-student relations (STUDREL), and student's attitude towards school (ATSCHL) were added to the model. The student-level predictors were group-mean centered. School-level predictors included in the model were student-teacher ratio (STRATIO), and teacher shortage (TCSHORT). These variables were grand mean-centered. The student and school-level characteristics were added to level-2 and 3, respectively,

$$\theta_{ij} = \beta_{0j} + \beta_{1j} * ESCS - \overline{ESCS_{.j}} + \beta_{2j} * Gender - \overline{Gender_{.j}}$$
$$+\beta_{3j} * STUDREL - \overline{STUDREL_{.j}} + \beta_{4j} * ATSCHL - \overline{ATSCHL_{.j}} + e_{ij} \quad (14)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * STRATIO - \overline{STRATIO} + \gamma_{02} * TCSHORT - \overline{TCSHORT} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

where gender =1 for females and 0 for males, e_{ij} is level-2 residual, and the variance component σ^2 captures the amount of variation that remains in the θ_{ij}'s within schools after accounting for ESCS, gender, STUDREL, and ATSCHL.The regression coefficients can be interpreted as follows:

β_{0j} is the mean latent variable capturing students' perceptions towards STIMREAD in school *j* (by virtue of group mean centering the predictors)

$\beta_{1j} = \gamma_{10}$ is the expected change in a student's STIMREAD score when ESCS increases 1 unit, holding constant all other student-level predictors.

$\beta_{2j} = \gamma_{20}$ is the expected change in the difference between male and female students in their STIMREAD scores holding constant all other student-level predictors.

$\beta_{3j} = \gamma_{30}$ is the expected change in a student's STIMREAD score when STUDREL increases one unit, holding constant all other student-level variables.

$\beta_{4j} = \gamma_{40}$ is the expected change in a student's STIMREAD score when ATSCHL increases one unit, holding constant all other student-level variables.

γ_{00} is, by virtue of grand-mean centering of the school-level predictors, the expected school mean STIMREAD score when STRATIO and TCSHORT values are equal to their respective grand mean values.

γ_{01} is the expected change in a school's mean STIMREAD score when student/teacher ratio (STRATIO) increases 1 unit, holding constant TCSHORT.

γ_{01} is the expected increase in a school's mean STIMREAD score when teacher shortage (TCSHORT) increases one unit holding constant STRATIO.

and τ_{00} represents the between-school variance that remains after taking into account the school-level predictors.

The within-school and between-school sample sizes across the various groups (see Table 5.5) the number of predictors we can add at each level. For example, in HP urban there were only 11 schools and 189 students. Therefore, only one predictor was added at a time at level-2. For other regions such as HP-rural, TN-rural, and TN-urban no more than 3 predictors were added to level-2 at a time. The means and SDs for all predictors are included in Table 5.6.

As mentioned previously, the PISA report found no significant relationship between ESCS and the PISA reading achievement score. To that end, in Table 5.10 we see that the posterior means of the coefficient for ESCS (i.e., the relationship between ESCS and exposure to STIMREAD within schools) in all but one region were approximately equal to 0; in HP urban the estimate of the coefficient was equal to -.16. ESCS was also not related to differences in school mean STIMREAD scores in the HP and TN's rural and urban schools. In an exploratory analysis I found that school mean ESCS is not related to β_{0j} differences among the rural and urban schools, and so was not included as a predictor in the school-level mdoel. I also found that the school mean STUDREL and school mean ATSCHL variables were were not related to school mean STIMREAD scores, and so I did not include them in the school-level model.

The posterior means for the coefficients for gender (β_{20}) were also found to be small across all groups except for the HP urban region. For HP urban the posterior mean of the coefficient for gender is equal to -0.20, and the 95% Bayesian interval ranges from -0.43 to 0.03. Thus female students in HP urban tend to have a 0.20 point lower STIMREAD score compared to male students, holding constant the other predictors in the student-level model

Table 5.10: Posterior means, SD, and 95% intervals for the fixed effects and variance components parameters across the four regions for the STIMREAD construct. The student and school-level characteristics are added to this model. The three-level model was fit separately to the rural and urban regions of HP (e.g., HP rural, HP urban) and separately to each region in TN.

Parameter of Interest	HP -Rural	HP- Urban*	TN -Rural	TN- Urban
	Posterior Mean (SD)	Posterior Mean (SD)	Posterior Mean (SD)	Posterior Mean (SD)
Fixed effects				
Grand mean (γ_{00})	0.24 (0.10) [0.03, 0.45]	0.03(0.14) [-0.21, 0.29]	-0.09 (0.04) [-0.18, -0.007]	-0.25 (0.13) [-0.52, 0.17]
Student variables				
ESCS (β_{10})	-0.01(.03) [-0.07, 0.05]	-0.16 (.06) [-0.28, -0.04]	-0.003 (0.02) [-0.05, 0.04]	-0.004 (.03) [-0.07, 0.06]
Gender_female (β_{20})	-0.026 (0.06) [-0.13, 0.07]	-0.20 (0.12) [-0.43, 0.03]	0.003 (0.04) [-0.09, 0.09]	0.02 (0.09) [-0.15, 0.21]
STUDREL (β_{30})	0.25 (0.03) [0.17, 0.31]	0.29 (0.06) [0.17, 0.42]	0.24 (0.02) [0.21, 0.28]	0.24 (0.02) [0.20, 0.29]
ATSCHL (β_{40})	0.05 (0.038) [-0.004, 0.11]	0.11 (0.06) [-0.004, 0.23]	0.10 (0.02) [0.05, 0.14]	0.03 (0.04) [-0.05, 0.11]
School variables				
STRATIO (γ_{01})	0.001 (0.02) [-0.03, 0.03]		0.0009 (0.02) [-0.03, 0.03]	-0.01 (0.01) [-0.03, 0.006]
TCSHORT (γ_{02})	-0.09 (0.07) [-0.25, 0.06]		0.05 (0.14) [-0.22, 0.34]	0.12 (0.10) [-0.08, 0.33]
Variance components				
Within-school (σ^2)	0.27 (0.03) [0.21, 0.33]	0.18 (0.04) [0.11,0.29]	0.17 (0.02) [0.15, 0.21]	0.21 (0.02) [0.17, 0.26]
Between-school (τ_{00}^2)	0.18 (0.05) [0.11, 0.30]	0.16 (0.07) [0.10,0.35]	0.11 (.010) [0.10, 0.14]	0.12 (.02) [0.10, 0.14]

Note:
1.For HP Urban school-level predictors were not included in the three-level model. There were identification issues when the school-level predictors were added to the model.
2. Common IRT parameters of the slope and thresholds were used for HP rural and urban, and another common set of IRT parameters were used for TN's rural and urban regions

With the lack of previous studies in this area of study, where the outcome variable is instructional practices, it is hard to establish why the ratings of exposure to STIMREAD practices were lower for females than males. However, there is abundant research in rural regions in India where girl students are primarily made to stay home while a male student attends school.

The posterior mean for the coefficient of STUDREL (β_{30}) was found to have a positive magnitude across all four regions. For HP rural the posterior mean for the coefficient of STUDREL was found to be 0.25, and the 95% interval ranged from 0.17 to 0.31. That is, for every one-unit increase in the STUDREL variable while controlling for all other level-1 variables we expect a 0.25 increase in students' perceptions of their exposure to STIMREAD practices. For HP urban we find the posterior mean for the STUDREL coefficient is 0.29 and the 95% interval ranged from 0.17 to 0.42. Similar values for the coefficients were noticed for TN. A possible reason for this pattern of results is that students who have higher scores on the STUDREL variable may feel more engaged in class, and more willing to answer questions compared to students with lower STUDREL scores, and as a result, may report having more exposure to STIMREAD practices than students with lower STUDREL scores.

Next, the variable student's attitude towards school (ATSCHL) was found to have a small posterior mean across the four regions. In particular, for TN rural we find that the posterior mean for the ATSCHL coefficient was found to be 0.10 and the lower 2.5% interval was found to be 0.05 and the 97.5% upper boundary was 0.14. This interval is narrow, and the positive coefficient indicates that when a student's attitude increases by one unit, their perception of the amount of exposure to STIMREAD practices increases by 0.10 points, holding constant all other within-school predictors. This variable is found to have a weaker positive relationship with the outcome variable in comparison to the STUDREL variable. (Recall that both STUDREL and ATSHCL have an OECD mean of zero and an SD of 1, and so an increase of 1 unit in these variables corresponds to a 1 SD increase, which is a sizable increase).

Similar to STUDREL, students who have higher scores on the ATSCHL variable perhaps are more engaged, may volunteer to answer teachers' questions, and participate in classroom

activities. As noted by Schweig (2016), students in a given classroom who differ in key background characteristics may have very different educational experiences in the classroom, and thus may have very different perceptions of their amount of exposure to STIMREAD practices.

For HP rural, let's calculate what the expected latent score would be for a typical student i in school j when predictors are included in the model. We make use of the level-2 model in Equation 14. The $\beta_{0j}$for an average school is 0.17. The regression coefficients for β_{30} and β_{40} are 0.25 and 0.05, respectively, and the mean for STUDREL and ATSCHL is 0.59 and 0.10, respectively. Therefore, for a student in a typical school with average values for STUDREL and ATSCHL (that is, when we set the STUDREL value equal to the mean STUDREL score and the ATSCHL value equal to the mean ATSCHL score), we find the expected STIMREAD score (θ_{ij}) to be 0.17. Interpreting this in terms of the latent trait variable (theta), we find that the expected probability of a student with this latent score responding to category 4 for item 6 ("in all lessons") – an item with a relatively large slope parameter – is 48%. Next, let's consider a one-unit increase in the STUDREL and ATSCHL variables. This results in an expected θ_{ij} value of 0.47. The expected probability of a student with a latent score of 0.47 responding to category 4 on item 6 is found to be 63%. For HP's urban region, the latent score would be just lightly smaller, a value of 0.44; the expected probability of a student with a latent score of 0.44 responding to category 4 on item 6 is found to be 61%.

In terms of the school-level variables, we find that the posterior means of the coefficients for both student-teacher ratio (STRATIO) and teacher shortage (TCSHORT) variables have small magnitudes across all regions. Note that in Figure 4.2 we saw that the TCSHORT for both HP rural and TN rural have some variation, however, we see that most principals reported that there is no shortage of teachers.

Next, we see that the posterior means of the within-school variance components decrease across all four regions when we add predictors to the model (see Tables 5.8 to 5.10). That is the student-level predictors in the level-2 model accounts for some of the differences in the students' perceptions towards the STIMREAD construct. For HP rural, the posterior mean for the within-school variance reduces to a value of 0.27 from 0.32, and for HP urban it reduces to 0.18 from 0.29, that is, a 37% reduction in the within-school variance component for HP urban. For TN rural and urban we see that the within-school variance reduces to a value of 0.17 and 0.21, from 0.25 and 0.28, respectively. Lastly, we did not see any change in the between-school variance after the addition of the school-level predictors.

5.5 Additional analysis

One question that may arise from the above three-level analysis is whether a "multilevel IRT" approach seems to be required. Would a summed score approach – a less-complex approach -- produce similar patterns of results? To this end, I analyzed the data for the STIMREAD construct using the summed score approach. For construct 1 (STIMREAD) the summed score outcome variable consisted of seven items (see Table 1 for the specific items). Two-level models were then specified and run using the HLM software program, in which the summed scores for the construct were treated as outcomes in a student-level model, and differences across schools in their mean summed score values were investigated via a school-level model. The student- and school-level predictors included in this model are similar to the ones in the three-level models discussed earlier.

The results for this set of analyses are presented in Appendix B, Tables B.2 to B.6. The tables present complete descriptive statistics for HP and TN, along with results from fitting a two-level HLM model to the STIMREAD outcome scores in each of the four regions. Tables B.5 and

B.6 describe the unconditional model (i.e., the model with no predictors), and the model with student- and school-level predictors, respectively. The results tabulated in Tables B.2 to B.6 have similar patterns to the results obtained from fitting a three-level model to the data (see Tables 5.8 to 5.10).

Using the summed scores as outcomes in the multilevel analyses allows us to assess the amount of variation within schools and across schools and investigate various relationships while "staying close to the data". However, the multilevel IRT approach has a few advantages.

First, a latent variable model (e.g., IRT models) provides a statistical framework for relating the observed responses to test items to the students' standing on unobserved (latent) variables. For example, in this study, students responded to seven items which together form the STIMREAD construct. A key point is that working with latent scores obtained via an IRT model (e.g., a graded-response model) allows us to make meaningful interpretations of these scores, as illustrated earlier in this chapter. For example, for a given latent score, we can obtain an expected probability of a person with that score experiencing exposure to a particular instructional practice (e.g., "The teacher helps students relate the stories they read to their lives") during all lessons (or during most lessons, or during some lessons, or never or hardly). These probabilities provide useful information about students' exposure to instructional practices of interest; they provide valuable, accessible, fairly concrete interpretations of the latent scores. In comparison, summed scores do not appear to provide information as interpretable and concerete.

Secondly, the responses on a particular test item can be explained (or modeled) using the information from the respondent and the item, allowing us to make inferences about item properties, for example, the item discrimination parameter providing information concerning how differences in student responses to a given item (e.g., 1, 2, 3 or 4) relate to differences in the

magnitudes of students' probabilities of experiencing the practice of interest never or hardly, during some lessons, during most lessons, or during all lessons.

Third, assessing the item properties is particularly important in large-scale assessments, such as in ECLS, PISA, or TIMSS, which are administered across multiple countries, since the construct might be used in subsequent years for testing. Theoretically, the latent scores are independent of the test as opposed to the observed scores, and sothe latent scores would allow the possibility of using the results across different tests in an analysis.

Fourth, when working with summed scores, the estimate of within-school variance (σ^2) would reflect both true score variance and error variance in the summed scores within schools. This, in turn, could also result in underestimates of the amount of between-school variance. However, in a comprehensive latent variable framework, the estimate of σ^2 provides us with an estimate of true score variance in STIMREAD within schools.

5.6 Summary of Findings

First, measurement invariance was established for the STIMREAD construct across the rural and urban regions of HP, and across the rural and urban regions of TN. Using the Wald statistics and the associated p-values from both sets of analyses, it was seen that no item had a p-value below 0.05. This indicates that the items appeared to be interpreted in the same manner by the students of the rural and urban regions of HP, and by the rural and urban regions of TN. This is vital when we want to compare students' latent traits across different regions and answer key substantive research questions. Furthermore, this set of analyses also helps us assess the fit of the graded response model before we build the three-level hierarchical model for comparisons of the sub-groups of interest. This phase of my dissertation yielded a set of common item parameter estimates that could be used in comparing students and schools in the rural and urban regions of

HP, and a common set of item parameter estimates that could be used in comparing students and schools in the rural and urban regions of TN

Moreover, in contrast to principals' responses to questions regarding various items including the possible shortages of qualified teachers, and rates of teacher absenteeism, the STIMREAD scale that was constructed using student responses provides a latent variable measure of students' exposure to key instructional practices, in this case, teachers' stimulation of reading engagement practices. An important point is that working with latent scores obtained via an IRT model (e.g., a graded-response model), makes possible meaningful, useful interpretations of these scores, as illustrated earlier in this chapter. For example, for a given latent score, we can obtain an expected probability of a person with that score experiencing exposure to a particular instructional practice (e.g., "The teacher helps students relate the stories they read to their lives") during all lessons (or during most lessons, or some lessons, or never or hardly).

Working with a latent variable measure of students' exposure to stimulation of reading engagement practices opens up an array of possibilities. In particular, we can use student latent variable exposure scores as outcomes in multilevel models. In TN rural, for example, we can estimate a grand mean exposure score, and estimate how much variation there is in STIMREAD exposure scores across schools (i.e., how much schools vary in their mean STIMREAD scores), and how much variation there is in student STMIREAD scores within schools. Moreover, we can investigate how differences in various school-level predictors relate to differences in school-mean STIMREAD scores, and how differences in various student-level predictors relate to differences in student STIMREAD scores within schools. Such analyses can be conducted in each region of interest (i.e., HP rural, HP urban and TN rural, TN urban).

Next, a three-level model was fit across the rural and urban regions of HP and the rural and urban regions of TN. The posterior mean for the grand mean (γ_{00}) for the STIMREAD construct represents the overall average latent score for students' exposure to the STIMREAD construct (θ_{ij}). The grand means values for the HP rural and urban regions were found to be 0.17 and 0.04, respectively. To relate these trait level estimates to the item response categories, we can calculate the expected probability of students to individual items and/or plot the expected item scores for an individual item. For item 6 (r*elate to lives*), for example, a student with a latent score of 0.17 has a 48% probability of responding to category 4, and a student with a latent score of 0.04 has a 42% probability of responding to category 4. Calculating the average of the expected scores across all seven items, we can find, for example, the predicted category that a student with a latent trait of zero, that is an average student, would respond to. (Please see Figure 5.5 above that shows the relationship between different latent trait scores for STIMREAD and students' expected item responses.)

The within-school variance for rural and urban regions of HP was found to be 0.32 and 0.29, respectively. The within-school variance of rural and urban regions of TN was also found to be close, a value of 0.25 and 0.28, respectively. Converting these variances to standard deviations, the standard deviations indicate that there is an appreciable amount of variation within schools, and we do not see much difference in this when comparing rural and urban regions. Similarly, we do not find large differences in the between-school variance while comparing the rural and urban regions of HP and the rural and urban regions of TN. Due to the lack of additional datasets on rural and urban regions in India on instructional practices, it is hard to elaborate on why we see no differences. One possibility is that the student responses on the items are unable to provide the necessary information to capture the differences among the rural and urban regions

– that is, we may require more than four categories for these items or we could use another latent construct capturing instructional practices to capture the similarities or differences among the regions.

We notice that the amount of variations between schools is smaller in comparison to the within-school variance. The schools across the rural and urban regions of both states are primarily in the low-socioeconomic areas, which could be one of the reasons for less variations between schools. We may see teachers with similar experiences or qualifications teaching students in these regions.

Third, student-level predictors such as ESCS, gender, students' attitude towards school (ATSCHL), and teacher-student relations (STUDREL) were added to the student level model, and student-teacher ratio (STRATIO) and teacher shortage (TCSHORT) variables were added at the school level model. The addition of the student- and school-level characteristics explained only a small amount of variation in the within-school and between-school variances in these instructional practices, and the school-level variables were found to have no relationship with the outcome variables outcomes for HP and TN (Walker, 2011). Two variables capturing students' attitudes towards school and classroom environments (ATSCHL and STUDREL) were found to have a positive relationship with the STIMREAD construct. Recall, STUDREL and ATSCHL help us better understand what the student-teacher relationship looks like and what students' perceptions are about their school. For HP rural, the coefficient for STUDREL[5] is found to have a posterior mean of 0.25 (nearly a quarter of an SD) and the 95% interval ranged from 0.17 to 0.31. For HP urban, the posterior mean is found to be 0.29 and the 95% interval ranged from 0.17

[5]Recall that both the STUDREL and ATSHCL scales follow an OECD standard normal distribution with a mean of zero and SD of 1.

to 0.42. Note that these intervals lie above a value of 0. Similarly, for both TN rural and TN urban, the posterior means for the STUDREL coefficients were found to be 0.24 with 95% intervals ranging from 0.20 to 0.28. The coefficient for the ATSCHL variable was found to have a small posterior mean across the rural and urban regions of HP and TN

For a one-unit increase in the STUDREL and ATSCHL variables, for HP rural we find the expected probability of a student with a latent score of 0.47 responding to category 4 is found to be 63%. For HP urban for a one-unit increase in the STUDREL and ATSCHL variables the latent score is just slightly smaller, a value of 0.44, and the expected probability of responding to category 4 is found to be 61%.

Lastly, as noted previously, we need to be cautious while making inferences about the results comparing the rural and urban regions of HP and the rural and urban regions of TN. Since students within schools were sampled randomly from lists of students that were incomplete, it was not possible to assess the representativeness of the resulting samples.

Multiple studies (e.g., MET study, TALIS) have tried to fill the gap in connecting instructional practices to student learning by making use of various tools (e.g., student and teacher surveys, teacher portfolios) to capture classroom practices and analyzing these data to study the impact of instructional practices on student learning – often used as predictors in the model. However, it can be challenging to conduct large-scale studies such as these in countries like India, where there is a fear of accountability and most schools are often public schools (run by the governement of India). More data sources and research is needed to understand the quality of school resources and teaching in various regions of India.

Chapter 6

Examining school-specific estimates of students' exposure to STIMREAD practices across public and private schools in rural regions

RQ3: What do the school-specific estimates of students' exposure to STIMREAD practices look like across public and private schools in TN's rural region?

In this chapter, I examine the school-specific estimates of students' exposure to STIMREAD practices across public and private schools in TN's rural region. The large sample size of students and schools in TN allows us to better understand what might be going on in the public and private schools in the rural region. Moreover, since the majority of the schools across India are located in rural regions, and are run by the government, this analysis might be helpful for taking a closer look at the data and showing how this approach can help in the identifying schools that may be thriving and those that may need assistance. Private schools are often rare in rural regions; however, parents prefer private schooling for their children due to better facilities and resources in comparison to public schools (Lewin, 2011).

In this chapter I first lay out the data and measures, then I fit a three-level model to the public and private schools, and assess and plot the school-specific estimates of exposure to STIMREAD practices with respect to students' socio-economic status, and lastly, present the results.

6.1 Data and Measures

For this set of analyses, the data are restricted to the public and private schools of TN's rural region. The sample size for this study consists of 944 students across 49 public schools and 395 students across 23 private schools in TN's rural region. That is, approximately on average there are 19 students per school in public schools in the TN rural region, and there are 17 students per school on average in the private schools in the TN rural region. In Table 6.1 I have tabulated

the sample size, means, and SDs for the items for the STIMREAD construct, and various student- and school-level variables for TN's public and private schools in the rural region. Recall, we need to be cautious while making inferences regarding the results for various schools in the sample since we lack information on how representative the students within a given school are of the population of children in that school.

Table 6.1: Descriptive statistics for the items for STIMREAD, and the student- and school-level predictors for TN rural regions' public and private schools.

Variables	**TN rural public school**			**TN rural private school**		
	N	Mean	SD	N	Mean	SD
1.Explain Expectations	922	2.61	0.98	391	2.64	0.96
2.Check Concentrating	919	2.97	0.94	390	3.05	0.89
3.Discuss work	915	2.87	1.00	390	2.85	0.96
4.Explain judgements	911	2.75	1.02	389	2.7	1.01
5.Ask if understood	910	2.96	0.94	391	3.02	0.92
6. Mark work	915	2.94	0.98	390	2.87	1.00
7.Student questions	920	3.07	0.96	391	2.95	0.96
8.Motivating questions	922	2.61	0.98	391	2.64	0.96
9.Immediate feedback	919	2.97	0.94	390	3.05	0.89
Student-level						
ESCS	939	-2.28	0.94	395	-1.67	1.16
Gender (female)	944	0.51	0.5	395	0.44	0.5
Student-teacher relations (STUDREL)	937	0.49	1.19	395	0.42	1.11
Attitude towards school (ATSCHL)	893	-0.23	0.82	381	-0.12	0.85
School-level						
Student-teacher Ratio (STRATIO)	49	35.29	13.27	22	33.4	12.71
Teacher Shortage (TCSHORT)	49	0.24	1.03	23	-0.32	0.76

6.2 Analysis and Model Specification

Using the data for the public and private schools for TN's rural region, a three-level model was fit in JAGS using MCMC. As before, for this analysis, the item parameters in the three-level

analysis in JAGS were set to the ML estimates of the item parameters obtained from fitting a GRM to the items. The model specification for this analysis is similar to the model described in Equations 13, i.e., an unconditional model is fit to the data to obtain an estimate of the grand mean (γ_{00}), the within-school variance component (σ^2), and the between-school variance component (τ_{00}).

Recall that the STIMREAD latent values for the individuals in a given school are treated as outcome values in a within-school model. A key parameter in the within-school model is β_{0j} which represents the mean STIMREAD value for the students in school *j*. In a level-3 (between-school) model, the β_{0j}'s are viewed as outcomes and are modeled as a function of a grand mean STIMREAD value for the schools in a given region and sector (e.g., TN rural public schools). (The within-school and between-school models can be expanded to include predictors.)

As noted in chapter 3, the estimate of β_{0j} for a given school based on the data for the sample of students in that school, will be shrunk a certain amount toward the grand-mean STIMREAD score for the schools in that region (e.g., TN rural public schools). The amount of shrinkage depends upon the amount of error variance connected with the estimate of β_{0j} based on the data for school *j*'s students (i.e., σ^2 / n_j) and upon the amount of between-school variance in the β_{0j}'s across the schools in that region (i.e., τ_{00}). When τ_{00} is very large relative to the error variance in the estimate of β_{0j}, the estimate of β_{0j}will be shrunk only slightly toward the grand mean. When τ_{00} is very small relative to the error variance in the estimate of β_{0j} – when the school mean STIMREAD scores for the schools in that region are clustered tightly together – the estimate of β_{0j}will be shrunk markedly toward the grand mean. Thus for each school we obtain what is referred to as a shrinkage estimator of β_{0j}. More specifically the analyses I conduct for this chapter yield a posterior distribution of the shrinkage estimator for each school. In the

Bayesian framework, this sort of shrinkage process is referred to as borrowing strength – or borrowing information – from other similar schools.

6.3 Results for public and private schools in TN's rural region

In Table 6.2 I have tabulated the posterior means along with the 95% intervals for the grand mean and the variance component parameters for the teachers' stimulation of reading engagement (STIMREAD) construct across the public and private schools of TN's rural region. The posterior mean for the grand mean (γ_{00}) parameter for public schools in TN and the posterior mean for the grand mean for private schools in TN were found to be 0.074 and 0.11, respectively. For the public schools, the lower boundary of the 95% credible interval was found to be -0.03 and the upper boundary was found to be 0.18. In the Bayesian framework, we can say that there is a 95% probability that the true value of the grand mean lies between -0.03 and 0.318. The lower boundary of the 95% interval for the private schools in TN's rural region, was found to be -0.14 and the upper boundary is 0.36, which is slightly wider than the interval for the public schools. The posterior means for the within-school variance (σ^2) for the public and private schools were found to be 0.41 and 0.35, respectively, and the posterior means of the between-school variances are 0.11 and 0.32, respectively.

Caterpillar plots for school-specific estimates for public and private schools

We take a closer look at the caterpillar plots (see Figure 6.1) that display the posterior means for the school mean exposure scores (i.e., the β_{0j}'s) and their corresponding 95% intervals, for the public and private schools in the rural region of TN. The β_{0j}'s can be viewed as school-specific latent variables capturing the extent to which students in each school report experiencing the STIMREAD teaching strategies. The caterpillar plots in Figure 6.1 help us see which schools

have the lowest posterior mean values of mean exposure to STIMREAD and those that have high posterior means on the STIMREAD outcome.

Table 6.2: Posterior means, SD's, and the 95% intervals for the grand mean and the variance component parameters for the STIMREAD construct across the public and private schools of TN's rural region. The three-level model was fit separately to each set of data using a common set of IRT parameters.

Parameter of Interest	**TN-rural public school**	**TN-rural private school**
	Posterior Mean (SD)	Posterior Mean (SD)
Fixed effects		
Grand mean (γ_{00})	0.074(0.05) [-0.03, 0.18]	0.11(0.13) [-0.14, 0.36]
Variance Components		
Within-school variance (σ^2)	0.41 (0.03) [0.35, 0.49]	0.35 (0.02) [0.26, 0.46]
Between-school variance (τ_{00})	0.11 (0.01) [0.10, 0.14]	0.32 (0.14) [0.12, 0.69]

The minimum value of the school-mean exposure for public schools was found to be -0.53, and the maximum value of the school-mean exposure was found to be 0.40. The expected probabilities of students in public schools with the minimum and maximum school-mean exposure values responding to category 4 on item 7 are found to be 23%, and 61%, respectively. For students in private schools, the expected probabilities of students with a minimum school-mean exposure value (-0.37) and maximum school-mean exposure value (0.52) responding to category 4 of item 7 are found to be 28% and 66%, respectively. The "outlying" provate school in Figure 6.1 is found to have a school-mean exposure value of 1.75, and the expected probability of a student with this latent trait is reponding to category 4 on item 7 is found to be 94%.

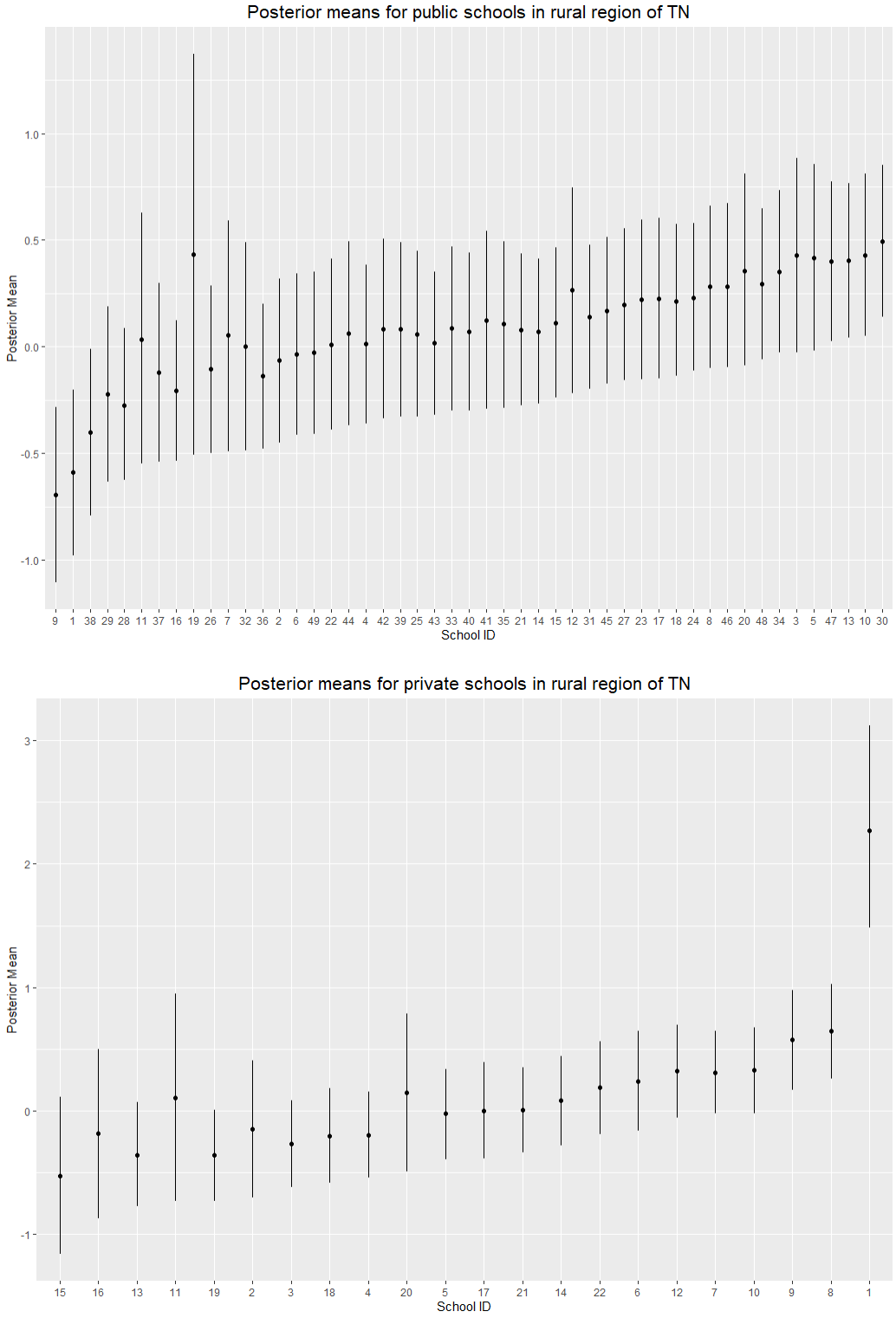


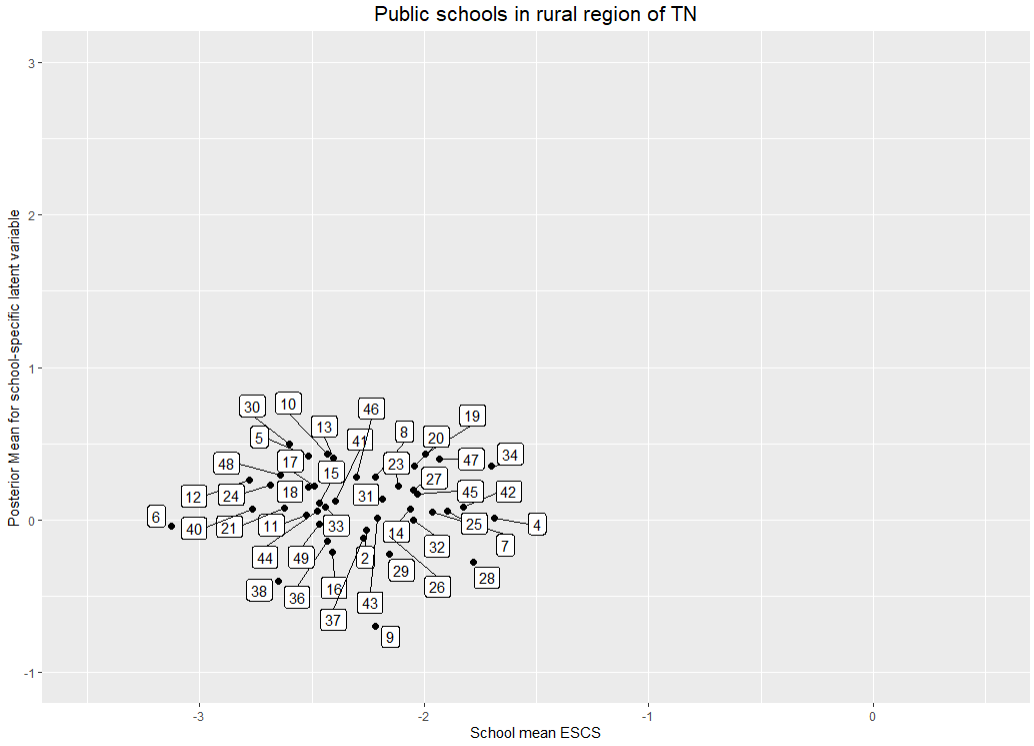
Figure 6.1: Caterpillar plots for the public schools (top plot) and private schools (bottom plot) in TN's rural regions for the STIMREAD construct.

Plots depicting the relationship between public schools and private schools and socio-economic status

Lastly, the posterior means for the school-specific variables (β_{0j}'s) for the public and private schools in the rural region of TN were plotted with respect to their school mean ESCS values in Figure 6.2. Both plots for public and private schools are on the same scale for the school mean ESCS (x-axis), thereby depicting the differences in the socio-economic status between the private and public schools in a rural region. These plots enable us to see if there are any systematic patterns between school mean ESCS and the magnitude of the posterior means of the β_{0j} values. We see that the public schools (see Figure 6.2 top plot) are largely concentrated in the lowest end of the socio-economic scale in the range of -3 to -1.6. This difference in the socio-economic status (ESCS) between the public and private schools in TN's rural region is also evident from the low mean of the ESCS variable for the public schools (e.g., a value of -2.28) in comparison to the ESCS mean for private schools (i.e., -1.67).

It is worth noting that school id 1, a private school that had a high posterior mean for exposure to teachers' stimulation of reading engagement in comparison to other schools is situated in a low socio-economic area (see Figure 6.2 bottom plot). Such schools are schools that we might want to take a closer look at via interviews with their principal and teachers. 8 students from this school participated in the study (students are 15year olds) and the school size is 246. However, we do not have data on whether the school is only a secondary school or a school with primary and secondary grades. The majority of the public schools seem to be concentrated in the lowest end of the socio-economic scale, whereas the private schools are more spread out, but still in low-income areas. Furthermore, in both the plot for the public schools and the plot for the

private schools, we do not see a systematic relationship between the posterior means and school ESCS values.



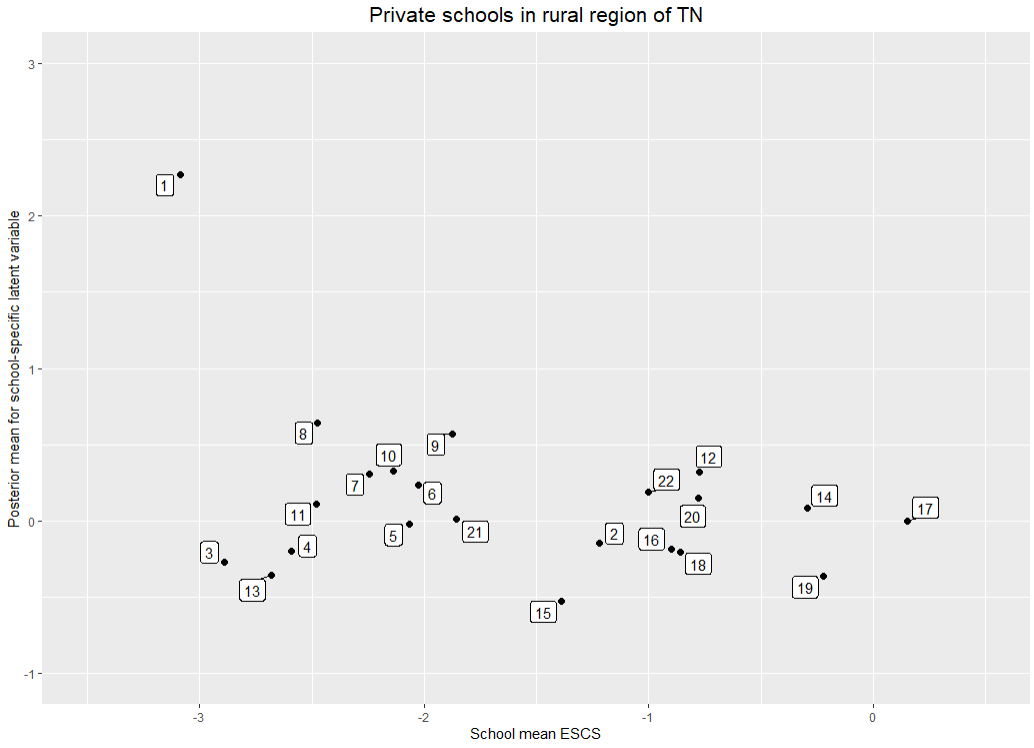


Figure 6.2: Posterior means of the school latent variable ($\beta_{00}'s$) for public and private schools in the rural region of TN, by their school mean ESCS values. The two plots have the same scale for school mean ESCS (x-axis), thereby bringing to light the differences in socioeconomic status between the private and public schools in the rural region.

6.4 Summary of Findings

A three-level model was fit to the TN rural-public schools and TN-rural private schools, where the STIMREAD construct was the outcome variable.

First, we see that the posterior means for school-mean STIMREAD scores for the public schools vary tightly around the grand mean STIMREAD score of the public sector, and this is especially the case for private schools, with the exception of an outlying private school with a large school mean STIMREAD score (see Figure 6.2). As noted previously, within a givcen school, students were randomly selected from incomplete lists of students, and as such it was not possible to assess the representativeness of the within-school samples. Thus caution needs to be exercised regarding the results, since they could be biased to some degree. The methodology employed to examine school-specific estimates of students' exposure to key instructional practices across public and private schools when using large-scale assessments can be valuable to other researchers, or officials in school districts to assess school's quality by identifying schools that need additional assistance and suppor.

To make these latent scores more interpretable, and to capture the differences between the schools in terms of students' exposure to the STIMREAD practices I calculated the expected probability of a student responding to, for example, category 4 across the seven items. The expected probability values allow us to interpret the relationship between the student latent trait (i.e., students' perceptions of exposure to STIMREAD in a school) and the students' responses. We can capture how probable they are to respond to a higher category based on their latent trait. For example, we notice that students across the public schools had a minimum 23% probability

of responding to category 4 on item 7 and a maximum of 61%, and students in private schools had a minimum expected probability of 28% of responding to category 4 of item 7 and a maximum of 94% in the case of the outlying school.. This implies that in some schools students' perception about the amount of exposure they had to STIMREAD teaching practices was as low as 23% of responding to the highest category, that is category 4 ("in all lessons").

Next, the between-school variance component for public schools and private schools suggests that there is considerable variation in students' perceptions towards their exposure to STIMREAD practices across both public schools and private schools. We see that there is appreciably more variation within schools than between schools based on the expeced probabilities for students in public and private schools.

Lastly, plotting the posterior means for the school-specific variable (β_{0j}'s) for the public and private schools in the rural region of TN with respect to their school mean ESCS revealed a private school with an estimated mean STIMREAD score that was much higher than the estimated mean STIMREAD scores for all other schools in the analysis. In addition, the majority of the public schools were found to be concentrated in the lowest end of the socio-economic scale, whereas the private schools were found to be more spread out, but still in low-socioeconomic areas. These plots allow us to, for example, take a closer look at the school resources for one school in comparison to other schools. We can pinpoint schools that might be in low socio-economic areas that tend to have low exposure to instructional practices or school resources, for example, school #6 among the public schools in a rural region, which has the lowest ESCS value among the public or private schools, and schools in low socioeconomic areas that tend to have high levels of exposure to key instructional practices, i.e., school #1 among the private schools,

which has a posterior mean of 1.75. In particular, schools such as school #1 among the private schools would be of special interest.

Chapter 7

Discussion

Access to quality instruction and the equal distribution of educational opportunities is fundamental for every child irrespective of their background. This dissertation study brings together substantive questions around access to instructional practices and school resources, and multilevel IRT models to illustrate how latent variables capturing students' perceptions towards instructional practices can be used as outcome variables to examine the extent of variation in exposure to such practices within and between schools. To my knowledge, this study is the first in the context of India to investigate access using items focusing on instructional practices as the outcome, instead of students' cognitive test scores or other school-level indicators such as enrollment numbers or teacher shortage variables. One of the motivations for this work, apart from India's poor performance in PISA 2009, is the lack of research on students' access or exposure to key instructional practices and other school resourcesusing large-scale assessments and employing these measures as the outcome variables in a study.

Improved access to various classroom instructional practices and school resources can not only generate greater interest in schooling but also encourage students to remain in schools and complete their primary education. Increasingly, families, particularly low-income families are hesitant to send their children to government schools while questioning the value of education and the quality of education (Kingdon, 2007; Maclean & Vine, 2003). Investments (e.g., monetary) made towards school resources, hiring of qualified teachers, and providing instructional materials, for example, can greatly improve access to educational opportunities for students, and will have a positive impact on student learning and retention of students in schools.

7.1 Main Takeaways from this dissertation study

For this study, I make use of the items from PISA 2009's non-cognitive background questionnaires completed by students and school principals. In particular, I take a closer look at various school resources (e.g., the percentage of qualified teachers in a school) and student's exposure to teachers' use of stimulation of reading engagement (STIMREAD) practices. The rural/urban divide in India has been discussed by many authors (e.g., Das & Zajonc, 2010a; Govinda & Bandyopadhyay, 2008), and access to educational opportunities and quality instruction has been a concern for students who live in remote areas of India with lack of access to schooling. This study illustrates how large-scale assessments can be utilized to study these various issues.

For RQ 1, based on previous literature across the rural and urban regions of India I looked at a few key school-level indicators such as teacher shortage, teacher absenteeism, and student absenteeism. The results in chapter 4 indicated a lack of teacher shortage and teacher absenteeism rates across schools, however, one question that arises is why students performed poorly in the PISA 2009 tests. There are at least two possibilities. (1) First, PISA rankings are based on test items in reading, mathematics, and science, and the teaching practices across the sample schools may not provide students enough knowledge to answer these high-level cognitive items. This issue may be particularly salient because nearly 80% of HP schools were in rural regions, and 68% in TN, and there are well known concerns about teaching quality in rural schools in india (e.g., Agrawal, 2014; Maclean & Vine, 2003). (2) Second the data reported by the school principals in the school questionnaires may not be entirely trustworthy— responses may be influenced by social desirability, where principals report in a manner they think makes their school look good.

For RQ2, first, I investigated measurement invariance for non-cognitive items from the student background questionnaire using an IRT approach across rural and urban regions of HP and rural and urban regions of TN. It was established that measurement invariance holds for the items that form the teachers' stimulation of reading engagement (STIMREAD) construct across rural and urban regions of HP and across the rural and urban regions of TN (see section 5.4.2 RQ2a: Assessing Measurement Invariance using IRT Models Across Different Regions). Measurement invariance allow us to make sensible inferences in comparing rural and urban regions in HP and TN in RQ2b.

Most previous research studies have examined measurement invariance for cognitive test items since they are the outcome variable(s); the current work, however, illustrates the use of an IRT approach to examine measurement invariance using non-cognitive items from a large-scale international assessment across regions within a country. This allows researchers to examine psychometric properties of the items, including equating test scales, and exploring differential item functioning to refine the items in the study for future research work.

Next, to examine RQ2b a three-level multilevel IRT model was fit to the student responses to the items that capture students' perceptions regarding teachers' use of stimulation of reading engagement (STIMREAD) construct. This section also brings to light the importance of working with latent scores obtained via an IRT model (e.g., a graded-response model), which allows us to make meaningful interpretations of these scores – e.g., for a given latent score, we can obtain an expected probability of a student with such a score experiencing a particular instructional practice (e.g., "The teacher helps students relate the stories they read to their lives") in all lessons (category 4) or most lessons (category 3), or never or hardly (category 1). These probabilities provide valuable information about students' exposure to instructional practices of interest and are

accessible interpretations of the latent scores in comparison to information provided by the summed scores, which might not be as meaningful. The latent score values can be related to the item responses via two approaches. First, we can calculate the expected probability of students endorsing particular categories for individual items. For example, in HP's rural region a student with a latent score of 0.17 (the grand mean) has a 48% probability of responding to category 4 on item 6 (r*elate to lives*) and in HP's urban region a student with a latent score of 0.04 (the grand mean) has a 42% probability of responding to category 4 (see section 5.4.3 RQ2b: Three-level Multilevel IRT Models: A Fully Bayesian Approach). Another approach is to plot the expected item score on a scale of 1 to 4 for each of the items for a range of different theta estimates, or one could plot an average expected score across the seven items (please see Figure 5.5 for an example).

As noted earlier, because within schools students were randomly sampled from incomplete lists, it was not possible to assess how representative the sample of students within a given school were of the population of children in that school. As such we need to be cautious about drawing conclusions based on the data for HP and TN (see Walker, 2011, p.104).

Lastly, for RQ3 I examined the school-specific estimates of students' exposure to STIMREAD practices across public and private schools in TN's rural region. We see that a student in a public school with a school-mean exposure value of 0.075 (the grand mean) has an expected probability of 47% of responding to category 4 on item 7 (*Build on Knowledge*). For private schools, an outlying school with a large posterior mean of 1.75 has an expected probability of 94% of responding to category 4 on item 7.

The expected probabilities of the minimum and maximum values of the school-mean exposure values for students in public schools ranged from 23% to 61% and for students in private

schools it ranged from 28% and 66%. These results suggest that students experience similar STIMREAD practices across public and private schools in TN's rural region. One may see these results if the students and schools participating in the study are from similar regions or have similar background characteristics, or are working with similar curricula. A key finding of this set of analyses suggests that a majority of the public and private schools were concentrated in the lowest end of the socio-economic scale, and private schools were found to be slightly more spread out.

Further clarification from principals and/ or teachers is needed to evaluate the results in Chapter 6. However, these analyses can be helpful to identify schools where exposure to key – practices is particularly high or low. Low exposure schools would require close attention in terms of additional school resources, instructional materials, or assistance with quality teachers to teach various courses. These results also indicate the need for a smaller and more focused follow-up qualitative study. Targeted interviews with students, principals, and key school officials could be the next step to understanding the similarities and differences we see across public and private schools, and shedding light on the outlying school with the large posterior mean exposure value.

One could make use of another data source, for example, either from India or another country to illustrate this methodology, where instructional practices are directly used as outcomes to capture what students are experiencing in classrooms instead of focusing on indicators such as teacher shortages or test scores, and to illustrate how equitably (or inequitably) students perceived exposure to key practices are distributed within schools and between schools.

Takeaways from the Methodological Approach and limitations of the study

The focus of this dissertation study has been on school resources and non-cognitive outcomes that have primarily been used as predictors in various studies. There are a few key

points that this dissertation brings forward on the methodological front. First, this study makes use of an MCMC approach to estimate a three-level multilevel model. While maximum likelihood provides us with estimates of parameters of interest and standard errors, MCMC yields the marginal posterior distributions of parameters of interest – a probability distribution for each parameter that provides us with various point estimates (e.g., a posterior mode, mean and median) for that parameter, as well as probabilities that the parameter of interest exceeds or lies below certain values, or lies within a particular range of values, and a 95% interval based on the lower .025 quantile of the posterior distribution and the upper .975 quantile. In this study, students' responses to the Likert-scale items are nested within students, and students are nested within schools. Using a multilevel-IRT framework allowed us to model these student responses, and student-level and school-level information simultaneously. Additionally, multiple software packages (e.g., Stata, Mplus, MLWin) enable us to implement MCMC in a broad variety of modeling settings, in particular high-dimensional settings. However, one needs to be cautious while using these approaches by using sensible priors, and appropriate starting values.

Secondly, in the case of multilevel models using MCMC, results can be sensitive to the choice of priors for the variance components, especially between-school variance components. That is, certain priors for variance components can potentially cause MCMC algorithms to derail, i.e., fail to converge. Specifically, caution must also be taken while specifying the lower bound of a prior for variance components, for example, a uniform prior as in this study. It was observed initially that a lower bound of 0.5 on the uniform prior for a between-school variance component, was too large and resulted in truncating the lower portion of the marginal posterior distributions of the between-school variance components at a value of .5. This was remedied by setting the lower bound to 0.1.

Third, care should be taken in assessing the convergence of these complex models and appropriate checks must be conducted to ensure that the model has converged (e.g., examining trace line plot, marginal posterior for each parameter). In some scenarios even though an MCMC algorithm seems to have converged, the posterior means and/ or standard deviations for certain parameters may be nonsensical. This may be due to the priors one is specifying, or the need to place some constraints among the model parameters to avoid identification issues, or perhaps large correlations between certain parameters may be causing problems. Therefore, even if a complex model (e.g., three-level models as in the current study) seems to have converged, it is crucial to inspect the marginal posteriors of each parameter before making inferences.

There are few limitations of this study. First, the addition of binary predictors to the school-level model (e.g., whether a school was a public school or private school) raised some potential problems in estimating key parameters in the three-level model when the level-1 model (i.e., the measurement model) is a graded-response model. Adding binary predictors to the model (e.g., private vs. public schools) results in a fully specified multiple group model, that is, all the group-specific item parameters fully capture the differences between the two groups (e.g., private and public schools), and so it is not possible to estimate regression coefficients capturing differences between private and public schools, for example, in their mean outcome scores, in analyzing the data for, say, TN Rural. This results in identification issues, as the fully specified multiple group model is not identified. The standard deviations associated with the posterior means of the regression coefficients for the level-three indicator variables in such situations were found to be extremely high, which serves as an indicator of the problem discussed in this section. This problem was one of the motivations to examine each of the sub-groups separately (e.g., TN rural and TN urban, and public schools and private schools in a particular region).

Second, is the concern about the PISA data and the kind of conclusions we can draw based on the available data for the various regions and countries. In the case of India, as stated in the previous chapters, there were some concerns about the information supplied by the school or district officials. Another factor to consider is that the lack of information regarding the actual number of schools across the rural and urban regions of HP and TN at the time of the study makes it difficult to assess if the samples of schools in those regions were representative of the populations of schools in the rural and urban regions of HP and TN. Thus it's a possibility that there might be schools that are performing better than the sample of schools available in this study. PISA does provide sample weights for various regions, however, few studies have incorporated these in a multilevel-IRT framework. But, this line of work can be explored in the future.

Lastly, more work is needed to better understand how these models can be used in applications where the outcome variables are Likert-scale items of instructional practices and what modeling considerations one needs to be mindful of. Furthermore, the methodology developed in connection with this work— the ideas presented in this research can be extended to other large-scale assessments (e.g., TIMSS, TALIS) to examine key research questions in a particular country and to conduct comparisons across countries to help us improve our understanding of school systems and key factors that impact students' academic growth and schooling experiences.

7.2 Future Directions

This study makes use of non-cognitive survey items modeled via multilevel IRT models. Future work is required to better understand the application of these models in various education settings using large-scale assessments. Three avenues of future research are described below.

First, variance components play a crucial in many studies. In this study both the within-school (level-2) and between-school (level-3) variance components are of particular substantive interest, and also often influence the estimation of other parameters of interest, including school-mean exposure scores, and the magnitude of 95% intervals for the coefficients of school-level predictors. A future direction for this work includes allowing the magnitude of the within-school variance component (which captures variation in students' perceptions of exposure to practices of interest in a given school) to vary across schools; that is we can allow σ^2 to vary. This can be accomplished using MCMC. (This would be an extension of work carried out by Kasim & Raudenbush (1998) where they allow the within-school variance components to vary across schools in two-level models.) Thus we might view an effective school as one in which the school's mean STIMREAD exposure score is high, and the amount of variation across the students in their STIMREAD scores is low.

Another possibility to explore is that we might view an effective school as one in which the mean outcome latent score for a school is fairly high, and in which the slope capturing the relationship between, for example, SES (or STUDREL or ATSCHL), and the outcome construct (e.g., STIMREAD) is flat (i.e., close to a value of 0). This would tell us that not only do the STIMREAD values for the students in the school tend to be high, but they are equitably distributed with respect to student SES. We could also focus on the distribution of STIMREAD values with respect to differences across students in their STUDREL values or ATSCHL values.

Second, it is vital to make use of another data source that captures some of the key student and school variables used in this dissertation. In the Indian context, there are two possible data sources. First, there is ASER, which primarily collects information on rural children and schools, and the second is the National Assessment Study (NAS), which is conducted by the government

of India. In addition to these quantitative indicators, a focused qualitative analysis including classroom observations, document analysis of teacher logs and teaching materials, and student interviews could allow us to get a complete picture of what the situation in schools for a particular region in India looks like.

Third, on the methodological front, we can examine if the estimates of the parameters of interest obtained via fitting a three-level multilevel IRT model to the items provide us with similar estimates to those obtained by fitting a two-level model where the outcome variable is the scaled score available via PISA 2009 (e.g., the composite index of STIMREAD). These scaled scores provided by PISA are obtained via a Partial Credit Model (PCM), which can then be used as an outcome in a two-level model. Note that if we work with estimates of the student latent variables of interest created in a separate analysis—that is, the scaled scores constructed via PCM we would then be using a two-level model in which the $\widehat{\theta_{ij}}$'s are treated as outcomes in the student-level model. For example, the STIMREAD scale is created using the seven items (see Table 5.1), which is a continuous scale with an estimate for every student ranging from a minimum of -3 to a maximum of +2. This variable can be employed as the outcome variable in a 2-level model instead of working with the items directly. However, if standard errors of measurement are not available for these estimates, then the estimate we obtain for σ^2 – the within-school variance component – would reflect both measurement error connected with the estimates and actual differences across students within schools in the extent to which they report experiencing certain instructional practices of interest. The estimate we obtain for σ^2 maybe fairly large, but a substantial portion of that estimate may reflect error. The results of analyses in this disserrtation can guide secondary analysts, who make use of scaled scores as the outcome.

Appendix A

```
#####################################################################
# JAGS code for the full model including the student- and school-level predictors.
# N total students ; J total Schools; K total items ; C_k is highest category #with C_k -1) thresholds
#
#  Note: In this model specification the values for “kappa” and “alpha” are fixed to the ML
#  estimates.
 #####################################################################

model{
# Level-1 Measurement Model
  for (i in 1:N){
    for (k in 1:K){
      Y[i, k] ~ dcat(p[i, k, 1:C])
    }

    ## Cumulative probabilities for item categories
    for (k in 1:K){
      for (c in 1:(C-1)){
        logit(P[i, k, c]) <- kappa[k, c] - alpha[k]*theta[i]
      }
      P[i, k, C] <- 1.0
    }
       theta[i] ~ dnorm(mu[i], sigma2inv)

   # P: category response probabilities
   # p: cumulative response probabilities
   # Item category probabilities
    for (k in 1:K){
      p[i, k, 1] <- P[i, k, 1]
      for (c in 2:C){
        p[i, k, c] <- P[i, k, c] - P[i, k, c-1]
      }
    }
  }

# group mean Centering the student-level predictors
for(i in 1:N){
cESCS[i] <- (ESCS[i] -  ESCS.grp.mean[i])
cGender[i]<-(Gender_Female[i] - Gender.grp.mean[i])
cSTUDREL[i] <- (STUDREL[i] -  STUDREL.grp.mean[i])
cATSCHL[i]<- (ATSCHL[i] - ATSCHL.grp.mean[i])
}
```

```
#Level-2 Model: Student level
   for (i in 1:N) {
mu[i]  <- beta00[SchoolID[i]] +beta10*cESCS[i]+ beta20*cGender[i]
              + beta30*cSTUDREL[i] + beta40*cATSCHL[i]
     }


# Level-3 Model: School level  (the school-level predictors are grand-mean centered)

  for (j in 1:J) {
    beta00[j] ~ dnorm(expbeta0[j],tau00inv)
    expbeta0[j] <-  gamma00 + gamma01*(STRATIO[j] -  mean (STRATIO[]))
                                    + gamma02*(TCSHORT[j] -  meanTCSHORT[]))
                                    + gamma03*ESCS.grp.mean[j]

   }

# Prior for fixed effects:
  gamma00 ~ dnorm(0,1.0E-5)
  gamma01 ~ dnorm(0,1.0E-5)
  gamma02 ~ dnorm(0,1.0E-5)
 gamma03 ~ dnorm(0,1.0E-5)

  beta10 ~ dnorm(0, 1.0E-5)
  beta20 ~ dnorm(0, 1.0E-5)
  beta30 ~ dnorm(0, 1.0E-5)
  beta40 ~ dnorm(0, 1.0E-5)


 # Prior specification for L2 and L3 variances
  sigma2  ~ dunif(0.1, 10)
  tau00 ~ dunif(0.1, 10)

# Creating variances from precisions
  sigma2inv <- 1/sigma2
  tau00inv <- 1/tau00

}
```

Appendix B

Table B.1: Item parameter estimates (slope and thresholds) for the STIMREAD construct (*teachers' stimulation of reading engagement)* across rural and urban regions of HP and TN. A graded response model was fit to the seven items for HP and TN data separately.

Items	Parameter	Estimates for HP Rural & Urban	Estimates for TN Rural & Urban
Item 1	a1	1.07	1.13
	k1	-2.59	-2.23
	k2	-0.54	0.13
	k3	0.63	1.08
Item 2	a2	1.69	1.49
	k1	-2.58	-2.35
	k2	-1.01	-0.75
	k3	0.22	0.53
Item3	a3	1.51	1.38
	k1	-2.22	-2.11
	k2	-0.98	-0.48
	k3	0.08	0.64
Item 4	a4	1.62	1.11
	k1	-1.94	-2.15
	k2	-0.75	-0.30
	k3	0.03	0.89
Item 5	a5	1.55	1.63
	k1	-1.98	-2.15
	k2	-1.03	-0.65
	k3	0.19	0.52
Item 6	a6	1.94	1.46
	k1	-1.84	-2.05
	k2	-0.66	-0.53
	k3	0.38	0.51
Item 7	a7	1.71	1.63
	k1	-1.98	-2.03
	k2	-0.63	-0.71
	k3	0.36	0.32

Table B.2: Descriptive statistics for the outcome variable (STIMREAD construct), and student and school-level predictors for HP

Variable	N	Mean	SD	Minimum	Maximum
Outcome					
Score_STIMREAD	728	21.58	4.36	7	28
Student level					
ESCS	728	-1.41	1.02	-3.8	1.35
Gender	728	0.52	0.5	0	1
STUDREL	728	0.66	0.89	-2.9	2.45
ATSCHL	728	0.1	0.97	-2.99	2.01
School-level					
PublicSchool	58	0.84	0.37	0	1
Student-teacher Ratio (STRATIO)	58	20.97	7.94	2.3	47.6
Teacher Shortage (TCSHORT)	58	-0.51	0.89	-1.02	2.65

Table B.3: Descriptive statistics for the outcome variable (STIMREAD construct), and student and school-level predictors for TN

Variable	N	Mean	SD	Minimum	Maximum
Outcome					
Score_STIMREAD	1761	20.39	4.18	7	28
Student level					
ESCS	1761	-1.89	1.1	-4.66	1.73
Gender	1761	0.54	0.5	0	1
STUDREL	1761	0.51	1.1	-2.9	2.45
ATSCHL	1761	-0.16	0.81	-2.99	2.01
School-level					
PublicSchool	110	0.69	0.46	0	1
Student-teacher Ratio (STRATIO)	110	35.53	20.08	3.26	197.6
Teacher Shortage (TCSHORT)	110	-0.18	0.95	-1.02	2.65

Table B.4: Descriptive statistics for the outcome variable, and student and school-level predictors for the rural and urban regions across HP and TN.

	HP-rural		HP-urban		TN-rural		TN-urban	
Variable Name	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ScoreSTIMREAD	21.62	4.48	20.73	4.68	19.93	4.81	19.95	4.57
Student-level *ESCS*	-1.59	0.94	-0.69	1.02	-2.09	1.06	-1.57	1.07
Gender	0.5	0.5	0.64	0.48	0.49	0.5	0.58	0.49
STUDREL	0.67	0.88	0.58	0.94	0.54	1.13	0.4	1.08
ATSCHL	0.09	0.99	0.09	0.92	-0.19	0.83	-0.17	0.82
School-level								
PublicSchool	0.91	0.28	0.58	0.94	0.7	0.46	0.67	0.47
Student-teacher Ratio (STRATIO)	20.55	7.67	22.76	9.21	34.5	13.28	37.13	27.66
Teacher Shortage (TCSHORT)	-0.46	0.84	0.09	0.92	0.08	0.98	-0.58	0.76

Table B.5: Estimates and standard errors for grand mean and the variance components for the rural and urban regions across HP and TN for the STIMREAD construct.

Parameters of interest	HP-rural	HP-urban	TN-rural	TN-urban
	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)
Fixed effects				
Grand mean (γ_{00})	21.94 (0.39)	20.79 (0.51)	19.93 (0.19)	19.97 (0.22)
Variance components				
Within-school (σ^2)	13.79	20.51	21.79	19.80
Between-school (τ^2_{00})	5.64	1.70	1.38	1.10

Table B.6: Means, and standard errors of the coefficient for the fixed effects and the variance components for the model with the student and school-level predictors for rural and urban regions across HP and TN for the STIMREAD construct.

Parameters of interest	HP-rural	HP-urban	TN-rural	TN-urban
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Fixed effects				
Grand mean (γ_{00})	21.95 (0.39)**	20.76 (0.33)**	19.93 (0.19)**	19.98 (0.21)**
Student variables				
ESCS (β_{10})	0.012 (0.16)	-1.16 (0.37)**	-0.09 (0.16)	-0.06 (0.16)
Gender (Female) (β_{20})	-0.028 (0.34)	-1.28 (0.75)	-0.05 (0.31)	0.51 (0.52)
STUDREL (β_{30})	1.07 (0.23)**	1.67 (0.39)**	1.58 (0.17)**	1.50 (0.16)**
ATSCHL (β_{40})	0.39 (0.27)	1.09 (0.39)**	0.85 (0.19)**	0.077 (0.25)
School variables				
STRATIO (γ_{01})	0.36 (0.39)	-0.08 (0.038)	0.008 (0.014)	-0.007 (0.007)
TCSHORT (γ_{02})	-0.007 (0.04)	-1.12 (0.33)**	0.17 (0.18)	-0.42 (0.33)
Variance components				
Within-school (σ^2)	12.79	15.39	18.03	17.23
Between-school (τ_{00}^2)	5.98	0.17	1.70	1.21

Note: **p<0.01, * p<0.05

.

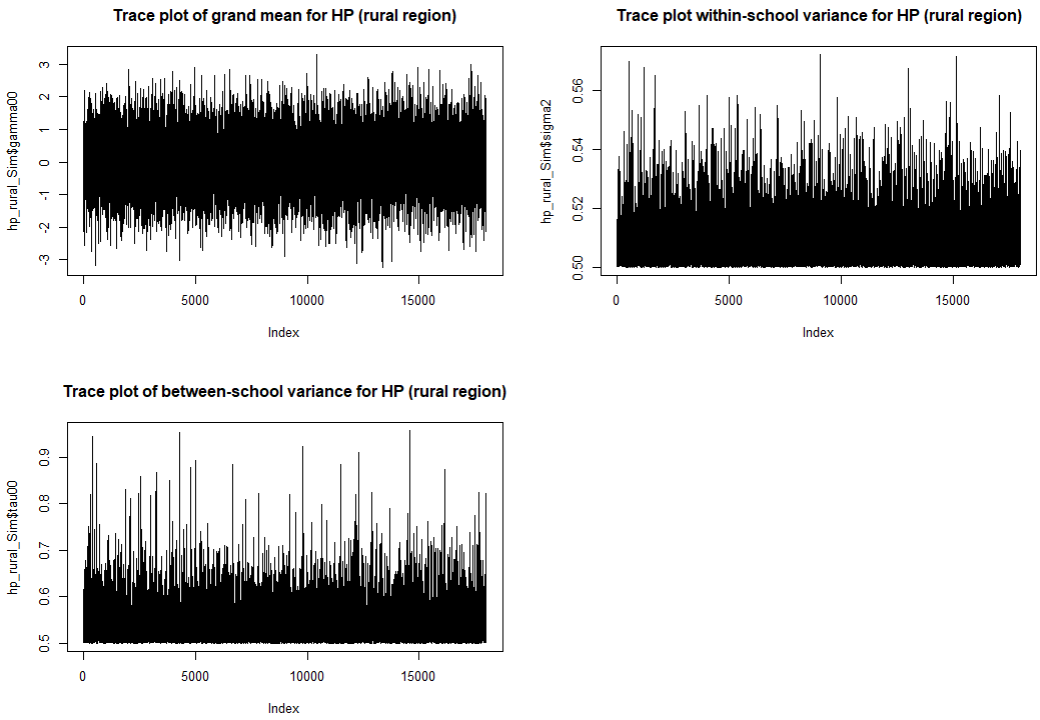


Figure B.1: Trace plots for the grand mean, within-school, and between-school variance components for HP's rural region.

Bibliography

Agrawal, T. (2014). Educational inequality in rural and urban India. *International Journal of Educational Development*, *34*, 11–19. https://doi.org/10.1016/j.ijedudev.2013.05.002

Ammermueller, A. (2005). *Educational Opportunities and the Role of Institutions* (SSRN Scholarly Paper ID 753366). Social Science Research Network. https://papers.ssrn.com/abstract=753366

Annual Status of Education Report (Rural) 2009. (2010). http://img.asercentre.org/docs/Publications/ASER%20Reports/ASER_2009/Aser2009ReportFull.pdf

Areepattamannil, S., shaljan. a@nie. edu. sg. (2014). International Note: What factors are associated with reading, mathematics, and science literacy of Indian adolescents? A multilevel examination. *Journal of Adolescence*, *37*(4), 367–372. https://doi.org/10.1016/j.adolescence.2014.02.007

Asadullah, M. N., 2,3, & Yalonetzky, G., 5. (2012). Inequality of Educational Opportunity in India: Changes Over Time and Across States. *World Development*, *40*(6), 1151–1163. https://doi.org/10.1016/j.worlddev.2011.11.008

Banerji, R., Bhattacharjea, S., & Wadhwa, W. (2013). The Annual Status of Education Report (ASER). *Research in Comparative and International Education*, *8*(3), 387–396. https://doi.org/10.2304/rcie.2013.8.3.387

BIMARU redux: NITI Aayog CEO says Bihar, Madhya Pradesh, Uttar Pradesh, Rajasthan keeping India backward. (2018, April 24). *The Financial Express*.

https://www.financialexpress.com/india-news/bimaru-redux-niti-aayog-ceo-says-bihar-madhya-pradesh-uttar-pradesh-rajasthan-keeping-india-backward/1143709/

Borman, G. D., & Dowling, M. (2010). Schools and Inequality: A Multilevel Analysis of Coleman's Equality of Educational Opportunity Data. *Teachers College Record*, *112*(5), 1201–1246.

Braeken, J., & Blömeke, S. (2016). Comparing future teachers' beliefs across countries: Approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item functioning. *Assessment & Evaluation in Higher Education*, *41*(5), 733–749. https://doi.org/10.1080/02602938.2016.1161005

Burstein, L. (1980). Chapter 4: The Analysis of Multilevel Data in Educational Research and Evaluation. *Review of Research in Education*, *8*(1), 158–233. https://doi.org/10.3102/0091732X008001158

Burstein, L. (1989). *Conceptual Considerations in Instructionally Sensitive Assessment*. https://eric.ed.gov/?id=ED341737

Cardichon, J., Darling-Hammond, Linda, L., Yang, M., Scott, C., Shields, P. M., & Burns, D. (2020). *Inequitable Opportunity to Learn: Student Access to Certified and Experienced Teachers* (p. 36). Palo Alto, CA: Learning Policy Institute.

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

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553–1623. https://doi.org/10.1093/qje/qju022

Chudgar, A., & Quin, E. (2012). Relationship between private schooling and achievement: Results from rural and urban India. *Economics of Education Review*, *31*(4), 376–390. https://doi.org/10.1016/j.econedurev.2011.12.003

Correnti, R., & Martínez, J. F. (2012). *Conceptual, Methodological, and Policy Issues in the Study of Teaching: Implications for Improving Instructional Practice at Scale*. https://doi.org/10.1080/10627197.2012.717834

Das, J., & Zajonc, T. (2010a). India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement. *Journal of Development Economics*, *92*(2), 175–187. https://doi.org/10.1016/j.jdeveco.2009.03.004

Das, J., & Zajonc, T. (2010b). India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement. *Journal of Development Economics*, *92*(2), 175–187. https://doi.org/10.1016/j.jdeveco.2009.03.004

Drèze, J., & Sen, A. (2002). *India: Development and Participation*. Oxford University Press.

Embretson, S. E., & Reise, S. Paul. (2000). *Item response theory for psychologists*. L. Erlbaum Associates; /z-wcorg/.

Fox, G. J. A., & Glas, C. A. W. (2016). Multilevel Response Models with Covariates and Multiple Groups. In W. J. Van Der Linden (Ed.), *Handbook of Item Response Theory, Volume One:*

Models (pp. 407–419). CRC Press. https://research.utwente.nl/en/publications/multilevel-response-models-with-covariates-and-multiple-groups

Fox, J. P. (2010). *Bayesian Item Response Modeling*. New York: Springer. https://link.springer.com/book/10.1007%2F978-1-4419-0742-4

Fox, J.-P. (2005). Multilevel IRT model assessment. *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*, 227–52.

Gamboa, L. F., & Waltenberg, F. D. (2015). Measuring Inequality of Opportunity in Education by Combining Information on Coverage and Achievement in PISA. *Educational Assessment*, *20*(4), 320–337.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). CRC press Boca Raton, FL.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. JSTOR.

Govinda, R., & Bandyopadhyay, M. (2008, July). *Access to Elementary Education in India* [Reports and working papers]. http://www.create-rpc.org/pdf_documents/India_CAR.pdf

Govinda, R., Varghese, N. V., Carron, Gabriel., International Institute for Educational Planning., & National Institute of Education Planning and Administration (India). (1993). *Quality of primary schooling in India: A case study of Madhya Pradesh*. International Institute for Educational Planning, UNESCO ; NIEPA, National Institute of Educational Planning and Administration; /z-wcorg/.

Hill, S., & Chalaux, T. (2011). *Improving Access and Quality in the Indian Education System* [OECD Economics Department Working Papers]. Organisation for Economic Co-operation and Development. http://www.oecd-ilibrary.org/content/workingpaper/5kg83k687ng7-en

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning* (pp. xv, 453). Lawrence Erlbaum Associates, Inc.

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project*. Bill & Melinda Gates Foundation. http://eric.ed.gov/?id=ED540960

Kingdon, G. G. (2007). The progress of school education in India. *Oxford Review of Economic Policy*, *23*(2), 168–195. https://doi.org/10.1093/oxrep/grm015

Kreiner, S., & Christensen, K. B. (2014). Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, *79*(2), 210–231. https://doi.org/10.1007/s11336-013-9347-z

Lewin, K. M. (2011). Expanding access to secondary education: Can India catch up? *International Journal of Educational Development*, *31*(4), 382–393. https://doi.org/10.1016/j.ijedudev.2011.01.007

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. L. Erlbaum Associates. http://www.gbv.de/dms/hbz/toc/ht000250483.pdf

Maclean, R. (2003). Equality of Opportunity in Education. In J. P. Keeves, R. Watanabe, R. Maclean, P. D. Renshaw, C. N. Power, R. Baker, S. Gopinathan, H. W. Kam, Y. C. Cheng,

& A. C. Tuijnman (Eds.), *International Handbook of Educational Research in the Asia-Pacific Region: Part One* (pp. 143–154). Springer Netherlands. https://doi.org/10.1007/978-94-017-3368-7_10

Maclean, R., & Vine, K. (2003). A Case Study of Learning Achievement in South Asia. In J. P. Keeves, R. Watanabe, R. Maclean, P. D. Renshaw, C. N. Power, R. Baker, S. Gopinathan, H. W. Kam, Y. C. Cheng, & A. C. Tuijnman (Eds.), *International Handbook of Educational Research in the Asia-Pacific Region: Part One* (pp. 143–154). Springer Netherlands. https://doi.org/10.1007/978-94-017-3368-7_10

Masri, Y. H. E., Baird, J.-A., & Graesser, A. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, *23*(4), 427–455. https://doi.org/10.1080/0969594X.2016.1218323

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation*, *12*(1), 53–74. https://doi.org/10.1080/13803610500392236

Muthén, B., & Asparouhov, T. (2018). Recent Methods for the Study of Measurement Invariance With Many Groups: Alignment and Random Effects. *Sociological Methods & Research*, *47*(4), 637–664. https://doi.org/10.1177/0049124117701488

MUTHÉN, B. O. (1994). Multilevel Covariance Structure Analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Muthén, L. K., & Muthén, B. O. (2017). Mplus user's guide (version 8) Los Angeles. *Los Angeles CA: Muthén & Muthén*.

National Academies of Sciences, E., and Medicine. (2019). *Monitoring Educational Equity*. The National Academies Press. https://doi.org/10.17226/25389

National Center for Education Statistics (DHEW), W., DC., General Research Corp., M., VA., & Killalea Associates, Inc., Arlington, VA. (1978). *Educational Opportunity. The Concept, Its Measurement, and Application. Highlights. Sponsored Reports Series*.

Oakes, J. (1989). What Educational Indicators? The Case for Assessing the School Context. *Educational Evaluation and Policy Analysis*, *11*(2), 181–199. https://doi.org/10.3102/01623737011002181

OECD. (2010). *TALIS 2008 Technical Report*. OECD. https://doi.org/10.1787/9789264079861-en

OECD. (2012). *PISA 2009 Technical Report*. OECD Publishing. https://doi.org/10.1787/9789264167872-en

Oliveri, M., Olson, BrentF., Ercikan, K., & Zumbo, BrunoD. (2012). Methodologies for Investigating Item- and Test-Level Measurement Equivalence in International Large-Scale Assessments. *International Journal of Testing*, *12*(3), 203–223. https://doi.org/10.1080/15305058.2011.617475

O'Sullivan, M. (2006). Lesson observation and quality in primary education as contextual teaching and learning processes. *International Journal of Educational Development*, *26*(3), 246–260. https://doi.org/10.1016/j.ijedudev.2005.07.016

Peske, H. G., & Haycock, K. (2006). *Teaching Inequality: How Poor and Minority Students Are Shortchanged on Teacher Quality: A Report and Recommendations by the Education Trust*. Education Trust. https://eric.ed.gov/?id=ED494820

Phillips, M., & Chin, T. (2004). School Inequality: In K. M. Neckerman (Ed.), *Social Inequality* (pp. 467–520). Russell Sage Foundation. http://www.jstor.org/stable/10.7758/9781610444200.17

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher*, *38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Plummer, M. (2011). *JAGS: a program for the statistical analysis of Bayesian hierarchical models by Markov Chain Monte Carlo*.

Pokropek, A., Borgonovi, F., & McCormick, C. (2017). On the Cross-Country Comparability of Indicators of Socioeconomic Resources in PISA. *Applied Measurement in Education*, *30*(4), 243–258. https://doi.org/10.1080/08957347.2017.1353985

R Core Team. (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. http://www.R-project.org/.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical Inference When Classroom Quality is Measured With Error. *Journal of Research on Educational Effectiveness*, *1*(2), 138–154. https://doi.org/10.1080/19345740801982104

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Rutkowski, L., & Rutkowski, D. (2016). A Call for a More Measured Approach to Reporting and Interpreting PISA Results. *Educational Researcher*, *45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100–100.

Schleicher, A. (2009). Securing quality and equity in education: Lessons from PISA. *PROSPECTS*, *39*(3), 251–263. https://doi.org/10.1007/s11125-009-9126-x

Schweig, J. (2014). Cross-Level Measurement Invariance in School and Classroom Environment Surveys: Implications for Policy and Practice. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/0162373713509880

Shavelson, R. J., McDonnell, L. M., & Oakes, J. (1989). *Indicators for Monitoring Mathematics and Science Education* [Product Page]. https://www.rand.org/pubs/reports/R3742.html

Shields, L., Newman, A., & Satz, D. (2017). Equality of Educational Opportunity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/equal-ed-opportunity/

Singh, R., & Sarkar, S. (2015). Does teaching quality matter? Students learning outcome related to teaching quality in public and private primary schools in India. *International Journal of Educational Development*, *41*, 153–163. https://doi.org/10.1016/j.ijedudev.2015.02.009

Venkatachalam, K. S. (2017). Why Does India Refuse to Participate in Global Education Rankings? *The Diplomat*. Retrieved from https://thediplomat.com/2017/01/why-does-india-refuse-to-participate-in-global-education-rankings/

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383–401. https://doi.org/10.1111/j.2044-8317.2012.02059.x

Wang, X., Bradlow, E. T., Wainer, H., & Muller, E. S. (2008). A Bayesian Method for Studying DIF: A Cautionary Tale Filled With Surprises and Delights: *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998607306080

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-Improved Wald Test for DIF Testing With Multiple Groups: Evaluation and Comparison to Two-Group IRT. *Educational and Psychological Measurement*, *73*(3), 532–547. https://doi.org/10.1177/0013164412464875