# Applying the Pseudo-Panel Approach to International Large-Scale Assessments: A Methodology for Analyzing Subpopulation Trend Data

Author: Martin Hooper

**Boston College**

**Lynch School of Education**

**Department of**

**Measurement, Evaluation, Statistics, and Assessment**

# Applying the Pseudo-Panel Approach to International Large-Scale Assessments: A Methodology for Analyzing Subpopulation Trend Data

**Dissertation**

**by**

**Martin Hooper**

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

**August 2017**

# Abstract


**Applying the Pseudo-Panel Approach to International Large-Scale Assessments: A Methodology for Analyzing Subpopulation Trend Data**

**Dissertation by Martin Hooper**

**Advisor: Ina V. S. Mullis, Ph.D.**

TIMSS and PIRLS assess representative samples of students at regular intervals, measuring trends in student achievement and student contexts for learning. Because individual students are not tracked over time, analysis of international large-scale assessment data is usually conducted cross-sectionally. Gustafsson (2007) proposed examining the data longitudinally by analyzing relationships between country-level trends in background constructs and trends in student achievement. Through longitudinal analysis of international large-scale assessment data, it becomes possible to mitigate some of the confounding factors in the analysis.

This dissertation extends this country-level approach to subpopulations within countries. Adapting a pseudo-panel approach from the econometrics literature (Deaton, 1985), the proposed approach creates subpopulations by grouping students based on demographic characteristics, such as gender or parental education. Following grouping, the subpopulations with the same demographic characteristics are linked across cycles and the aggregated subpopulation means are treated as panel data and analyzed through longitudinal data analysis techniques. As demonstrated herein the primary advantages of the subpopulation approach are that it allows for

analysis of subgroup differences, and it captures within-country relationships in the data that are not possible to analyze at country level.

Illustrative analysis examines the relationship between early literacy activities and PIRLS reading achievement using PIRLS 2001 and PIRLS 2011 data. Results from the subpopulation approach are compared with student-level and country-level cross-sectional results as well as country-level longitudinal results. In addition, within-country analysis examines the subpopulation-level relationship between early literacy activities and PIRLS reading achievement, multiple group analysis compares regression coefficient estimates between boys and girls and across parental education subgroups, and mediation analysis examines the extent that partaking in early literacy activities can explain differences between boys and girls in PIRLS reading achievement.

# Acknowledgements

Before coming across the pseudo-panel literature, this dissertation began with the seemingly unconventional idea that it was possible to analyze international large-scale assessment data longitudinally through the use of aggregated subgroup means. From this early stage, the members of my committee were very helpful in cultivating and forming this idea into what has become this dissertation.

I would especially like to thank my dissertation advisor Dr. Ina V. S. Mullis for always believing in me and supporting me, and my readers Dr. Michael O. Martin, Dr. Henry Braun, and Dr. Jan-Eric Gustafsson. Dr. Martin provided regular feedback during key stages of dissertation development, including fielding calls over the Christmas holiday. Dr. Braun led my doctoral cohort's dissertation seminar, and provided helpful ideas from the early stages of the process. Dr. Gustafsson was generous from afar with his time and expertise, explaining new methods of differences-in-differences and providing insightful correspondence on important issues related to this dissertation.

Truth be told, I could not have gotten through the rigors the ERME (MESA) program without the unwavering support of my wife, Yeny Pardini Gonzalez. Yeny has always encouraged me to dream big and she supported me immensely during these five years of balancing full-time employment and doctoral studies.

To my dear lil' Chloe Celine—that one day in the 2030's or 2040's you find this acknowledgements page when you google your dad, and you learn that having more time to spend with you inspired the early mornings and late nights needed to complete this dissertation.

Thank you to my parents, for always prioritizing my education and encouraging me to follow my own path—whether it be travelling the world in my 20's or starting a doc program in my 30's.

Also, I would like to thank my Massachusetts family—the Maury-Nolan family (Dan, Beth, Ruby, Kiki, and Penny) for making us feel at home in Boston and Aunt Kathleen for her visits.

I am also grateful for the support of friends/colleagues at the TIMSS & PIRLS International Study Center including Victoria Centurino, Paul Connolly, Kerry Cotter, Christine Hoage, and Kathleen Holland. I would especially like to thank my questionnaire teammate, Bethany Fishbein, who was always willing to lend a hand at work when I was pressed to make my dissertation deadlines.

Thank you to all of the professors of the ERME (MESA) department.

I would also like to thank Trude Nielson and Rolf Strietholt, who in the early stages of developing this dissertation helped me catch up on the differences-in-differences literature as applied to international large-scale assessment data.

# Table of Contents

Chapter 1: Introduction ..... 1
Chapter 2: Theoretical Background ..... 8
2.1 Longitudinal Analysis and Causality ..... 10
Cross-Sectional Analysis and International Large-Scale Assessments ..... 12
Randomized Control Trials ..... 14
Longitudinal Designs ..... 16
2.2 Technical Details on Techniques for Longitudinal Data Analysis ..... 19
Structural Equation Modeling Approach ..... 24
2.3 PIRLS as a Longitudinal Study ..... 28
2.4 Difference-in-Differences ..... 30
Theoretical Background of Difference-in-Differences ..... 31
Difference-in-Differences Applied to International Large-Scale Assessments ..... 33
Argument for Gustafsson's (2007) Difference-in-Differences Approach ..... 35
Common Trends and Difference-in-Differences ..... 38
Change Over Time in the Explanatory Variables ..... 40
Issues Related to Aggregation ..... 41
Structural Equation Modeling Approach to Difference-in-Differences ..... 45
2.5 Proposed Subpopulation Approach ..... 46
Econometrics Theory behind the Subpopulation Approach ..... 48
Implementing the Subpopulation Approach ..... 49
Pseudo-Panel Analysis Applied to International Large-Scale Assessment Data ..... 52
Proposed Methodology ..... 54
Chapter 3: Analysis Methodology ..... 57
3.1 Description of the PIRLS Assessment and Database ..... 57
PIRLS International Database ..... 59
3.2 Preparing Data for Analysis ..... 59
Reading Achievement ..... 60
Early Literacy Activities ..... 61
Time-Varying Covariates ..... 63
Demographic Variables Used for Creating Subpopulations and as Covariates ..... 66

Weighting ....68
Missing Data ....69
Significance Testing ....71
3.3 Analysis Overview ....72
Phase 1: Multilevel Cross-Sectional Analysis of Each Cycle of PIRLS Data ....73
Phase 2: Country Cross-Sectional Analysis of Each Cycle of PIRLS Data ....75
Phase 3: Country Difference-in-Differences Analysis ....76
Phase 4: Subpopulation Difference-in-Differences Analysis ....77
Phase 5: Analysis of Within-Country Relationships ....81
Phase 6: Comparisons of Fixed-Effects Coefficient Estimates across Groups ....84
Phase 7: Mediation Analysis through a Random-Effects Model ....88
Chapter 4: Results of Analysis ....98
4.1 Phase 1: Multilevel Cross-Sectional Analysis of Each Cycle of PIRLS Data ....98
4.2 Phase 2: Country Cross-Sectional Analysis on Each Cycle of PIRLS Data ....101
4.3 Phase 3: Country Difference-in-Differences Analysis ....107
4.4 Phase 4: Subpopulation Difference-in-Differences Analysis ....111
4.5 Phase 5: Analysis of Within-Country Relationships ....115
4.6 Phase 6: Comparisons of Fixed-Effects Coefficient Estimates across Groups ....127
4.7 Phase 7: Mediation Analysis through a Random-Effects Model ....135
Chapter 5: Discussion ....140
5.1 Comparing Cross-Sectional and Longitudinal Approaches across Levels of Aggregation ....141
5.2 Capturing Within-Country Relationships ....145
5.3 Examining Differential Relationships across Subgroups ....148
5.4 Mediation Modeling Applications ....150
5.5 Difference-in-Differences and Causal Inference ....151
5.6 Limitations and Future Research ....154
5.7 Conclusions ....156
References ....158
Appendix A: Additional Analysis Details ....165

# List of Tables

Table 3.1: Countries and Provinces Included in the Analysis .....60
Table 3.2: Items Measuring Early Literacy Activities.....62
Table 3.3: Recoding of Reports on Duration of Preprimary Attendance .....64
Table 3.4: Items Measuring Parent Like Reading. .....66
Table 3.5: Amount of Missing Data for Each Explanatory Variable .....69
Table 3.6: Explanatory Variables Included in Each Phase 1 Model.....75
Table 4.1: Results from the Multilevel Analysis Conducted in Phase 1 .....99
Table 4.2: Parameter estimates for Phase 2 Country-Level Cross-Sectional Analysis .....104
Table 4.3: Evaluating the Effect of Outliers on Model 7 Early Literacy Activities Coefficient Estimates.....106
Table 4.4: Parameter estimates for Phase 3 Country-Level Difference-in-Differences .....110
Table 4.5: Evaluating the Effect of Outliers on Model 9 Early Literacy Activities Coefficient Estimates.....111
Table 4.6: Parameter estimates for the Phase 4 Subpopulation Approach .....113

# List of Figures

Figure 2.1: Excerpt from PIRLS 2011 International Results in Reading, Exhibit 4.6: Early Literacy Activities Before Beginning Primary School ....13

Figure 2.2: Path Model for Fixed-Effects Approach ....25

Figure 2.3: Path model for the Random-Effects Approach ....26

Figure 2.4: Path Model for the Random-Effects Approach with Observed Time-Invariant Covariate ....28

Figure 2.5: Excerpt from PIRLS 2011 International Results in Reading, Exhibit 1.4: Trends in Reading Achievement ....29

Figure 2.6: Excerpt from PIRLS 2006 International Report, Exhibit 3.1: Index of Early Home Literacy Activities (EHLA) with Trends ....30

Figure 2.7: Illustration of the Common Trends Assumption for Difference-in-Differences ....31

Figure 2.8: Trend Lines in Reading Achievement across PIRLS 2001, 2006, and 2011 ....39

Figure 2.9: Excerpt from PIRLS 2011 International Results in Reading, Exhibit 1.7: Trends in Reading Achievement by Gender ....47

Figure 3.1: Classification Rules for Creating Subpopulations within Each Country for Each Cycle ....77

Figure 3.2: Path Model for the Subpopulation Fixed-Effects Analysis ....80

Figure 3.3: Null Model for Multiple Group Analysis Where Regression Coefficients are Constrained to be Equal across Boys and Girls ....86

Figure 3.4: Alternative Model for Multiple Group Analysis Where Regression Coefficients Vary Across Boys and Girls ....87

Figure 3.5 Conceptual Diagram Illustrating Mediation Modeling ....89

Figure 3.6: Path Model for the Fixed-Effects Approach Highlighting the Covariance ....92

Figure 3.7: Path Model Examining the Relationship Between Gender and PIRLS Reading Achievement ....94

Figure 3.8: Path Model for Predicting PIRLS Reading Achievement by Gender and Early Literacy Activities ....94

Figure 3.9: Path Model for Analyzing the Mediation Effect ....96

Figure 4.1: Graphs Showing the Country-Level Relationship between Mean Early Literacy Activities and Mean Reading Achievement for Each PIRLS Cycle ....103

Figure 4.2: Scatterplot Displaying Country-Level Changes in Average Early Literacy Activities and Average Reading Achievement Between PIRLS 2001 and PIRLS 2011 ....108

Figure 4.3: Scatterplot Showing Country-Level Changes in Average Early Literacy Activities and Average Reading Achievement Between PIRLS 2001 and PIRLS 2011 (Divided into Quadrants) .....109

Figure 4.4: Subpopulation Relationship Between Changes in Average Early Literacy Activities Scores and Changes in Average Reading Achievement .....112

Figure 4.5: Path Model with the Subpopulation Fixed-Effects Model Results (Model 10) .....114

Figure 4.6: Subpopulation Relationship Between Changes in Average Early Literacy Activities Scores and Changes in Average Reading Achievement (Color Coded) .....117

Figure 4.7: Highlighting the Five Countries with the Largest Standard Deviation Between the Subpopulations in Changes in Average Early Literacy Activities Scores .....118

Figure 4.8: Highlighting the Five Countries with the Smallest Standard Deviation Between the Subpopulations in Changes in Average Early Literacy Activities Scores .....119

Figure 4.9: Relationship Between Changes in Average PIRLS Reading Achievement and Changes in Average Early Literacy Activities Scores for Countries with Coefficients over 0.3 .....122

Figure 4.10: Relationship Between Changes in Average PIRLS Reading Achievement and Changes in Average Early Literacy Activities Scores for Countries with Coefficients Between 0 and 0.3 .....123

Figure 4.11: Relationship Between Changes in Average PIRLS Reading Achievement and Changes in Average Early Literacy Activities Scores for Countries with Negative Coefficients .....124

Figure 4.12: Subpopulation-Level Relationship Between Country-Centered Changes in PIRLS Reading Achievement and Country-Centered Changes in Early Literacy Activities Scores .....126

Figure 4.13: Relationship Between Changes in Average Early Literacy Activities and Changes in Average Reading Achievement by Gender of Subpopulation .....128

Figure 4.14: Relationship Between Changes in Average Early Literacy Activities and Changes in Average Reading Achievement for Highest Parental Education Groups .....132

Figure 4.15: Relationship Between Changes in Average Early Literacy Activities and Changes in Average Reading Achievement for the No High School Parental Education Group .....133

Figure 4.16: Relationship Between Changes in Average Early Literacy Activities and Changes in Average Reading Achievement for the College Parental Education Group .....134

Figure 4.17: Path Model with the Subpopulation Random-Effects Model Results .....136

Figure 4.18: Path Model Representing the Subpopulation Relationship Between Gender and PIRLS Reading Achievement .....137

Figure 4.19: Path Model Representing the Subpopulation Relationship Between PIRLS Reading Achievement and the Explanatory Variables Gender and Early Literacy Activities .....138

Figure 4.20: Path Model Representing Early Literacy Activities Mediating Relationship between Gender and PIRLS Reading Achievement .....139

# Chapter 1: Introduction

In the summer of 2016, the National Academy of Education convened two workshops focusing on future opportunities and challenges for international large-scale assessments. At the workshops, a number of researchers highlighted important opportunities available for analysts to take better advantage of the longitudinal nature of the repeated cross-sectional design of these assessments (Chmielewski & Dhuey, 2017; Gustafsson, 2016; Rutkowski, 2016a).

With an eye to optimizing the soundness of inferences from international large-scale assessment results, this dissertation proposes and illustrates a new subpopulation approach for examining international large-scale assessment data longitudinally. Adapting an econometrics pseudo-panel methodology first developed by Nobel Laureate Angus Deaton (1985) to the educational assessment context, the new subpopulation approach provides the opportunity for complex analysis of subgroup differences using trend data.

In the context of economic household surveys, Deaton (1985) noticed that in many countries individual longitudinal data were non-existent, but there was an ample supply of repeated cross-sectional survey data. In his seminal paper, Deaton argued that samples from survey cross-sections could be divided in subpopulations by time-invariant characteristics such as demographic variables, and the aggregated means on the variables of interest for these subpopulations could be treated as individual data and analyzed through panel analysis approaches. Over the past 30 years, this methodology has been implemented widely in economics and other disciplines.

Like the household survey context, in international assessments like TIMSS, PIRLS, and PISA there is an abundance of data collected cross-sectionally across repeated waves, but given the complexities and costs of tracking individuals over time, longitudinal data at the student level is virtually non-existent. Typically, researchers analyzing large-scale assessment data only analyze one cross-section—providing a snapshot of the relationships in the data at the time of testing. However, if it were possible to treat subpopulations as individuals and implement longitudinal analysis approaches, the analysis would be able to reap the benefits of the longitudinal data structure—namely, the mitigation of some of the confounding factors in the data.

The proposed subpopulation methodology extends the country difference-in-differences approach proposed by Jan-Eric Gustafsson (2007). Given that TIMSS and PIRLS measure trends at country-level for both achievement scales and background data, Gustafsson (2007) proposed aggregating data for explanatory and outcome variables to country-level and with this aggregated dataset, pooled across countries and cycles, applying a fixed-effects regression model. This regression model would thereby estimate the country-level relationships between changes across assessment cycles in the background variable (X) and the achievement variable (Y). Gustafsson's approach built upon a study by Hanushek and Wößmann (2006), where the authors implemented a similar approach to analyze the effect of tracking on educational inequity across countries.

By pooling data across countries and examining change across cycles, Gustafsson (2007) contended that this longitudinal approach was able to strengthen causal interpretations drawn from analysis of international large-scale assessment data. Following Gustafsson's (2007) and Hanushek and Wößmann's (2006) applications of this approach to international large-scale

assessment data, numerous papers have applied the approach through the country-level analysis technique (Gustafsson, 2013; Gustafsson & Nilsen, 2016; Hanushek, Link, & Wößmannn, 2013; Liu, Bellens, Gielen, Van Damme, & Onghena, 2014; Rosén & Gustafsson, 2014; Rosén & Gustafsson, 2016).

In his initial paper on this approach, Gustafsson (2007, p. 61) noted that this difference-in-differences methodology may also be used with subpopulation data:

> Analysis of trend data can be extended in many other interesting ways. The total sample of students, for example, could be broken into results for different subgroups, such as gender, language spoken at home, and socioeconomic background.

Many educational equity research questions examine subgroup differences, such as differences in achievement between boys and girls and differences in achievement between socioeconomic groups. Such research questions are difficult to evaluate with country-level approaches since between-group differences are lost in the aggregation process. For these questions, it is necessary to extend the difference-in-differences approach to analyze subpopulation data.

Applying Deaton's (1985) pseudo-panel methodology to the international large-scale assessment context allows for longitudinal analysis, such as difference-in-differences analysis, at subpopulation level. Questionnaires from international large-scale assessments collect data on student demographic characteristics, and these demographic data can be used to classify students into subpopulations for each cycle. After classifying students into subpopulations, the student data on the background variables (X) of interest and the outcome variable (Y) can be aggregated to subpopulation level, and subpopulation data can be paired across cycles on their demographic characteristics. After linking the data across cycles, the data can be treated a pseudo-longitudinal

data—with each subpopulation having a score for each cycle on the background variable and outcome variable.

The new approach has two primary advantages:

(1) The subpopulation approach provides an option for longitudinal data analysis at lower levels of aggregation than the country-level difference-in-differences approach, making it possible to incorporate a level of important subgroup relationships into the analysis; and

(2) The subpopulation approach provides an opportunity for modeling subgroup differences longitudinally, such as analysis of differences in fixed-effects coefficients across subgroups or mediation analysis.

To illustrate this new approach, analysis was conducted to examine the relationship between early literacy activities and PIRLS reading achievement using PIRLS 2001 and PIRLS 2011 data. PIRLS (Progress in International Reading Literacy Study) is the worldwide standard for assessing reading achievement at the fourth grade, and was first administered in 2001 and since then it has been conducted every five years—2001, 2006, 2011, and 2016, with the PIRLS 2016 results to be released in December 2017. PIRLS works well for this example analysis because it follows a repeated cross-sectional design and is designed for measuring trends in both student achievement and contextual data.

Accompanying the PIRLS reading assessment are a set of questionnaires that provide information on each students' context for learning. Students complete a student questionnaire that provides key demographic information as well as information on their experiences in school

and their attitudes toward reading, and their parents complete a home questionnaire providing retrospective reports on the students' early childhood experiences as well as other information on the home context for learning, including additional demographic information. Teachers and school principals also complete questionnaires providing information on the school and classroom contexts for learning.

Using demographic characteristics on parental education and student sex collected through the PIRLS questionnaires in PIRLS 2001 and PIRLS 2011, six subpopulations were created within each country for each cycle. Within each subpopulation, early literacy activities data and reading achievement data were aggregated to subpopulation level producing an aggregated mean score on both variables, and then the subpopulations were paired over time allowing for longitudinal analysis.

Analysis was conducted across seven phases. The purpose of the first three phases is to create a baseline for comparisons with the subpopulation approach in Phase 4. The Phase 1 analysis examines the relationship between early literacy activities and PIRLS reading achievement through cross-sectional analyses of student-level data pooled across countries, and the Phase 2 analysis examines the relationship between early literacy activities and PIRLS reading achievement through analysis of cross-sectional data aggregated to country-level. In Phase 3, the relationship between countries' early literacy activities averages and PIRLS reading achievement averages was estimated using Gustafsson's (2007) country-level difference-in-differences approach, and in Phase 4 differences-in-differences analysis was conducted with subpopulation data.

Phase 5 analyzes subpopulation variation within countries, and looks at whether evidence of analysis-relevant relationships can be found by analyzing subpopulation-level variability alone. The analyses graphically explore the relationship between early literacy activities and PIRLS reading achievement among the six subpopulation units within each of the 21 countries. In another analysis, the subpopulation-level variance is decomposed from the between-country variance and using only subpopulation-level variance the relationship between changes in early literacy activities scores and PIRLS reading achievement is examined across the 21 countries.

Phase 6 and Phase 7 demonstrate more complex applications of the subpopulation approach. Phase 6 illustrates the potential for testing whether regression coefficients are the same across subpopulations, in this case comparing regression coefficients across gender and highest parental education groups, and Phase 7 demonstrates the use of the subpopulation approach in mediation analysis to explain the gender achievement gaps.

As the objective of this dissertation is to demonstrate the subpopulation methodology, the primary role of the example analysis is to illustrate the proposed approach. As such, Chapter 2 outlines the theoretical background for applying the subpopulation approach to international large-scale assessment data. The chapter begins by first describing why longitudinal analysis should provide more sound inferences than cross-sectional approaches, and then the chapter provides the technical details for a number of longitudinal analysis approaches relevant to difference-in-differences analysis and the current applications of the subpopulation approach. With this background at hand, the chapter then examines Gustafsson's (2007) difference-in-differences analysis and describes the subpopulation approach proposed herein.

Chapter 3 details the illustrative analysis. The chapter provides an overview of the PIRLS assessment and the PIRLS variables included in the analysis, and provides a short summary of the theoretical background for the example analysis on the relationship between early literacy activities and reading achievement. The chapter also describes the seven phases of the analysis, including the procedure for creating the subpopulations.

Chapter 4 provides the results of each of the seven phases of the analysis, and Chapter 5 discusses the lessons learned from the analysis in light of the theoretical background presented in Chapter 2.

# Chapter 2: Theoretical Background

This dissertation develops a new approach for examining international large-scale assessment data from a longitudinal perspective by adapting Deaton's (1985) pseudo-panel approach. In so doing, the dissertation expands upon the difference-in-differences framework for conducting longitudinal analysis on international large-scale assessment data put forward by Gustafsson (2007).

This dissertation contributes to the burgeoning literature on strengthening the causal argument for analysis of international large-scale assessment data (Robinson, 2014; Rutkowski, 2016b; Rutkowski and Delandshere, 2016; Schlotter, Scherdt, & Wößmannn, 2014; Strietholt, Gustafsson, Rosén, & Bos, 2014). Causal inference has long been a goal for international large-scale assessments. The founders of the IEA first proposed international large-scale assessments as a way to identify policies and practices that can be implemented to improve educational achievement (Husén, 1967). It was believed that by examining inputs and outputs across countries and cultures, it would be possible to view the world as an educational laboratory, allowing "comparisons to be made with means more powerful and more sure than artificially set up and costly experimental situations within one country or culture (Husén, 1967, pp. 27-28)."

The largest international large-scale assessments—TIMSS, PIRLS, and PISA, follow a repeated cross-sectional design, assessing representative samples in a country at regular intervals (4 years for TIMSS, 5 years for PIRLS, and 3 years for PISA). In such a design, each participant is only sampled once and there is no random assignment to treatment and control groups. Without random assignment, it is difficult to justify causal claims.

Because a repeated cross-sectional design is not ideal for causal interpretation, often complicated statistical techniques need to be employed to strengthen causal claims, such as propensity score techniques, regression discontinuity designs, or instrumental variable approaches. A more straightforward way to aid causal interpretation is through difference-in-differences analysis, an econometrics technique that takes advantage of the longitudinal structure of the design and allows for causal inference when certain assumptions are fulfilled (Angrist & Pischke, 2009, 2015; Woolbridge, 2010). Gustafsson (2007) and Hanushek and Wößmann (2006) have made inroads into conducting difference-in-differences analysis at country-level using international large-scale assessment data.

Compared to cross-sectional approaches, the advantage of Gustafsson's (2007) and Hanushek and Wößmann's (2006) difference-in-differences approach and the subpopulation extension proposed herein is based upon the idea that through longitudinal data analysis it is possible to mitigate some of the threats to causal inference. To understand the contribution of the subpopulation approach to analysis of international large-scale assessment necessitates foundational knowledge in three areas:

1. Longitudinal data analysis and causal inference;
2. Longitudinal data analysis techniques; and
3. Difference-in-differences analysis.

Because this dissertation does not assume the reader to have this foundation, Section 2.1 provides an overview of the importance of longitudinal data analysis for causal inference, Section 2.2 provides the technical details for relevant longitudinal data analysis techniques,

Section 2.3 examines how PIRLS can be viewed as a country-level longitudinal study, and Section 2.4 explains difference-in-differences analysis and its application to international large-scale assessment data. With this context at hand, Section 2.5 details the new subpopulation approach.

As explained in the Chapter 1, this dissertation is methodologically-focused, and the purpose for including the example analysis is to illustrate the methodology proposed herein. As such, this background chapter provides the theoretical foundation for the method, and the methodological chapter that follows in Chapter 3 outlines the operational methodology behind the example analysis and includes a brief summary of the substantive theory on the relationship between early literacy activities and PIRLS reading achievement. Although the relationship between early childhood education and reading achievement is not the focus of this chapter, when appropriate this chapter does use research on early childhood education within the examples provided.

## 2.1 Longitudinal Analysis and Causality

At the heart of policy-relevant educational research, including research through international large-scale assessments, is the goal of improving educational outcomes. One of the primary ways that research can contribute to improving outcomes is by identifying policies and practices that work and recommending the implementation of these policies or practices. The challenge is that the claim that a policy or practice works is inherently causal, and therefore the claim is strengthened by the quantity and quality of causal evidence available to support it.

There is great debate within the research community about what constitutes causal evidence (Pearl, 2000; Holland, 1986;  Rubin, 1974; Shadish, Cook, & Campbell, 2002), and

there are numerous frameworks to evaluate causal evidence. As summarized by Cook and Campbell (1979) and Shadish et al. (2002), a straightforward framework that can be used to evaluate causality is that of the nineteenth century philosopher John Stuart Mill, who outlined three conditions for causal inference:

(1) The cause must come before the effect;

(2) There is a relationship between the cause and effect; and

(3) No other plausible causal agents could have produced the effect.

These three conditions from John Stuart Mill are used throughout this dissertation to evaluate the validity of causal claims.

To make Mill's three conditions more tangible, consider the situation where a researcher hypothesizes that more frequently engaging children in early literacy activities in the home increases reading achievement at the fourth grade. One would need to show that engaging children in early literacy activities occurs prior to the measure of reading achievement (condition 1), that increases in early literacy activities are associated with increases in achievement (condition 2), and that all rival hypotheses that could have caused this increase in reading achievement can be plausibly negated (condition 3).

If early literacy activities are defined as activities that children engage in with their parents before starting primary school, as defined by PIRLS, it would be safe to conclude that children engage in these activities before they take the fourth grade PIRLS assessment, fulfilling the first condition. Likewise, if there is a statistical association between early literacy activities and reading achievement, then it is also possible to conclude that the second condition is

fulfilled. Nevertheless, it is difficult to fulfill the third condition—that the increases in achievement could not have been caused by factors other than early literacy activities, such as a good preschool program, a good reading teacher, or outside-of-school reading lessons.

**Cross-Sectional Analysis and International Large-Scale Assessments**

Cross-sectional analysis measures a person or unit at just one particular time point, and provides a snapshot of the statistical relationships at that one time point. In the context of fourth grade reading achievement, this snapshot can provide descriptions of the contexts for learning of high achieving and low achieving students.

Since the first TIMSS study in 1995 and the first PIRLS study in 2001, the international reports are a great resource for policymakers across the world (Martin, Mullis, Foy, Hooper, 2016; Mullis, Martin, Foy, Hooper, 2016; Mullis, Martin, Foy, & Drucker, 2012). The reports not only document relative achievement across educational systems and trends in achievement results for each educational system, but also provide information on the background characteristics that are associated with student achievement. This background information when coupled with the achievement results provides important insights into factors related to student achievement.

The analyses in the international result reports tend to be primarily descriptive in nature and typically focus on within-country relationships—factors that relate to achievement within most of the participating countries. Figure 2.1 provides an excerpt from PIRLS 2011 report (Mullis et al., 2012, p. 126) focusing on the relationship between early literacy activities and student achievement. Nine items reported by parents on their engagement with their children in literacy activities before beginning primary school were scaled through Rasch item response theory methodology to provide a measure of early literacy activities for each student (Martin,

Mullis, Foy, & Arora, 2012). Students were then classified into regions of the scale based on their parents' reports of their engagement in these activities, and the relationship between the mean achievement of students in each of these regions can be compared within each country. As can be seen in the figure, across the countries listed, students whose parents "Often" engaged them in early literacy activities tended to have higher academic achievement than those whose parents "Sometimes" engaged them.

**Figure 2.1: Excerpt from *PIRLS 2011 International Results in Reading*, Exhibit 4.6: Early Literacy Activities Before Beginning Primary School**

**Exhibit 4.6: Early Literacy Activities Before Beginning Primary School**

PIRLS 2011 4th Grade

*Reported by Parents*

Students were scored according to their parents' frequency of doing the nine activities on the *Early Literacy Activities* scale. Students **Often** engaged in early literacy activities had a score on the scale of at least 10.7, which corresponds to their parents "often" doing five of the nine activities with them and "sometimes" doing the other four, on average. Students **Never or Almost Never** engaged in such activities had a score no higher than 6.2, which corresponds to parents "never or almost never" doing five of the nine activities with them and "sometimes" doing the other four, on average. All other students had parents who **Sometimes** engaged them in early literacy activities.

| Country | | Often | | Sometimes | | Never or Almost Never | | Average Scale Score |
|---|---|---|---|---|---|---|---|---|
| | | Percent of Students | Average Achievement | Percent of Students | Average Achievement | Percent of Students | Average Achievement | |
| Russian Federation | | 61 (1.3) | 576 (2.7) | 38 (1.2) | 558 (3.4) | 1 (0.3) | ~ ~ | 11.1 (0.06) |
| Northern Ireland | s | 59 (1.3) | 582 (3.5) | 41 (1.4) | 559 (3.7) | 0 (0.2) | ~ ~ | 11.2 (0.06) |
| New Zealand | s | 55 (1.0) | 567 (2.7) | 44 (1.0) | 529 (2.5) | 1 (0.1) | ~ ~ | 11.0 (0.05) |
| Australia | s | 52 (1.4) | 555 (3.0) | 46 (1.3) | 528 (3.4) | 1 (0.3) | ~ ~ | 10.8 (0.06) |
| Georgia | | 52 (1.4) | 498 (2.6) | 47 (1.3) | 479 (4.0) | 1 (0.2) | ~ ~ | 10.7 (0.06) |
| Canada | r | 51 (0.9) | 566 (1.9) | 48 (0.9) | 541 (1.8) | 1 (0.1) | ~ ~ | 10.7 (0.04) |
| Ireland | | 50 (0.9) | 569 (2.3) | 49 (0.8) | 542 (2.6) | 1 (0.1) | ~ ~ | 10.8 (0.04) |
| Croatia | | 50 (0.9) | 562 (2.2) | 49 (0.9) | 544 (1.9) | 0 (0.1) | ~ ~ | 10.7 (0.03) |
| Slovenia | | 48 (1.2) | 543 (2.3) | 51 (1.2) | 522 (2.6) | 0 (0.1) | ~ ~ | 10.6 (0.04) |
| Israel | r | 48 (1.0) | 563 (3.0) | 51 (1.0) | 534 (3.5) | 1 (0.2) | ~ ~ | 10.6 (0.04) |
| Italy | | 48 (0.9) | 553 (2.4) | 51 (1.0) | 537 (2.6) | 1 (0.2) | ~ ~ | 10.5 (0.03) |
| Slovak Republic | | 47 (0.9) | 547 (2.9) | 51 (0.9) | 530 (2.5) | 2 (0.6) | ~ ~ | 10.5 (0.05) |
| Trinidad and Tobago | | 47 (1.1) | 497 (4.0) | 52 (1.1) | 456 (4.1) | 1 (0.3) | ~ ~ | 10.5 (0.05) |
| Malta | | 45 (0.9) | 507 (1.9) | 54 (0.9) | 463 (2.7) | 1 (0.2) | ~ ~ | 10.4 (0.04) |
| Spain | | 44 (1.0) | 528 (2.7) | 55 (1.0) | 507 (2.7) | 1 (0.2) | ~ ~ | 10.4 (0.03) |
| Poland | | 43 (0.8) | 544 (2.8) | 56 (0.8) | 514 (2.1) | 1 (0.3) | ~ ~ | 10.4 (0.03) |
| Hungary | | 43 (0.8) | 553 (2.8) | 56 (0.8) | 535 (3.2) | 1 (0.4) | ~ ~ | 10.3 (0.04) |
| Czech Republic | | 40 (1.0) | 555 (2.6) | 60 (1.0) | 542 (2.3) | 1 (0.2) | ~ ~ | 10.3 (0.03) |
| Netherlands | s | 40 (0.8) | 559 (3.1) | 60 (0.8) | 551 (2.0) | 1 (0.2) | ~ ~ | 10.2 (0.03) |
| Bulgaria | | 39 (1.4) | 559 (3.1) | 51 (1.0) | 529 (3.7) | 9 (1.4) | 455 (15.3) | 9.7 (0.12) |
| Romania | | 38 (1.5) | 529 (4.1) | 54 (1.3) | 494 (4.5) | 8 (1.0) | 423 (8.9) | 9.9 (0.09) |
| Germany | r | 38 (0.9) | 555 (2.8) | 61 (0.9) | 543 (2.2) | 1 (0.2) | ~ ~ | 10.2 (0.03) |
| Norway | | 37 (1.4) | 524 (2.5) | 63 (1.4) | 500 (2.2) | 1 (0.2) | ~ ~ | 10.0 (0.06) |
| France | | 36 (0.7) | 536 (2.6) | 63 (0.7) | 515 (2.7) | 1 (0.2) | ~ ~ | 10.0 (0.03) |
| Lithuania | | 36 (0.9) | 541 (1.9) | 63 (0.9) | 524 (2.5) | 2 (0.2) | ~ ~ | 10.0 (0.03) |
| Austria | | 35 (1.0) | 543 (2.1) | 63 (1.1) | 523 (2.5) | 1 (0.2) | ~ ~ | 10.0 (0.03) |
| Portugal | | 35 (1.1) | 558 (2.8) | 63 (1.1) | 535 (2.6) | 2 (0.4) | ~ ~ | 10.0 (0.05) |
| Sweden | | 34 (1.0) | 562 (2.9) | 64 (1.0) | 537 (2.2) | 2 (0.2) | ~ ~ | 9.9 (0.04) |
| Colombia | | 34 (1.1) | 457 (5.7) | 63 (1.0) | 448 (3.8) | 3 (0.4) | 409 (11.0) | 9.9 (0.06) |
| Denmark | | 32 (0.9) | 567 (2.2) | 67 (0.9) | 550 (1.9) | 1 (0.2) | ~ ~ | 9.9 (0.03) |

SOURCE: IEA's Progress in International Reading Literacy Study – PIRLS 2011

Source: Mullis et al. (2012, p. 126), Copyright © 2012 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Through such bivariate analysis, it is not possible to negate the hypothesis that other factors could have caused these achievement differences, and therefore not possible to fulfill Mill's third condition. To rule out these other plausible causal agents, researchers often include covariates in the model. The problem, however, is that it is never possible in cross-sectional analysis to include enough covariates to negate all of the plausible explanations for the statistical relationships in the data.

**Randomized Control Trials**

It is generally accepted that identifying cause and effect is best justified through randomized control trials (Shadish et al., 2002). In the simplest randomized control trials, researchers randomly assign participants to treatment and control conditions—those in the treatment group receive the intervention and those in the control group maintain the status quo. At the end of the intervention, both groups are tested and the treatment effect is considered to be the difference between the treatment group and the control group on the outcome measure.

Consider a randomized control trial evaluating the treatment effect of a special type of preschool program for four-year-olds on literacy outcomes at Kindergarten entry. To conduct such an experiment in its simplest form, one would randomly assign four-year-olds to treatment and control conditions, and then provide the preschool instruction to the treatment group but not to the control group. At Kindergarten entry, the two groups would be tested and the difference in average scores between the groups would be considered the treatment effect.

In the case of this preschool experiment, preschool predated the reading test, fulfilling Mill's first condition. Assuming that students who went to the preschool achieved the higher literacy score, it can be concluded that there is a positive statistical association between

preschool attendance and higher literacy scores, fulfilling Mill's second condition. In addition, random assignment allows one to assume that prior to beginning preschool, the treatment and control groups were probabilistically equivalent, and therefore assume that both groups would be equivalent in terms of factors influencing their later test scores other than those related to the treatment—fulfilling the third condition. Probabilistic equivalence means that any differences between the groups prior to the experiment were due to random chance alone.

Nevertheless, in this randomized control trial, a number of assumptions need to be made in order to justify a causal claim. A primary assumption is that the randomization worked to make the groups equal, not just probabilistically equivalent, on any factors that could influence their later test scores. Although it is understood in statistics that over infinite replications randomization ensures equivalence, for most experiments there is only one randomization and no replications. As such, despite randomization there may be differences between the treatment and control groups before the treatment is applied. Such differences are more likely to exist when the number of subjects participating in the experiment is relatively small.

Another assumption is that other than the preschool instruction treatment, the groups had equivalent experiences from the beginning of the treatment through the literacy test at Kindergarten entry, and therefore all other possible causes of the higher reading scores can be plausibly negated.

The point is that even in randomized control trials there are numerous threats to the validity of causal inferences, and when a treatment effect is found, it is often difficult to know

for certain whether the effect is caused by the treatment itself, random chance, or a rival causal agent within the experimental process.[1]

In addition to these threats to the validity of randomized control trials, it also can be difficult to use experimental designs to evaluate certain research questions due to ethical or feasibility issues. For example, if one wanted to conduct an experiment to estimate the efficacy of early literacy activities in the home, it would be unethical to ask parents in the control group to withhold engaging in early literacy activities with their children. Following from this, if parents in the control group are allowed to continue status quo educational activities with their children, then parents in both the treatment and control group would have engaged with their children in such activities. Assuming that this engagement in early literacy activities is effective, the treatment effect would be underestimated since both groups engaged in the literacy activities.

In a real-world example, Duncan and Magnuson (2013) notice in their meta-analysis that the effect size of preschool attendance has decreased since the 1960's. They posit that the decrease in effect size could be linked to the increasing early childhood educational activities of the control group. In recent years, those in the control group may be more likely to attend another preschool (other than the intervention preschool) or be raised in a more educationally-centered home environment.

**Longitudinal Designs**

For both randomized control trials as well as for other analyses, longitudinal data can provide numerous benefits. In the context of randomized control trials, additional measurement points can serve a number of purposes. Prior to the administration of the treatment, a pretest can be

[1] This review lists some of the more common critiques of randomized control trials. For more information, please see a recent critique by Ginsberg and Smith (2016).

used to ensure that the randomization functioned to make groups equivalent, at least on the outcome of interest, by examining whether the treatment and control groups have equal scores on the construct prior to the intervention. In addition, a number of measures can be taken during the intervention process to evaluate differences in change over time. By examining change over time, it is possible to evaluate the growth trajectory of the students, and create a picture of how differences between the treatment and control groups evolve over time. Outcome measures can also be collected following the completion of the treatment making it possible to longitudinally examine the long-term effects of the program.

For many longitudinal designs, participants are not randomly assigned to treatment and control conditions. In such designs, data are collected at multiple time points for members of the sample. Because participation in the treatment is not random, it cannot be determined whether those receiving the treatment are equivalent to those not receiving the treatment. As such, in longitudinal studies without randomization many feel it is impossible to rule out the threat of omitted variable bias (Holland, 1986), casting doubt on potential for causal inference.

Nevertheless, analysis of longitudinal data is considered to be superior to cross-sectional analysis because it becomes possible to control for all variables that do not change over time, whether these variables are measured or unmeasured. In the longitudinal design literature, variables that do not change over time are referred to as time-invariant variables, and, in contrast, variables that change over time are time-variant variables. In analysis of individual data, time-invariant variables could include demographic characteristics such as gender, race, or social class, or could be related to educational activities and experiences that happened prior to the first measurement point.

Longitudinal designs are especially advantageous when reverse causality is present. As outlined by Gustafsson (2010), in education, it is common to provide struggling students beneficial learning conditions such as assigning them to small classes or providing them with additional instructional time or homework. In these circumstances, lower prior achievement causes differential access to the supposedly efficacious educational opportunity leading to a reverse causality in the data—the analyst strives to estimate the effect of reduced class size on achievement but the measured effect is confounded by the effect of prior achievement on class assignment. For researchers conducting cross-sectional analysis it is difficult to partial out the effect of differential access to the educational practice from the actual effect of the practice. Cross-sectional analysis often shows such practices to have little relationship with achievement or even a negative relationship with achievement, but in reality the estimated treatment effect is likely biased downwards due to selection bias—lower achieving students being more likely to receive the treatment.

One way to negate the influence of reverse causality is through a measure of prior achievement. If a variable measuring prior achievement were available, it would be possible to control for prior achievement and come closer to estimating an unbiased treatment effect.

Consider the case of estimating the effect of homework on student achievement. Analysis of one cross-section of TIMSS data does not show a positive relationship between time spent on homework and student achievement. One explanation for this is that struggling students spend more time on homework than high achieving students. If a prior achievement measure were available, it would be possible to better assess whether the duration of time spent on homework is related to achievement since prior ability could be used as a control.

## 2.2 Technical Details on Techniques for Longitudinal Data Analysis

There are a number of ways to analyze longitudinal data, including adding prior achievement as a control variable and conducting analysis through fixed-effects and random-effects approaches. This section outlines the technical details of these three longitudinal analysis techniques. An understanding of the distinctions between these techniques as applied to individual-level data provides a foundation for comprehending the longitudinal analysis approaches applicable to country and subpopulation longitudinal analyses. This section covers the theory behind these techniques as generally applied in longitudinal data analysis at the individual level. Chapter 4 analysis will demonstrate applications of the fixed-effects and random-effects approaches, as well as the corresponding structural equation models associated with these approaches.

One popular way to analyze longitudinal data is to control for prior achievement within a regression model (Keppel & Wickens, 2004; Morgan & Winship, 2015). In the scenario where there are two measures, a prior measure and final measure, researchers often use the prior measure as the control variable so that they can control for prior differences when estimating the (treatment) effect of an indicator variable:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i \qquad i = 1, ..., I; \qquad \textbf{Equation 2.1}$$

Where $Y_i$ is the outcome variable for person $i$, $\beta_0$ is the intercept, $X_{1i}$ is the measure on the indicator variable for person $i$ (i.e., the variable of interest—in causal designs the treatment effect), $\beta_1$ is a regression weight for the measure $X_{1i}$, $X_{2i}$ is the prior measure for person $i$, $\beta_2$ is a regression weight for the prior measure $X_{2i}$, and $e_i$ is the residual term associated with the difference between person $i$'s expected $Y_i$ value and person $i$'s observed $Y_i$ value.

As in the analysis of covariance (ANCOVA) methodology geared toward experimental designs where $X_{1i}$ indicates the administration of the treatment, orthogonality between $X_{1i}$ and $X_{2i}$ is preferred because it implies that the treatment $X_{1i}$ is uniformly administered across the distribution of the prior measure $X_{2i}$. In such situations, the prior measure $X_{2i}$ is assumed to explain variance in $Y_i$ that is unrelated to $X_{1i}$—thereby, decreasing the standard error associated with coefficient $\beta_1$ and increasing statistical power.

In situations where there is a substantial covariance between $X_{1i}$ and $X_{2i}$, the model conditions the $\beta_1$ coefficient estimates on prior scores $X_{2i}$ and thereby controls for between-person differences prior to the treatment. Nevertheless, covariance between $X_{1i}$ and $X_{2i}$ is less than ideal because in this case the partial regression coefficient $\beta_1$ may be estimated to have a quite different value then it would if $X_{2i}$ were not included in the model, and the covariance between $X_{1i}$ and $X_{2i}$ could also lead to a large standard error associated with the $\beta_1$ partial regression coefficient.

Another common method is a random-effects model. In a random-effects model, the score of a person at each time point is conceptualized to be composed of a combination of time-variant characteristics and time-invariant characteristics.

$$Y_{ti} = \beta_{t0} + \beta_1 X_{ti} + \gamma_1 W_i + \mu_i + e_{ti} \qquad i = 1, ..., I;\ t = 1, ..., T \qquad \textbf{Equation 2.2}$$

Where $Y_{ti}$ is the outcome variable for person *i* at time *t*, $\beta_{t0}$ is the intercept that varies across time, $X_{ti}$ is the value on the predictor variable for person *i* at time *t*, $\beta_1$ is a regression weight

associated with the predictor variable $X_{ti}$, $W_i$ is the value of person *i* on a predictor variable that is static across time, $\gamma_1$ is a regression weight associated with $W_i$, $\mu_i$ is a random variable with a certain probability distribution representing person-specific, time-invariant deviations from the model for person *i*, and $e_{ti}$ represents time-specific deviations from the model for person *i*.

An important assumption of the random effects approach is that $X_{ti}$ is uncorrelated with $\mu_i$. When $X_{ti}$ and $\mu_i$ are correlated the random effects model provides biased estimates of $\beta_1$, because the relationship between the explanatory variable of interest and the outcome variable is confounded by unmeasured differences between people that are static across time.

One way to control for this confounding between $X_{ti}$ and $\mu_i$ is by controlling for all time-invariant individual characteristics and focusing analysis on the relationship between changes in $X$ and changes in $Y$. Models focusing analysis on changes over time are called fixed-effects models. [2] As described by Allison (2004), one example of such a model is the first-difference approach. The model begins with an equation representing each time point, similar to the random effects model from Equation 2.2:

$$Y_{1i} = \beta_{10} + \beta_1 X_{1i} + \gamma_1 W_i + \mu_i + e_{1i} \qquad \textbf{Equation 2.3}$$

$$Y_{2i} = \beta_{20} + \beta_1 X_{2i} + \gamma_1 W_i + \mu_i + e_{2i}$$

[2] Econometricians often distinguish between the first-difference and fixed-effects approaches—the latter includes a set of dummy-variable fixed effects to control for unmeasured time-invariant characteristics (see Equation 2.5). Because this dissertation exclusively focuses on analysis across two time points and both approaches provide identical unstandardized coefficient estimates when analysis is conducted across two time points, this dissertation considers the first-difference approach to be under the umbrella of the fixed-effects model and therefore refers to it as a "fixed-effects approach."

From there, the first equation is subtracted from the second equation to become:

$$(Y_{2i} - Y_{1i}) = (\beta_{20} - \beta_{10}) + \beta_1(X_{2i} - X_{1i}) + (e_{2i} - e_{1i}) \quad \textbf{Equation 2.4}$$

The terms that remain the same over time, $\gamma_1 W_i$ and $\mu_i$ have been differenced out of the latter equation. The advantage of the fixed-effects model is that unobserved student characteristics, $\mu_i$, are controlled for by differencing them out, and by differencing out $\mu_i$ the bias associated with estimates of $\beta_1$ due to the covariance between $\mu_i$ and $X_{ti}$ is eliminated from the model. Basically, the model examines whether changes in $X_i$ are associated with changes in $Y_i$. Changes in $e_i$ represent all of the other time-variant variables not included in the model as well as any random variation.

The standard fixed-effects approach from the econometrics literature, which provides identical unstandardized coefficient estimates for $\beta_1$ to the first-difference estimator for measurement across two time points, uses a set of $K$ dummy variables ($Z$) to partial out the variability associated with between-person differences. In this model, the outcome measure $Y_{ti}$ is regressed on $X_{ti}$, a dummy variable $C_{ti}$ representing the time point for the analysis, and the set of dummy variables $Z_k$:

$$Y_{ti} = \beta_{t0} + \beta_1 X_{ti} + \beta_2 C_{ti} + \sum_{k=1}^{K} \gamma_k (Z_k) + e_{ti} \quad \textbf{Equation 2.5}$$

$$i = 1, \ldots, I; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K;$$

A $Z_k$ dummy variable is associated with every person but one—the one person not represented by a dummy variable serves as a reference in the estimation process. These dummy variable are

commonly referred to as "fixed effects," and the coefficient $\gamma_k$ is estimated for each of the $K$ individuals. $\gamma_k$ describes the adjusted time-invariant difference on $Y_i$ between person $k$ and the reference individual. $Z_k$ subsumes all of the measured ($W_i$) and unmeasured ($\mu_i$) time-invariant characteristics. By including $Z_k$ in the analysis, the model controls for all of the time-invariant characteristics of person $k$, thereby ensuring that $X_{ti}$ is uncorrelated with time-invariant unobserved individual characteristics. $C_{ti}$ is another dummy variable taking on the value of 1 for one of the two time points in the analysis and 0 for the other time point. The regression coefficient $\beta_2$, associated with $C_{ti}$, represents the adjusted mean difference in $Y_i$ across the two time points.

Fixed-effects models are advantageous because they control for all time-invariant omitted variables, since both measured $W_i$ and unmeasured $\mu_i$ are differenced out of the model in the first-difference approach. Consequently, the fixed-effects model focuses estimation on the relationship between $Y_{ti}$ and the time-varying covariate $X_{ti}$ at the expense of not being able to estimate the contribution of the time-invariant observed variable $W_i$. The random-effects model, on the other hand, allows for the estimation of the contribution of $W_i$ but can provide biased estimates when there is substantial covariance between $\mu_i$ and $X_{ti}$.

For researchers looking into moving from a fixed-effects model to a random-effects model, traditionally a Hausman test can provide evidence of the degree of bias introduced to the $\beta_1$ coefficient estimates by use of the random effects model. The Hausman test compares the coefficient estimates from the fixed-effects model with those in the random-effects model to examine whether the random-effects model unduly biases the estimation of the coefficients.

**Structural Equation Modeling Approach**

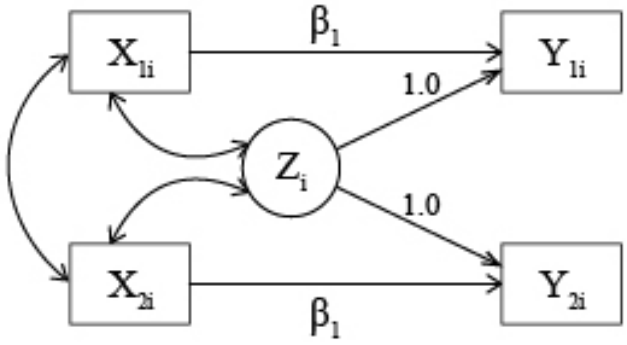
In addition to these traditional approaches for analyzing longitudinal data, Allison and Bollen (1997) devised a way to estimate the fixed-effects and random-effects models within the structural equation modeling framework, and since this initial paper a number of publications have further elaborated on this framework (Allison, 2005, 2009; Bollen & Brand, 2008, 2010). Estimating fixed-effects and random-effects models through the structural equation modeling framework is advantageous for a number of reasons. Namely, the structural equation modeling approach offers more flexibility in testing assumptions, allows for more complicated models within the same analysis including those with numerous dependent variables, and provides the opportunity for including latent variables as explanatory or outcome variables. Because Gustafsson and Nilsen (2016) have recently applied the structural equation modeling approach to difference-in-differences analysis using international large-scale assessment data, and these approaches are utilized in the analysis in this dissertation, this section provides an overview of how fixed-effects and random-effects analysis is employed through the structural equation modeling approach.

Figure 2.2 shows the path model of the structural equation modeling version of the fixed-effects analysis when there are measures at two time points. The dependent variables are $Y_{1i}$ and $Y_{2i}$ and the time-varying independent variables are $X_{1i}$ and $X_{2i}$. $Z_i$ is a latent variable representing all time-invariant characteristics ($W_i$ and $\mu_i$) of $Y_{ti}$. In contrast to the regression-based fixed-effects model, in the structural equation modeling approach the latent variable $Z_i$ exists theoretically for person *i,* but a coefficient describing the relationship between $Z_i$ and $Y_{ti}$ is not estimated in the model. Instead, the paths between $Z_i$ and $Y_{1i}$ and $Z_i$ and $Y_{2i}$ are fixed to 1,

meaning that $Z_i$ is equally related to both $Y_{1i}$ and $Y_{2i}$.[3] Allowing $Z_i$ to covary with $X_{1i}$ and $X_{2i}$ controls for all measured and unmeasured time-invariant characteristics and as such ensures that the correlation between $X_{ti}$ and time-invariant person-specific characteristics do not bias the estimation of $\beta_1$. $\beta_1$ thereby represents the relationship between changes in $X_{ti}$ and changes in $Y_{ti}$, and the structural equation modeling approach provides identical estimates of $\beta_1$ when compared with regression-based fixed-effects analysis in Equations 2.4 and 2.5.

**Figure 2.2: Path Model for Fixed-Effects Approach**

$$Y_{ti} = \beta_{t0} + \beta_1 X_{ti} + Z_i + e_{ti}$$




Source: Modified from Allison and Bollen (1997, p.6)

According to the proponents of the structural equation modeling approach (Allison & Bollen, 1997; Allison, 2005, 2009; Bollen & Brand, 2008, 2010), a major advantage it has over the regression-based approach is that fit statistics facilitate testing to what extent the data matches the model and to what extent these constraints should be relaxed. For example, it may be the case that the effect of $X_{ti}$ on $Y_{ti}$ is different at each of the time points. In this case,

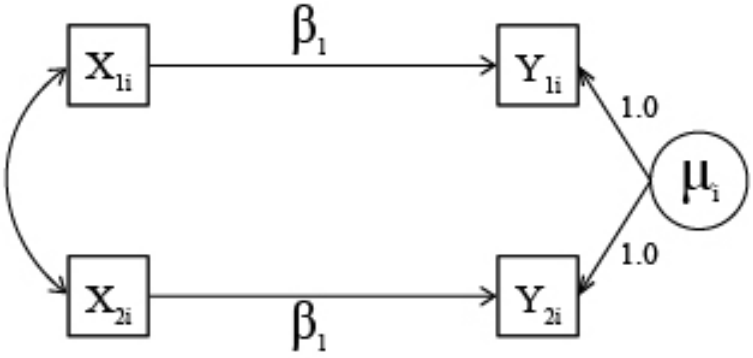
[3] Generally, the paths are fixed to 1. Fixing the paths to another number will solely change the scale of the $Z_i$ residual estimates (but would not affect the scale of the $\beta_1$ estimates) .

constraining $\beta_1$ across time may be unrealistic, and therefore this constraint can be relaxed and unique coefficients can be estimated for each time point.

From the structural equation modeling perspective, as can be seen in Figure 2.3, the difference between the random-effects model and the fixed-effects model is simply that the latent variable is no longer allowed to covary with $X_{ti}$. To maintain continuity with the regression-based random-effects model, the latent variable from Figure 2.2 $Z_i$ is renamed $\mu_i$ because it now represents unmeasured random effects.[4]

**Figure 2.3: Path Model for the Random-Effects Approach**

$$Y_{ti} = \beta_{t0} + \beta_1 X_{ti} + \mu_i + e_{ti}$$




Source: Modified from Allison and Bollen (1997, p.6)

Similar to the corresponding regression-based approaches, researchers should be careful in transitioning from the fixed-effects to the random-effects structural equation modeling approaches to ensure the random-effects model does not introduce undue bias to the coefficient

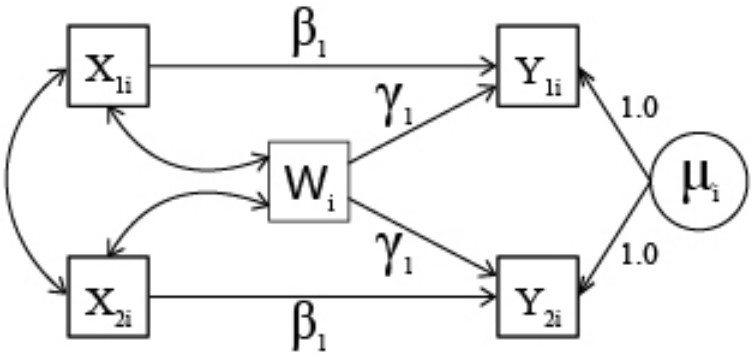
[4] In the structural equation modeling random effects approach across two time points, it is not necessary to include $\mu_i$ in the model. The random-effects analysis will provide the same estimates of $\beta_1$ and fit statistics if $Y_{1i}$ and $Y_{2i}$ are allowed to covary. Nevertheless, $\mu_i$ is maintained in the model because it facilitates illustrating the difference between the random-effects and fixed-effects approaches.

estimates. Instead of using a Hausman test to examine this confounding, with the structural equation modeling approach it is possible to examine the relationship within the fixed-effects approach between $X_{ti}$ and $Z_i$. The strength of this covariance is easily identifiable in the structural equation modeling fixed-effects approach (Figure 2.2) through an examination of the magnitude of the covariance and its significance as well as through a comparison of fit statistics across the random- and fixed-effects approaches. Referring back to Figure 2.2, when $X_{ti}$ is approximately orthogonal to $Z_i$, the estimate of $\beta_1$ remains unbiased in the random-effects model. Bollen and Brand (2010) argue that examining this covariance is a more direct evaluation than the Hausman test since the coefficient estimates examined by the Hausman test are not a measure of this confounding but instead a consequence of it.

Like its regression-based counterpart, the random-effects model through the structural equation modeling approach allows for estimation of the effects of the time-invariant observed covariate $W_i$, as can be seen in Figure 2.4. By convention the observed covariate $W_i$ is allowed to covary with $X_{1i}$ and $X_{2i}$, as would be the case in a regression model.

**Figure 2.4: Path Model for the Random-Effects Approach with Observed Time-Invariant Covariate**

$Y_{ti} = \beta_{t0} + \beta_1 X_{ti} + \gamma_1 W_i + \mu_i + e_{ti}$




Source: Modified from Allison and Bollen (1997, p.6)

## 2.3 PIRLS as a Longitudinal Study

Keeping this background in longitudinal analysis in mind, PIRLS can be conceptualized as a country-level longitudinal design, where the same country unit is measured at multiple time points. Conceptualizing PIRLS longitudinally, PIRLS in essence examines the growth in achievement for a country and subgroups. Figure 2.5 shows an excerpt from the *PIRLS 2011 International Results in Reading* report (Mullis et al., 2012, p. 48), that focuses on trend analysis—how a country's mean achievement changes over time. For example, Bulgaria's reading achievement in 2011 was found to be significantly lower than it was in PIRLS 2001 and PIRLS 2006, with the mean score in Bulgaria being 532 in PIRLS 2011, compared with 547 in PIRLS 2006 and 550 in PIRLS 2001.

**Figure 2.5: Excerpt from *PIRLS 2011 International Results in Reading*, Exhibit 1.4: Trends in Reading Achievement**

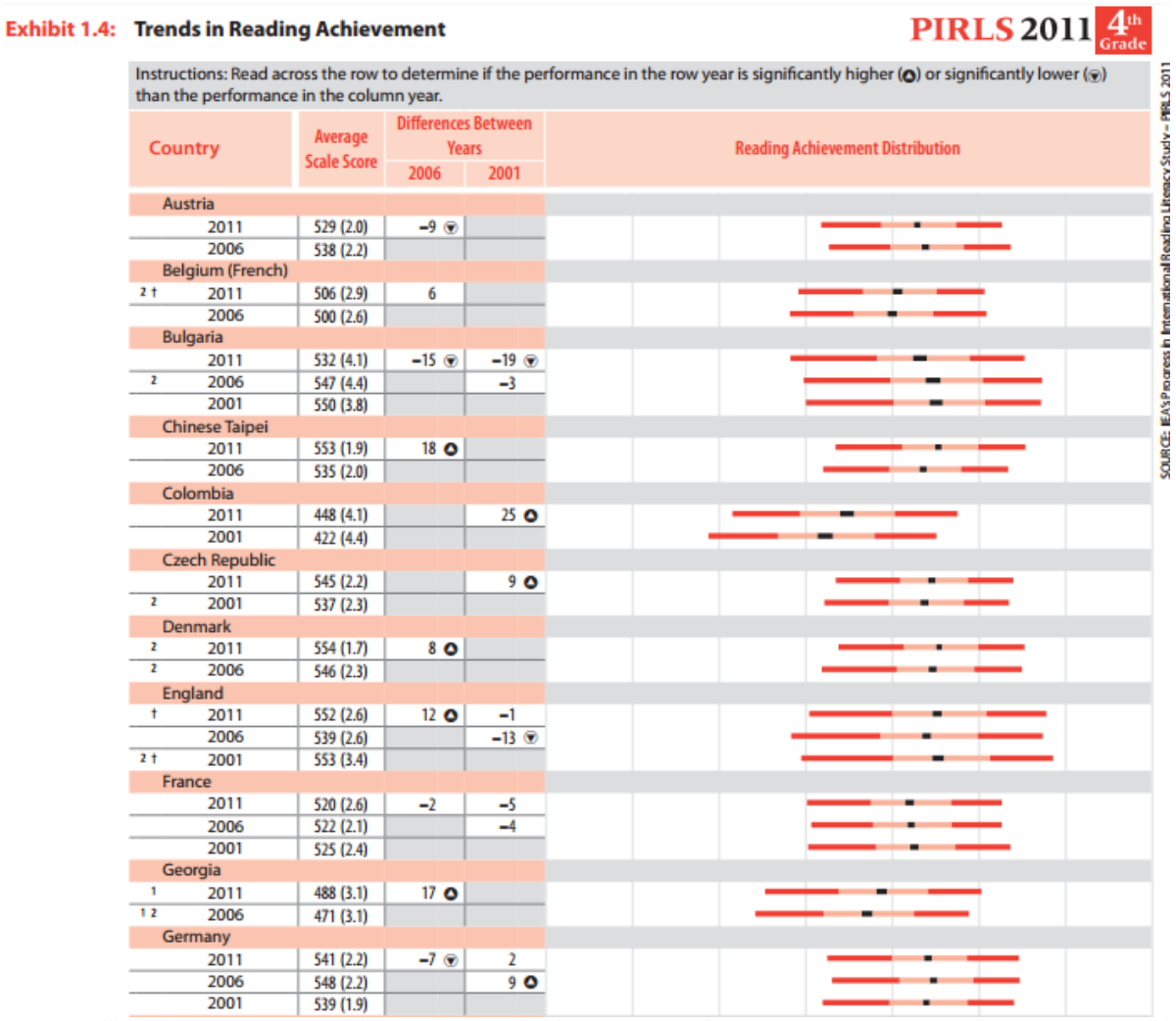


**Exhibit 1.4: Trends in Reading Achievement**

PIRLS 2011 4th Grade

Instructions: Read across the row to determine if the performance in the row year is significantly higher (▲) or significantly lower (▼) than the performance in the column year.

| Country | Average Scale Score | Differences Between Years: 2006 | Differences Between Years: 2001 | Reading Achievement Distribution |
|---|---|---|---|---|
| Austria | | | | |
| 2011 | 529 (2.0) | −9 ▼ | | |
| 2006 | 538 (2.2) | | | |
| Belgium (French) | | | | |
| 2 † 2011 | 506 (2.9) | 6 | | |
| 2006 | 500 (2.6) | | | |
| Bulgaria | | | | |
| 2011 | 532 (4.1) | −15 ▼ | −19 ▼ | |
| 2 2006 | 547 (4.4) | | −3 | |
| 2001 | 550 (3.8) | | | |
| Chinese Taipei | | | | |
| 2011 | 553 (1.9) | 18 ▲ | | |
| 2006 | 535 (2.0) | | | |
| Colombia | | | | |
| 2011 | 448 (4.1) | | 25 ▲ | |
| 2001 | 422 (4.4) | | | |
| Czech Republic | | | | |
| 2011 | 545 (2.2) | | 9 ▲ | |
| 2 2001 | 537 (2.3) | | | |
| Denmark | | | | |
| 2 2011 | 554 (1.7) | 8 ▲ | | |
| 2 2006 | 546 (2.3) | | | |
| England | | | | |
| † 2011 | 552 (2.6) | 12 ▲ | −1 | |
| 2006 | 539 (2.6) | | −13 ▼ | |
| 2 † 2001 | 553 (3.4) | | | |
| France | | | | |
| 2011 | 520 (2.6) | −2 | −5 | |
| 2006 | 522 (2.1) | | −4 | |
| 2001 | 525 (2.4) | | | |
| Georgia | | | | |
| 1 2011 | 488 (3.1) | 17 ▲ | | |
| 1 2 2006 | 471 (3.1) | | | |
| Germany | | | | |
| 2011 | 541 (2.2) | −7 ▼ | 2 | |
| 2006 | 548 (2.2) | | 9 ▲ | |
| 2001 | 539 (1.9) | | | |

SOURCE: IEA's Progress in International Reading Literacy Study – PIRLS 2011

Source: Mullis et al. (2012, p. 48). Copyright © 2012 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

It is also possible to measure trends on background variables. Figure 2.6 shows an excerpt of an exhibit from *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007) measuring trends in early literacy activities. For this exhibit, students were classified into three regions (high, medium, and low) based on their parents' reports of their participation in early literacy activities. The "difference in percent from 2001" column shows the difference between the percentage classified to that region in 2006 compared with the percentage classified

in 2001, and the corresponding arrow shows whether the 2006 percentage is significantly higher or lower than the 2001 percentage.

**Figure 2.6: Excerpt from *PIRLS 2006 International Report*, Exhibit 3.1: Index of Early Home Literacy Activities (EHLA) with Trends**

Exhibit 3.1 **Index of Early Home Literacy Activities (EHLA) with Trends** — PIRLS 2006 4th Grade

| Countries | | High EHLA | | | Medium EHLA | | | Low EHLA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 2006 Percent of Students | Average Achievement | Difference in Percent from 2001 | 2006 Percent of Students | Average Achievement | Difference in Percent from 2001 | 2006 Percent of Students | Average Achievement | Difference in Percent from 2001 |
| Scotland | s | 85 (1.1) | 547 (3.5) | 3 (1.6) | 14 (1.1) | 522 (8.2) | –2 (1.4) | 2 (0.4) | ~ ~ | 0 (0.5) |
| *Canada, Nova Scotia* | | 77 (0.8) | 553 (2.3) | ◊ ◊ | 20 (0.8) | 523 (3.4) | ◊ ◊ | 3 (0.3) | 510 (7.8) | ◊ ◊ |
| Russian Federation | | 75 (1.0) | 573 (3.2) | 9 (1.6) ▲ | 20 (0.8) | 548 (4.3) | –6 (1.3) ▼ | 4 (0.4) | 520 (6.7) | –3 (0.8) ▼ |
| New Zealand | s | 74 (1.0) | 560 (2.0) | 5 (1.5) ▲ | 22 (0.9) | 519 (3.8) | –4 (1.4) ▼ | 4 (0.4) | 501 (8.0) | –1 (0.7) |
| Israel | | 73 (1.2) | 526 (4.3) | x x | 22 (1.0) | 531 (5.8) | x x | 5 (0.4) | 531 (7.8) | x x |
| *Canada, Ontario* | r | 71 (1.3) | 563 (3.0) | 1 (1.6) | 23 (1.0) | 541 (4.2) | –2 (1.3) | 6 (0.6) | 539 (8.4) | 1 (0.8) |
| *Canada, British Columbia* | r | 71 (1.2) | 570 (2.9) | ◊ ◊ | 23 (1.0) | 547 (4.3) | ◊ ◊ | 6 (0.5) | 539 (6.7) | ◊ ◊ |
| *Canada, Alberta* | r | 70 (1.2) | 573 (2.5) | ◊ ◊ | 25 (1.1) | 554 (3.9) | ◊ ◊ | 5 (0.6) | 516 (6.4) | ◊ ◊ |
| Hungary | | 69 (0.9) | 560 (3.1) | 7 (1.4) ▲ | 26 (0.8) | 541 (3.7) | –6 (1.3) ▼ | 5 (0.5) | 525 (7.8) | –1 (0.7) |
| Spain | s | 68 (1.1) | 530 (2.5) | ◊ ◊ | 26 (1.0) | 506 (4.0) | ◊ ◊ | 6 (0.6) | 487 (6.4) | ◊ ◊ |
| Macedonia, Rep. of | r | 67 (1.0) | 460 (4.3) | 6 (1.7) ▲ | 27 (0.8) | 431 (5.0) | –4 (1.3) ▼ | 6 (0.5) | 414 (9.3) | –2 (1.1) |
| Trinidad and Tobago | | 67 (1.2) | 457 (5.1) | ◊ ◊ | 27 (0.9) | 416 (5.3) | ◊ ◊ | 6 (0.6) | 363 (10.4) | ◊ ◊ |
| Georgia | | 66 (1.5) | 481 (3.6) | ◊ ◊ | 26 (1.2) | 461 (4.3) | ◊ ◊ | 8 (1.0) | 458 (11.5) | ◊ ◊ |
| Italy | | 65 (1.0) | 561 (2.7) | 3 (1.4) ▲ | 28 (0.9) | 545 (3.9) | –2 (1.3) | 7 (0.6) | 531 (6.1) | –1 (0.7) |
| Slovak Republic | | 65 (1.1) | 542 (2.2) | 2 (1.5) | 30 (0.8) | 524 (3.3) | –2 (1.3) | 5 (0.6) | 475 (15.6) | 0 (0.8) |
| Netherlands | s | 64 (1.2) | 561 (1.8) | 9 (1.6) ▲ | 30 (1.0) | 547 (2.8) | –7 (1.4) ▼ | 6 (0.6) | 544 (5.1) | –2 (0.9) ▼ |
| *Canada, Quebec* | r | 64 (1.1) | 544 (3.0) | 3 (1.7) | 30 (1.0) | 523 (3.6) | –2 (1.6) | 6 (0.6) | 517 (6.1) | –1 (0.9) |
| Slovenia | | 64 (0.9) | 532 (2.3) | 6 (1.4) ▲ | 31 (0.8) | 510 (3.2) | –6 (1.3) ▼ | 5 (0.3) | 503 (5.2) | –1 (0.6) |

SOURCE: IEA Progress in International Reading Literacy Study (PIRLS) 2006

Source: Mullis et al. (2007, p. 109). Copyright © 2007 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

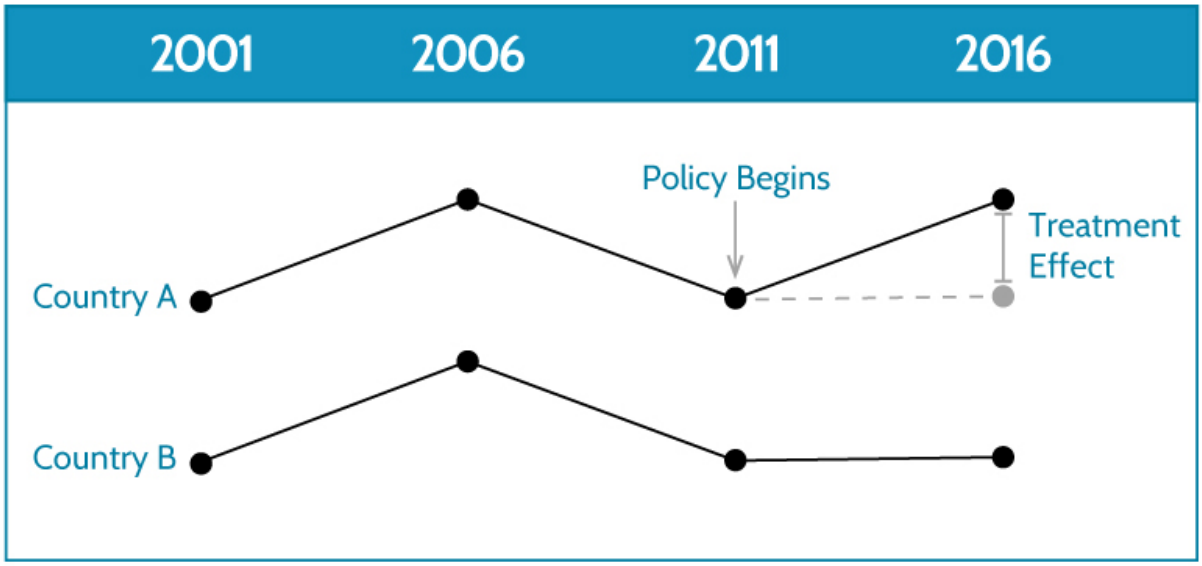
## 2.4 Difference-in-Differences

Given that international large-scale assessments like PIRLS measure trend at country-level for both achievement scales and background data, Gustafsson (2007) proposed analysis through a difference-in-differences approach. This methodology treats international large-scale assessments as longitudinal studies and uses pooled data across countries to examine whether country-level trends in background data (e.g., early literacy activities) are associated with country-level trends in achievement (e.g., reading achievement).

**Theoretical Background of Difference-in-Differences**

Difference-in-differences is based on the assumption of common trend (Angrist & Pischke, 2009, 2015; Schlotter et al., 2014). When comparing two entities, such as states or countries, difference-in-difference analysis assumes that the entities would have common trend—meaning that in the absence of the change to the policy or practice, the two entities would progress in parallel. Figure 2.7 shows a hypothetical example of two countries with common trend across PIRLS 2001, 2006, and 2011, and a deviation from that trend for PIRLS 2016. If the policy in question was first implemented in say 2011, following data collection in PIRLS 2011, then the researcher would have grounds to argue that this deviation in the trend line was due to the implementation of this policy. In difference-in-differences, the counterfactual is the growth rate of Country B, and the treatment effect is considered to be the difference between Country A's actual score in 2016 and what Country A's score would have been if Country A had the same increase or decrease from 2011 to 2016 as Country B.

**Figure 2.7: Illustration of the Common Trends Assumption for Difference-in-Differences**



Difference-in-differences has its foundation in econometrics. One well-known difference-in-differences study was conducted by Card and Krueger (1994). In economics, many assume that raising the minimum wage decreases the employment rate, and to test this Card and Krueger (1994) compared the relationship between the minimum wage and the employment rate through a natural experiment involving the fast food industry in New Jersey and Pennsylvania. As described in Angrist and Pischke (2009), this natural experiment was based on a policy change by New Jersey to raise the minimum wage from $4.25 to $5.05 in April 1992. Card and Krueger (1994) compared the change in New Jersey's employment rate at fast food restaurants following the introduction of the increased minimum wage to the change in Pennsylvania's employment rate, where there was no change in the minimum wage. The authors found that the employment rate increased slightly in New Jersey and decreased in Pennsylvania, and they concluded that raising the minimum wage does not necessarily decrease the employment rate.

Viewing Card and Krueger's (1994) analysis from Mill's causal perspective, Card and Krueger can assume that the policy change came before the outcome measure since the policy change was implemented in April 1992, and the employment data for the outcome variable was collected in November-December 1992—fulfilling Mill's first condition. Because Card and Krueger (1994) assume that the trend lines would have been parallel if this policy were not implemented implies that the third condition is also fulfilled—all relevant causal agents other than the policy change affecting employment in New Jersey also affected Pennsylvania. However, despite the design, a causal claim could not be drawn since Mill's second condition was not fulfilled—there was not a statistical relationship between the raising of the minimum wage and a decreased employment rate in New Jersey.

Difference-in-differences can also be implemented with multiple entities such as states or countries. Angrist and Pischke (2009) provide the example of Card (1992), which uses a multistate approach to conduct a natural experiment around the federal government's increase in the minimum wage from $3.35 to $3.80 in 1990. In the United States, the minimum wage in states must be at least as high as the federal minimum wage. However, many states have minimum wages above the federal minimum wage. To analyze this variation in minimum wage across states, Card (1992) aggregated data to state-level and examined whether there was a relationship between the estimated percentage of teenagers in a state making less than the minimum wage before 1990, and changes in teen employment after the minimum wage increase. Similar to the findings of Card and Krueger (1994), Card (1992) found no evidence that increases in the minimum wage decreased teen employment. The Card (1992) analysis is more similar to the difference-in-differences approach of Gustafsson (2007), in the sense that there are many states being analyzed, and the explanatory variable in the study was continuous.

**Difference-in-Differences Applied to International Large-Scale Assessments**

In their work of applying difference-in-differences to international large-scale assessment data, Hanushek and Wößmann (2006) analyzed the relationship between educational tracking and inequality in student achievement using data from TIMSS or PIRLS at the fourth grade as a covariate and data from TIMSS at the eighth grade or PISA at the age of 15 as the outcome variable. Several analyses were conducted with different combinations of explanatory and outcome variables. Hanushek and Wößmann (2006) assumed tracking to be a country-level policy and classified countries on whether tracking was implemented. The authors aggregated data to country-level and compared changes from the lower grade to the upper grade in the dispersion (standard deviation, top quartile v. bottom quartile, top 5% v. bottom 5%) of student

achievement in tracked countries and untracked countries. Their results showed larger dispersion, and therefore more inequality, at the higher grades (eighth grade for TIMSS and fifteen-year-olds for PISA) in tracked countries after controlling for the dispersion in primary school. Hanushek and Wößmann (2006) implemented difference-in-differences hierarchically, by using the fourth grade achievement data as a "pretest" control in analysis of eighth grade TIMSS measures or the 15-year-old PISA measures.

Gustafsson (2007) extended the work of Hanushek and Wößmann (2006) by adapting a difference-in-differences approach where each country-grade combination is the unit of analysis—examining whether countries' trend changes in mean achievement at one particular grade can be predicted by their trend changes in an explanatory variable.

It is important to note the distinction between the Hanushek and Wößmann's (2006) vertical approach to difference-in-differences and Gustafsson's (2007) horizontal approach. Hanushek and Wößmann (2006) employ a regression approach that uses the fourth grade results as a pretest control. This approach does not necessitate that the pretest measure be on the same scale as the posttest, and therefore this approach works well in the Hanushek and Wößmann (2006) context since the fourth grade studies (PIRLS and TIMSS) are on different metrics than TIMSS at the eighth grade and PISA. The primary advantage of the Hanushek and Wößmann (2006) approach is that the same cohort can be measured over time. For example, in one set of their analyses, Hanushek and Wößmann (2006) compared fourth grade country-level outcomes in TIMSS 1995 and eighth grade country-level outcomes in TIMSS 1999. If it can be assumed that the composition of the cohort did not change between the early grades and the later grades, then the Hanushek and Wößmann (2006) approach would control for cohort effects.

Since Gustafsson (2007) looks at the same grade over time, the metric remains the same across cycles (PIRLS achievement in 2001 is on the same metric as PIRLS achievement in 2006 and 2011), making it more defensible to employ the fixed-effects or random-effects approach to measure change over time. The Gustafsson (2007) method also aligns well with reporting in international assessments like TIMSS and PIRLS that report trend as changes in a country's achievement at a particular grade over time (as shown in Figures 2.5 and 2.6). However, because the Gustafsson approach measures a different cohort for each cycle, it is more susceptible to bias linked to cohort effects, such as from changes in a country's fourth grade population across cycles due to immigration or emigration.

**Argument for Gustafsson's (2007) Difference-in-Differences Approach**

This section details the argument for Gustafsson (2007) differences-in-differences approach and the following sections provide a critique of the approach.

Gustafsson (2007) argues that by aggregating data to country-level and assuming that a country is a stable unit for analysis, it is possible to control for country characteristics that do not change over time. Such country characteristics could include relative wealth compared with other countries, cultural characteristics that remain the same over time, and educational policies that are static across cycles.

Gustafsson (2007) validates his argument for this approach by providing two examples—one focusing on the relationship between student age and TIMSS mathematics achievement and the other focusing on the relationship between class size and TIMSS mathematics achievement.

To investigate the effect of student age, Gustafsson (2007) analyzed fourth grade and eighth grade TIMSS mathematics data for countries that participated in 1995 and 2003.

Aggregating data to country-level and examining each cycle separately through cross-sectional analysis, Gustafsson (2007) found no correlation between country-level achievement and student age, meaning that countries with older students did not have higher achievement in 1995 or 2003 than countries with younger students. Then, Gustafsson (2007) examined the age data as an explanation for trend gains over time, regressing changes in average mathematics achievement on changes in average student age. The regression analysis showed an unstandardized regression coefficient of 38, meaning that a one-year change in average student age between cycles is associated with a 38-point change in average student achievement. Although many countries had little change in student age, Latvia, Lithuania, Korea, and Romania had sizable changes, and in Latvia and Lithuania in particular these changes were associated with gains in student achievement. Gustafsson (2007) conducted another analysis at the fourth grade and found similar results, showing a regression coefficient of 71.[5] To test whether the results were overly influenced by outliers, the eighth grade data were reanalyzed after removing the highly influential countries from the model, and the results showed a regression coefficient of 28 at the eighth grade—large but smaller in magnitude than the original coefficient of 38 with the outlier countries. Reanalyzing the fourth grade data after removing outliers, Gustafsson (2007) reported a correlation coefficient similar to the correlation coefficient from the original analysis.[6]

[5] An informative scatterplot of the eighth grade data can be found in Figure 3.1 of Gustafsson (2007, p. 49), and a scatterplot of the fourth grade data can be found in Figure 3.2 of Gustafsson (2007, p. 51).
[6] Another regression coefficient was not reported for the re-analysis of the fourth grade data.

Gustafsson (2007, p. 47) explains the paradoxical results that in cross-sectional analysis student age is not related to student achievement but is related to changes in achievement across cycles:

> The correlation between age and achievement is influenced by cultural and economic factors and by such matters as country differences in starting age. These factors, which are omitted variables in the analysis of cross-sectional data, may conceal a true correlation between student age and achievement. However, the correlation between change in achievement and change in age within countries keeps these factors associated with countries constant, thereby allowing the correlation between age and achievement to appear.

In a second set of analyses, Gustafsson (2007) examines the relationship between class size and mathematics achievement. International large-scale assessment data have generally shown class size to have little relationship with student achievement (Hanushek & Wößmannn, 2017). As described by Gustafsson (2007), these results differ from those of randomized control trials, like the Tennessee STAR experiment, where 12,000 primary school students were randomly assigned to smaller or larger class sizes. The Tennessee STAR experiment found an effect size of 0.25, implying that a seven student decrease in average class size was associated with a quarter of a standard deviation increase in student achievement.

Gustafsson (2007) analyzed change in fourth grade mathematics achievement and change in class size across the TIMSS 1995 and TIMSS 2003 cycles through the difference-in-differences fixed-effects approach. Initial cross-sectional analysis of each cycle separately showed no relationship between country-level class size and mathematics achievement. However, when examining changes across cycles, an unstandardized regression coefficient of -4.4 was found, meaning that a decrease in average class size of seven students was associated with a 31 point increase in average student achievement. Since one international standard

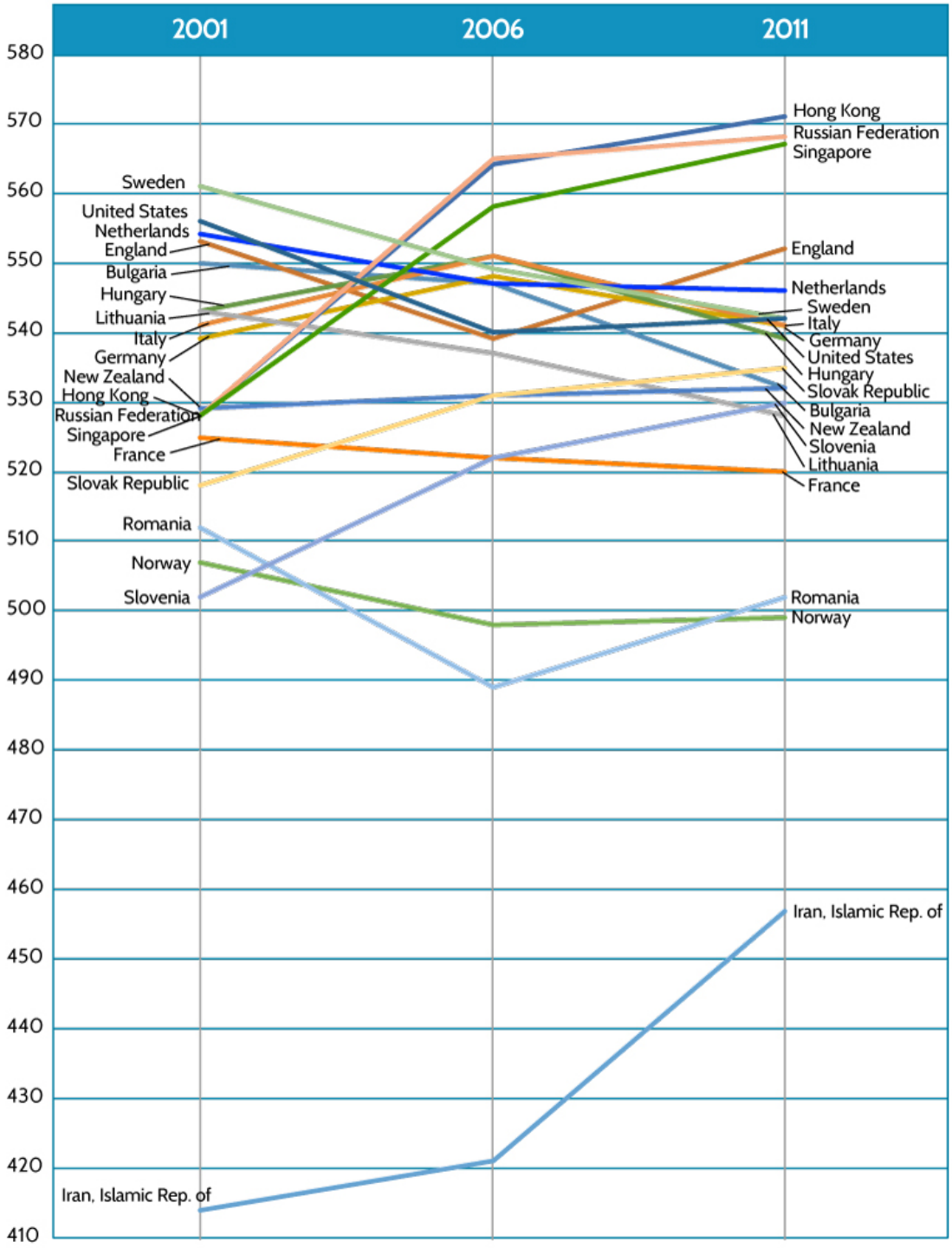
deviation in TIMSS 1995 was equal to 100 points on the TIMSS scale, the results align closely with those of the Tennessee Star Experiment.

Based on these results, Gustafsson (2007, p. 60) concludes that the "cross-sectional data yielded biased results, while the results from the longitudinal country-level analyses were reasonable and compatible with results obtained in studies using other methodological approaches." At the same time, Gustafsson cautioned that this approach has its limitations as well, as it can be difficult to control for influence of time-varying covariates through the country-level fixed-effects approach. In difference-in-differences, the potential effect of time-varying covariates is negated through the common trend assumption—however, as is discussed in the coming section, it is difficult to fulfill the common trend assumption when analyzing international comparative assessment data.

**Common Trends and Difference-in-Differences**

The strongest assumption behind causal inferences from difference-in-differences is that there are no other time-varying variables omitted from the analysis that could have caused the results. In classical difference-in-differences analysis, this assumption is fulfilled by showing common trend. Common trend is a difficult assumption to fulfill using large-scale assessment data because trends tend to fluctuate across countries. Figure 2.8 shows the trends of the 18 countries that participated in PIRLS in 2001, 2006, and 2011. As can be seen in the graph a number of countries increase and others decline but there does not seem to be parallel trend lines across countries. Thus, the common trend assumption that underpins causal inference for difference-in-differences does not seem to hold across the PIRLS countries.

**Figure 2.8: Trend Lines in Reading Achievement across PIRLS 2001, 2006, and 2011**




Given that parallel trend is difficult to assume, researchers can strengthen the assumption that there are no time-variant omitted variables by including numerous covariates in the model

representing rival hypotheses. However, given the small sample size, which is equal to the number of trend countries, including numerous covariates is not always sensible because it limits the statistical power of the analysis—each covariate means losing a degree of freedom.

The problem of sample size and the consequential loss of statistical power results from the aggregated design of the country-level difference-in-differences analysis. In the context of international large-scale assessments, each country collects data from at least 4,000 students across 150 schools per cycle. To perform country-level difference-in-differences, for each cycle these data are combined into one country-level mean for the achievement outcome and one country-level mean for each explanatory variable. As detailed in Chapter 3, the country-level analysis for this dissertation combines data from nearly 190,000 students across 21 countries into just 2 measurement points on each variable for each of the 21 countries—losing a lot of data and statistical power in the process.

**Change Over Time in the Explanatory Variables**

An assumption of the difference-in-differences approach, and other fixed effects models, is that the explanatory variable changes over time. Variables that are supposedly time-varying but show little change over time have been called "sluggish" by Wilson and Butler (2007). Sluggish explanatory variables should not be analyzed using the fixed-effects difference-in-differences approach because they could lead the researcher to make Type II error—incorrectly retaining the null hypothesis. Gauging the sluggishness of an explanatory variable can be challenging for researchers as random sampling error and measurement error can add random variation to essentially unchanged variables, masking the presence of a sluggish variable.

It should be noted that fixed-effects approaches do not assume changes in the outcome variable over time. From a causal perspective, changes in an explanatory variable and no change to the outcome variable would imply no relationship between the two variables—meaning the researcher would retain the null hypothesis, and these results would provide evidence that the explanatory variable is not causally related to the outcome variable.

**Issues Related to Aggregation**

In addition to the statistical power issues linked to aggregation, aggregation can also lead to conceptual problems when the level of analysis does not match the level of interpretation. It is generally considered best practice that the level of analysis corresponds with the research question, and multilevel analysis be employed when analyzing cross-level research questions.

Nevertheless, to find stable units that allow for longitudinal analysis of international large-scale assessment data, researchers employing country difference-in-differences aggregate data to country-level to examine research questions about student-level or classroom-level relationships. For example, Gustafsson (2007) examines the relationship between class size and mathematics achievement through an analysis of country-level relationships.

Using country-level data to draw inferences about lower-level relationships makes difference-in-differences susceptible to aggregation effects where one draws incorrect inferences about lower-level relationships using data aggregated to higher-levels. Robinson (1950) first identified this effect when he examined the relationship between the percentage of African-Americans in a region and percentage of people who are illiterate. Aggregating data to regional level, he found a correlation of 0.95 whereas using individual-level data he only found a correlation of 0.20. In another analysis, he found that the correlation between being foreign born

and illiterate was -0.11 when computed at individual level and 0.53 when computed at state level In this seminal paper, Robinson concluded that correlations at aggregate levels (i.e., ecological correlations) differ from those at lower levels of aggregation.

One conceptual issue related to ecological correlations is that examining aggregate data can sometimes hide the true individual-level relationship. Essentially, the process of aggregation has the potential to provide misleading results because it homogenizes the data by subsuming the individual heterogeneity into mean estimates for group. As an example, in Gustafsson's (2007) eighth grade class size analysis it is unclear whether class size reduction in Latvia, Lithuania, Korea, and Romania was uniform across the population of students or was implemented in some schools and classrooms more than others. As such, it is difficult to know whether the students who benefitted from the class size reduction policy when compared to their peer group from the previous cycle were the same students whose achievement increased.

In analysis of cross-sectional data, one reason why researchers often see different regression coefficients upon aggregation is that grouping criteria is often related to the dependent variable independent of the explanatory variable, and this can lead to biased estimates of regression coefficient(s) (Hannan & Burstein, 1974; King, 1997). In contrast, the regression coefficients remain unbiased when grouping is either random or solely related to the independent variable (only related to the dependent variable through its relationship with the independent variable). King (1997) recommends grouping based on the independent variable because it is considered much more efficient than grouping randomly.

Equation 2.6 shows the basic regression model:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad \textbf{Equation 2.6}$$

Where $Y_i$ is the value on the outcome variable for person *i*, $\beta_0$ is the intercept term, $X_i$ is the value on the explanatory variable for person *i*, $\beta_1$ represents the relationship between $X_i$ and $Y_i$, and $e_i$ represents person *i*'s deviations on $Y_i$ not explained by $X_i$. As summarized by King (1997), aggregation bias arises when grouping occurs on $Y_i$ because through the aggregation process, the relationship between $X_i$ and $Y_i$ is confounded with the relationship between $e_i$ and $X_i$.

The aggregated version of Equation 2.6 can be seen in Equation 2.7 below.

$$\bar{Y}_g = \beta_0 + \beta_1^* \bar{X}_g + e_g \qquad \textbf{Equation 2.7}$$

Where $\bar{Y}_g$ represents each group's mean score on the outcome variable, $\bar{X}_g$ represents each groups's mean score on the explanatory variable, $\beta_1^*$ represents the relationship between $\bar{X}_g$ and $\bar{Y}_g$ when the data are analyzed at aggregate level, and $e_g$ represents the residual term.

Returning to the example of class size at aggregate levels, in an idealized situation the grouping variable can create uniform groups based on class size. In this hypothetical scenario, one group could be composed of students in classes with 30 students, another group in classes with 31 students, and another group in classes with 32 students, etc. In this scenario, each group

would be homogenous on $X_i$ and the estimation of $\beta_1^*$ would provide an unbiased estimate of $\beta_1$ (King, 1997).

In another hypothetical scenario, grouping could capture all of the heterogeneity in $Y_i$, so that one group is composed of all of the students with a score of 531 and another group is composed of all of the students with an average score of 530. In this scenario, there is homogeneity within each group on $Y_i$ and likely heterogeneity on $X_i$ within each group. As synthesized by King (1997), those with high values $Y_i$ would likely have high values on both $X_i$ and $e_i$ and similarly those with low values on $Y_i$ would likely have low values on $X_i$ and $e_i$, and this correlation between $X_i$ and $e_i$, would lead to biased estimates of $\beta_1^*$.

The third possibility for cross-sectional aggregation is to the group based on a random variable or any variable that is orthogonal to both $X_i$ and $e_i$. In this scenario, as group sizes decrease and conversely the number of groups increase the estimates of $\beta_1^*$ will converge to the estimates of $\beta_1$.

Transferring these lessons to the longitudinal context for analysis of international large-scale assessment data, the challenge is to use a variable or set of variables for grouping that at student-level are related to change in the explanatory variable and only related to change in student achievement through this change in the explanatory variable. When the grouping variable is related to changes in achievement independent of changes in the explanatory variable, a correlation between changes in the explanatory variable and the error term may lead to aggregation bias.

In country difference-in-differences, grouping is based solely on the student's country. As such, aggregation bias is introduced when being from a particular country predicts differences in student achievement not explained by the explanatory variable. Estimating the degree of aggregation bias through a longitudinal horizontal approach is complex, but it seems unlikely in most analyses that the country the students' reside would predict student-level changes on the explanatory variable without having a statistical association with the error term—obviously, the amount of aggregation bias would depend on the explanatory variable included in the analysis. As covered in the coming chapters, finding a suitable grouping variable is also a problem for the subpopulation approach—in the international large scale assessment context it is difficult to find demographic variables that are related to the explanatory variable and orthogonal to the error term.

**Structural Equation Modeling Approach to Difference-in-Differences**

A recent extension to Gustafsson's (2007) difference-in-differences approach is the adaptation of Allison and Bollen's (1997) structural equation modeling approach to apply difference-in-differences to analysis of international large-scale assessment data. Gustafsson and Nilsen (2016) used country-level data to examine the relationship between changes in instructional practices across TIMSS 2007 and TIMSS 2011 and changes in achievement across 38 trend countries. For each of the instructional practices they examined, they conducted analyses through both fixed-effects and random-effects models and compared the differences in coefficient estimates as well as differences in model fit. Overall, Gustafsson and Nilsen (2016) found that teachers' education level and their professional development have significant effects on mathematics achievement.

Gustafsson and Nilsen (2016) also conducted multiple group analysis, examining whether any of the 21 measures of instructional quality had a differential effect across OECD and non-

OECD countries. Similar to analyzing an interaction effect, the Mplus multiple group function when applied to estimating differences in path coefficients examines variation in the slopes of the regression lines across groups. Gustafsson and Nilsen (2016) conducted a test of the null model where the OECD and non-OECD countries had coefficients that were constrained to be equal by comparing fit statistics between this null model and an alternative model where the path coefficients were allowed to vary across OECD and non-OECD countries. A $\chi^2$ difference test was conducted to test whether the model with the varying slopes fit significantly better than the constrained model. Gustafsson and Nilsen (2016) did not find any significant differences between OECD and non-OECD countries on any of the instructional practices.

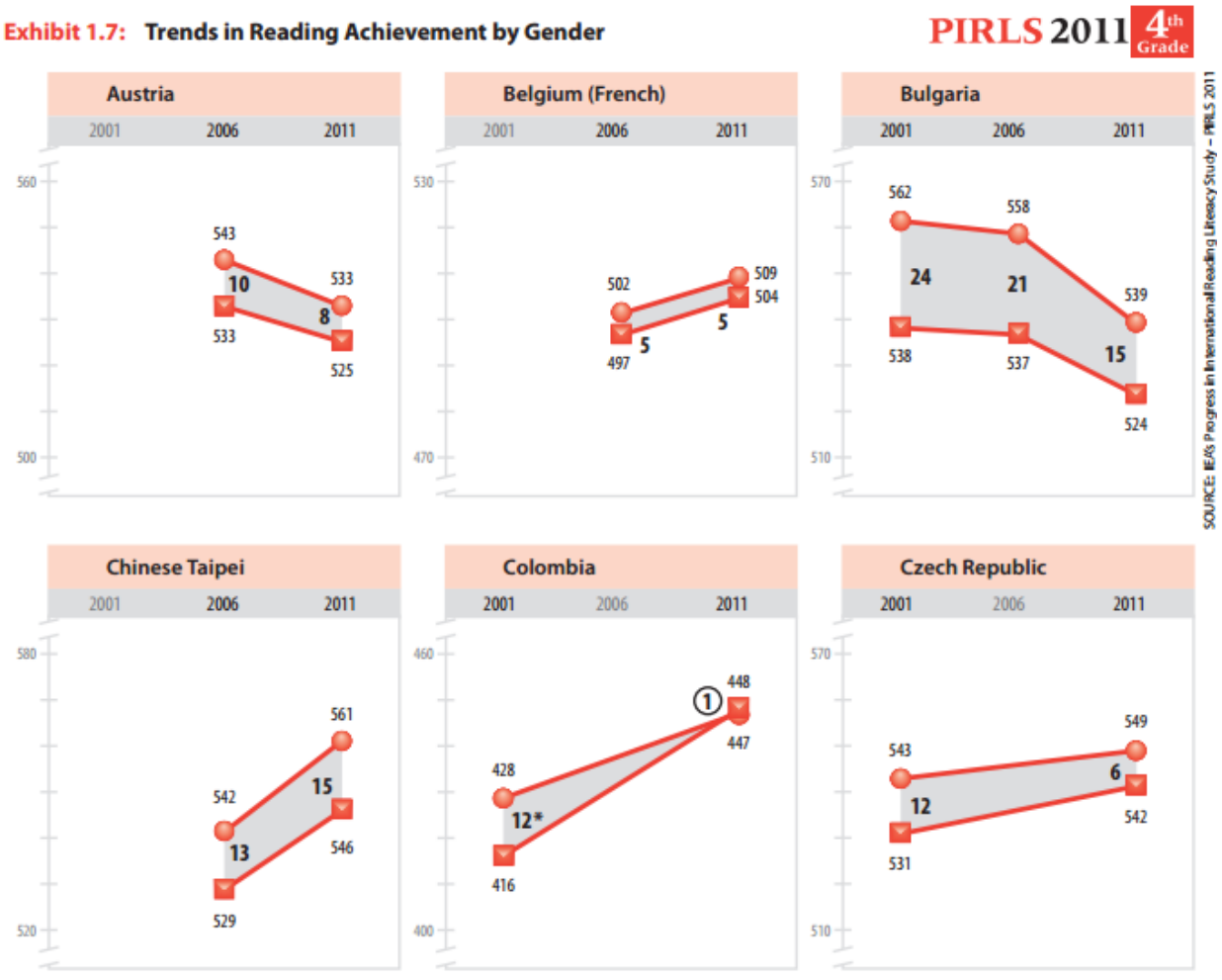
Gustafsson and Nilsen's (2016) multiple group approach is especially pertinent to the subpopulation approach because it provides a way to examine subgroup differences through a fixed-effects approach. As explained previously, in fixed-effects analysis it is not possible to estimate a coefficient representing group effects, because group effects are collinear to the fixed effects. However, it is still possible in the fixed-effects approach to estimate differences in relationships between variables across groups. Gustafsson and Nilsen's (2016) application of the multiple group analysis, which extended the Allison and Bollen (1997) methodology, allows for analysis of between-group differences in the slopes of the fixed-effects regression lines.

## 2.5 Proposed Subpopulation Approach

Building upon the country-level difference-in-differences methodology, this dissertation proposes a subpopulation approach for analyzing international large-scale assessment data. Just as the PIRLS methodology can provide accurate estimates of country achievement and background trends over time, it can also provide accurate estimates of subpopulation achievement trends over time. As an example, Figure 2.9 shows another excerpt from the *PIRLS*

*2011 International Results in Reading* report (Mullis et al., 2012, p. 55)—this exhibit illustrating trends in achievement over time for girls and boys. In the figure, girls are represented by the circles and boys are represented by the squares. For Colombia, the star around the 2001 figure signifies that the reading achievement gap decreases between 2001 and 2011, meaning that boys have closed the reading gap in Colombia since 2001. In this example, the unit of analysis is a subpopulation within each country—e.g., boys in Colombia, girls in Colombia, boys in Austria, girls in Austria.

**Figure 2.9: Excerpt from *PIRLS 2011 International Results in Reading*, Exhibit 1.7: Trends in Reading Achievement by Gender**




Source: Mullis et al. (2012, p. 55). Copyright © 2007 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Because it is possible to accurately estimate changes in subpopulation scores over time, it also becomes possible to model these changes through fixed-effects or random-effects difference-in-differences approaches. The idea is that by dividing countries into subpopulations allows for new analysis applications that explore between-group differences such as multiple-group analysis and even mediation modeling. Because the subpopulation approach can capture some of the within-country heterogeneity in the data, examining student-, school-, or classroom-level research questions using subpopulation data also allows for a closer correspondence between the unit of analysis and the research question than the country-level approach.

**Econometric Theory behind the Subpopulation Approach**

The pseudo-panel approach was first proposed in the econometrics literature by Angus Deaton (1985). Deaton (1985, p.109) referred to these subpopulations as "cohorts," and defined a cohort "as a group with fixed membership, individuals of which can be identified as they show up in the survey."[7] In creating these subpopulations, Deaton (1985) argued that researchers should identify important differences in the population at the individual level—differences that are lost upon further aggregation—and researchers should use these variables to group individuals into subpopulations. According to Deaton (1985, 1997), the pseudo-panel approach is ideal in situations when individual-level longitudinal data are unavailable, and he argued that such data may be preferred over actual individual-level longitudinal data since individual data often suffers from the effects of attrition.

[7] Since "cohort" has a distinct meaning in the context of education, to avoid confusion this dissertation refers to Deaton's "cohorts" as "subpopulations" or "subgroups."

**Implementing the Subpopulation Approach**

In pseudo-panel analysis, the researcher uses criteria to divide the population into subgroups. As described by Verbeek (2008a), the variables used to create the subpopulations should be time-invariant so that individuals from one subpopulation cannot move between subpopulations over time. The subpopulation criteria must also be exhaustive so that each individual is assigned to at least one subpopulation and mutually exclusive so that each individual is solely assigned to one subpopulation. Verbeek (2008a) notes that assignment variables used to create the subpopulations are often demographic variables such as age, gender, or location. For example, in the seminal application of this approach, Browning, Deaton, and Irish (1985) created their subgroups based on age of the head of household and whether the head of household worked as a "manual" or "nonmanual worker." In so doing, Browning et al. (1985), assumed that type of work was time-invariant.

In choosing the variables for creating the subpopulations, researchers are advised to maximize heterogeneity between subgroups and minimize heterogeneity within subgroups (Baltagi,1995), prioritizing variables that are associated with changes in the explanatory variables. If it is possible to assume that each subpopulation is homogeneous, at least with regard to the explanatory time-varying covariate, then the coefficient estimates from the regression analysis should be similar to those estimated if the actual individuals were tracked over time.

Although the econometrics on pseudo-panel approaches does not go into detail on how to be avoid aggregation bias *per se*, it has been argued that subpopulation identifiers should fulfill the conditions of an instrumental variable—be correlated with the explanatory variable and uncorrelated with the error term (Moffitt, 1993; Verbeek, 2008b). These instrumental variable

conditions correspond to the grouping conditions necessary for avoiding aggregation bias in the cross-sectional context.

A final requirement for subpopulation identifiers in the pseudo-panel approach is that they should be able to produce subpopulations of approximately the same size—meaning each individual has approximately the same probability to be placed in each of the subpopulations (Verbeek & Nijman, 1992). However, Verbeek & Nijman (1992) acknowledge that differences in sample size across subpopulations can be counterbalanced through the use of weights.

A variable commonly used in pseudo-panel studies is year of birth/age (Baltagi,1995; Blundell, Megir, & Neves, 1993; Browning et al., 1995; Moffitt, 1993) . Such studies typically take a vertical approach to their longitudinal analysis and examine how cohorts grouped based on birth year progress over time. The authors contend that this variable representing year of birth/age is correlated with the explanatory variable because it represents a cohort effect (Verbeek, 2008b).

A noticeable divergence between the pseudo-panel literature and the literature on grouping in the cross-sectional context is on the use of variables orthogonal to both the explanatory variable and the outcome variable. Verbeek (2008b, p. 376) argues for using variables correlated to the explanatory variable as subpopulation identifiers to ensure that there is sufficient variation in the explanatory variable to conduct the analysis:

> Suppose, as an extreme example, that cohorts are defined on the basis of a variable that is independent of the variables in the model. In that case, the true population means $X_{ct}$ would be identical for each cohort *c* (and equal to the overall population mean) and the only source variation that is left in the data that is not attributable to measurement error would be the variation in $X_{ct}$ over

> time. If these population means do not change over time, all variation in the observed cohort average $\bar{X}_{ct}$ is measurement error.[8]

Verbeek's (2008b) extreme example, however, does not take into account the more likely scenario that grouping the population finely based on a variable that is less relevant to the analysis will converge to the individual estimate of the regression coefficients as the subpopulations become smaller. This divergence in the two strands of the literature could be linked to concern among econometricians that as subsamples become smaller the difference scores that are the foundation of the longitudinal analysis become less reliable.

As described by Baltagi (1995) and Verbeek and Nijman (1992), there is a tradeoff between increasing the number of subpopulations by including numerous demographic characteristics as subpopulation identifiers and ensuring there are sufficient subjects in each subpopulation to provide stable estimates over time. Assuming there is a plentiful supply of subpopulation identifiers, increasing the number of subgroups through the use of time-invariant variables related to the explanatory variable allows the model to capture more of the relevant heterogeneity in the data. Nevertheless, the more fine-grained the stratification of the subpopulations, the sample size in each subpopulation generally becomes smaller and consequently the aggregated mean estimates corresponding to each subpopulation at each time point become less stable—producing imprecise coefficient estimates.

Since Deaton's (1985) paper, there have been concerns that the subpopulation means may be unstable—adding additional error to the data. Because the same individuals are not actually

[8] In the pseudo-panel literature, the error associated with not sampling the same individuals at each time point is referred to as measurement error. However, it seems more like sampling error as the error is not directly attributable to the instrument and measurement that is taking place.

tracked over time, but rather these subpopulations are tracked, mean estimates become less precise as subpopulation sample sizes decrease. Deaton (1985) proposed an error-in-variables estimator, which does not assume a large sample per subpopulation, but it has been rarely applied by researchers (Verbeek, 2008b).

Building on the work of Deaton (1985), Verbeek and Nijman (1992) found bias related to this error was minimal when there were at least 100 subjects in each subpopulation, and many applied researchers use this as a rule of thumb for creating subpopulations for their analysis. However, there is some disagreement in the literature about subpopulation size with Devereux (2007) contending that each subpopulation should include around 2000 or more individuals.

**Pseudo-Panel Analysis Applied to International Large-Scale Assessment Data**

The proposed subpopulation approach examines the relationship between changes in an explanatory variable and changes in an outcome variable across subpopulations and countries horizontally at the same grade level.

Up to this point, the only variant on the pseudo-panel approach applied to international large-scale assessment data examines only one country's data using fourth grade TIMSS and PIRLS data as a control in analysis of TIMSS eighth grade data or PISA data for fifteen-year-olds (De Simone, 2013; Choi, Gil, Mediavilla, and Valbuena, 2016a, 2016b).

De Simone (2013) analyzed Italian TIMSS 2003 data at the fourth grade and Italian TIMSS 2007 data at the eighth grade to examine the extent to which boys advantage over girls at eighth grade in mathematics and science can be explained by differences in their prior achievement at the fourth grade. De Simone (2013) also investigates whether differences between foreign-born and Italian-born students and differences between students of higher and

lower socioeconomic status can be explained by differences that were already present at the fourth grade. As such, De Simone (2013) analyzes the secondary school (grades 5-8) contribution to these gaps after controlling for projected student prior achievement at the fourth grade.

De Simone (2013) follows a two-stage imputed regression strategy. For the first stage, the author takes advantage of TIMSS items that measure time-invariant characteristics and repeat across the fourth grade and eighth grade questionnaires, and uses TIMSS 2003 data to predict the fourth grade TIMSS scores for each eighth grade student. The items used to predict the fourth grade scores are: student sex (male v. female), student birth location (born in Italy v. born abroad), age child moves to Italy, parents birth location (born in Italy v. born abroad), region of residence, and books in the home. Basically, the predicted score from the regression analysis is the mean score at the fourth grade for a student in a particular subpopulation.

In the second stage, these predicted fourth grade scores are matched with each eighth grade student based upon each student's demographic variables. A variable representing this imputed fourth grade score is added as a control variable in the analysis of the eighth grade data. After adding the predicted fourth grade score as a control, De Simone (2013) examines fluctuations in achievement by gender, socioeconomic status, and birth location to draw inferences about whether between-group differences in eighth grade achievement have been exacerbated or mitigated over the course of secondary school. The author concludes that boys' advantage in mathematics at eighth grade was present at fourth grade, but their advantage in science becomes larger between fourth and eighth grades. De Simone (2013) also finds that socioeconomic differences are exacerbated between fourth and eighth grade, but the foreign-born disadvantage decreases between fourth and eighth grade.

In a similar study, Choi et al. (2016a) examine the effect of grade retention on PISA 2012 reading achievement of Spanish students after controlling for prior achievement using a projected score on PIRLS 2006. Similar to De Simone (2013), Choi et al. (2016a) use demographic and household variables that are common across PIRLS and PISA to estimate prior PIRLS 2006 achievement for each subgroup. They then convert both the PISA and PIRLS scores into international z-scores and subtract the projected PIRLS score for each Spanish student from their PISA score. The Choi et al. (2016a) regression model examines whether grade retention predicts higher PISA scores than PIRLS scores after controlling for a number of covariates. Choi et al. (2016a) confirm that grade retention in secondary school has a negative relationship with achievement even after controlling for prior achievement.

In a separate paper, Choi et al. (2016b) examine when gender, socioeconomic, and regional achievement gaps begin in Spain. The authors link PIRLS 2006 scores for Spanish students with PISA 2012 scores by following the two-stage approach of first predicting PIRLS 2006 achievement for each PISA participant based on common demographic information across PIRLS and PISA, and then using the predicted PIRLS scores as a control in the PISA analysis. They find that educational inequality begins primarily at the lower grades but increases through secondary school.

**Proposed Methodology**

Up to this point, the application of the Deaton (1985) pseudo-panel approach to the international large-scale assessment context has been limited. Using international large-scale assessment data, a variant on the technique has been applied for analysis of one-country's data, hierarchically using the fourth grade results as a control in analysis of eighth grade TIMSS achievement or 15-year-old PISA achievement (De Simone, 2013; Choi et al., 2016a, 2016b).

The proposed subpopulation approach expands upon the Gustafsson horizontal approach to difference-in-differences, and allows researchers to analyze data at a lower level of aggregation than country-level. The approach is especially relevant for researchers looking to draw internationally generalizable inferences to research questions focused on subpopulation differences across countries, such as differences between boy and girls or differences between socioeconomic groups, or researchers looking to analyze data at a lower level of aggregation than country level.

As detailed in Chapter 3, country-identifiers in combination with demographic variables can be used to create the subpopulations and then the mean scores on the explanatory variable and outcome variable can be computed for each subpopulation. Following aggregation, data for each subpopulation are paired across cycles (PIRLS 2001 subpopulation data paired with PIRLS 2011 subpopulation data) so that each subpopulation has an average score in each cycle. In creating the subgroups, special attention should be paid to ensure that subpopulations are of sufficient size for stable estimates of subpopulation means.

Following the pairing of the subpopulations, the data can be treated as longitudinal, and the fixed-effects and random-effects methodologies can be applied, as well as more complex structural equation modeling techniques to identify between subgroup differences. Cluster-robust standard errors, such as those produced through the Huber-White sandwich estimator, should be used since subpopulations are nested within a country and would likely have correlated error terms.

To illustrate the pseudo-panel approach, this dissertation analyzes the relationship between early literacy activities and PIRLS reading achievement. The next chapter details this analysis and Chapter 4 displays the results.

# Chapter 3: Analysis Methodology

To exemplify the advantages of the subpopulation approach to difference-in-differences, this dissertation provides an illustrative analysis, and the methodology for this analysis is explained in this chapter. The analysis focuses on the relationship between early literacy activities and PIRLS reading achievement. The purpose of the illustrative analysis is threefold—to demonstrate the approach, to examine whether the approach is able to capture additional relationships in the data, and to explore new opportunities for longitudinal subgroup analysis available through the approach.

## 3.1 Description of the PIRLS Assessment and Database

The PIRLS assessment is designed to provide country-level comparisons of student achievement in reading at the fourth grade. To accomplish this, sampling and weighting procedures are followed to ensure that the PIRLS sample is representative of the national population as a whole. As described by Joncas and Foy (2012), PIRLS follows a two-stage cluster sampling design, first randomly sampling schools from the country's fourth grade population of schools and then randomly sampling a fourth grade classroom (or two) within the school and assessing all of the students within that classroom. To increase efficiency, PIRLS also stratifies the sample—the use of stratification criteria can aid in ensuring a representative sample and decreasing the sampling error. Stratification is also employed to disproportionately sample certain groups. Given the complex sampling design, it is recommended that PIRLS weights be used in analysis to ensure that the results provide unbiased estimates of the population as a whole.

As described by Martin, Mullis, and Foy (2015), the PIRLS international reports provide policymakers with a complete country-level picture of reading achievement at the fourth grade.

The reports detail student achievement across the two PIRLS reading purposes-reading for literary experience and reading to acquire and use information-and across the two processes-retrieval and straightforward inferencing and interpreting, integrating, and evaluating.

According to Martin et al. (2015), in order to accurately measure the full breadth of these purposes and processes, PIRLS encompasses 10 passages with corresponding achievement items and a total testing time of eight hours. However, testing fourth graders for eight hours is unrealistic in most countries. To minimize the testing time for each individual student, PIRLS utilizes matrix sampling—dividing the test passages and corresponding items into blocks, systematically placing two blocks in each assessment booklet, and then randomly assigning booklets to individual students. Through matrix sampling, each student is only assessed on two of the passages and the corresponding items, requiring a testing time of 80 minutes per student.

As described by Foy, Brossman, and Galia (2013), despite the operational advantages of matrix sampling, it comes at a cost. Because each student only takes a subset of the assessment, it is not possible to estimate the scores of individual students with a high degree of precision. For this reason, PIRLS like TIMSS, PISA, and NAEP employs plausible value methodology, and in lieu of estimating individual scores for each student, five multiple imputed values are generated from his/her estimated ability distribution. Through analysis across all five plausible values, it is possible to accurately estimate population and subpopulation parameters, as well as the level of uncertainty around these estimates.

### PIRLS International Database

One of the primary outputs of each PIRLS assessment is a large international database, which includes both the achievement data and background data associated with each student who participated in the assessment. Within the database, achievement results in the form of five plausible values are provided for each student, as well as the contextual data collected through the context questionnaires. A User Guide accompanying the database details best practices for using the data including how to use plausible values and weighting (Gonzalez & Kennedy, 2003; Foy & Drucker, 2013). The IEA's IDB Analyzer facilitates merging the PIRLS data files and conducting basic analysis through SPSS.

## 3.2 Preparing Data for Analysis

To illustrate the proposed methodology, the dissertation used two PIRLS International Databases, the PIRLS 2001 International Database and the PIRLS 2011 International Database. Based on trend reporting from PIRLS 2011 (Mullis et al., 2012), there were 19 countries and 2 benchmarking participants that have trend between PIRLS 2001 and PIRLS 2011 and administered all of the variables used in this analysis across both cycles. These participants are listed in Table 3.1. The Canadian provinces of Quebec and Ontario were benchmarking participants in both cycles—participating as non-country entities. Since education is centralized at the provincial level in Canada, for the purposes of this research it is assumed that the results from these two provinces are independent—meaning that the provinces are treated as if they are individual countries. The datasets for these 21 entities include 91,834 students in the 2001 sample and 97,799 students in the 2011 sample, for a combined sample of 189,633 students. Because 19 of the 21 entities are countries, when discussing the entities as a group, this dissertation refers to them as "countries."

**Table 3.1: Countries and Provinces Included in the Analysis**

| | | |
|---|---|---|
| Bulgaria | Iran, Islamic Republic of | Russian Federation |
| Colombia | Italy | Singapore |
| Czech Republic | Lithuania | Slovak Republic |
| France | Netherlands | Slovenia |
| Germany | New Zealand | Sweden |
| Hong Kong, SAR | Norway | Ontario, Canada |
| Hungary | Romania | Quebec, Canada |

This dissertation analyzes changes over time between PIRLS 2001 and PIRLS 2011. Because the fixed-effects methodology depends on changes over time in the explanatory variable, it was decided it would be best to allow ten years for changes to occur by analyzing changes between PIRLS 2001 and PIRLS 2011 instead of analyzing changes across a five year interval between PIRLS 2006 and PIRLS 2011.

**Reading Achievement**

Across both PIRLS 2001 and PIRLS 2011, reading achievement was represented by the five PIRLS overall reading plausible values. All five plausible values were used for the analysis to ensure standard errors account for the imputation variance.

**Early Literacy Activities**

The primary variables of interest for the illustrative analysis quantify parents' responses to how frequently they engaged their child in seven early literacy activities before beginning primary school. The data were collected through the PIRLS *Learning to Read Survey*, which is completed by the parents of students taking the PIRLS assessment. Cross-sectional analysis in the PIRLS international reports have long found a relationship between frequency of parents engaging their children in early literacy activities and PIRLS reading achievement (Mullis, Martin, Gonzalez, & Kennedy, 2003; Mullis et al., 2007; Mullis et al., 2012), aligning with a wide breadth of research on the positive association between participating in early childhood learning activities and educational outcomes (Gustafsson, Hansen, & Rosén, 2013; Hart & Risley, 1995; Melhuish et al., 2008; Sénéchal & LeFevre, 2002). Epigenetic research also confirms that engaging children in speech from an early age and reading to children improves their cognitive development and language skills (Nelson & Sheridan, 2011; Weisleder & Fernard, 2013).

Table 3.2 shows the seven items measuring early literacy activities that were included in both the 2001 and 2011 questionnaires. Items include how often parents read books with their children and do other literacy activities including singing songs, telling stories, playing with alphabet toys, playing word games, writing letters or words, and reading aloud signs.

Reliability analysis on the combined dataset, with data pooled across across countries and cycles, showed a Cronbach's α coefficient of 0.75. Principal components analysis was implemented to assess the dimensionality of the response patterns to these items, and the analysis showed evidence that the set of items measured a sufficiently unidimensionality construct. All items had unrotated component loadings on the first component of at least 0.50, with all items

but “sing songs” having loadings above 0.60. Also, the first component explained 40% of the variance.

After combining the student-level data across countries and cycles, *Conquest* software was used to scale the seven items measuring early literacy activities using a one-parameter Partial Credit Model (Masters, 1982). The Rasch infit statistics for each of the items were reviewed, the values (infit < 1.25) confirm alignment between the items and the model. Through scaling, each student was provided a score on a logit scale. Following scaling, the logit values were transformed into z-scores, with an international weighted mean across countries and cycles of 0 and a standard deviation of 1.

**Table 3.2: Items Measuring Early Literacy Activities**

Stem: Before your child began primary/elementary school, how often did you or someone else in your home do the following activities with him or her?

Response categories: Often, Sometimes, Never or almost never

a) Read books
b) Tell stories
c) Sing songs
d) Play with alphabet toys (e.g., blocks with letters of the alphabet)
e) Play word games
f) Write letters or words
g) Read aloud signs and labels

Source: IEA's Progress in International Reading Literacy Study, PIRLS 2001/PIRLS 2011 Home Questionnaire, Copyright © 2001, 2011 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

In the last few years, the PIRLS 2011 early literacy activities scale has been used a number of times in exemplar analysis to illustrate methodologies for analyzing international large-scale assessment data. Punter, Glas, and Meelissen (2016) analyzed the PIRLS 2011 scale, among other scales, to illustrate the authors’ psychometric framework for modeling parental

involvement. Punter et al. (2016, p. 91) concluded that the early literacy activities scale "seems to work identically in a large number of countries and cultures…and meets the minimum standard for a survey" across all PIRLS 2011 countries. Caro (2015) also recently used the 2011 version of this scale to illustrate a technique for causal mediation analysis of international large-scale assessment data. In a similar vein, this scale is used to illustrate the new subpopulation methodology proposed in this dissertation.

**Time-Varying Covariates**

As outlined in Chapter 2, in order to strengthen causal interpretations, researchers should provide evidence that other variables, either measured or unmeasured, did not contribute to the changes in the outcome variable (e.g., reading achievement) that are being attributed to the causal agent (e.g., early literacy activities). For this reason, time-varying covariates measuring duration of preprimary education and parents like reading were included in the regression models. These two variables were chosen for this analysis because they provide alternative early childhood explanations for differences in PIRLS fourth grade reading achievement.

Preprimary education refers to education at ISCED Level 0, or before children begin the first grade of primary school (UNESCO, 2012). This includes any mandatory or optional educational programs provided by educational systems across the world, such as Kindergarten in the United States. Cross-sectional bivariate analysis from PIRLS results reports have consistently shown a positive relationship between duration of preprimary education and reading achievement (Mullis et al., 2012). Similarly, Duncan and Magnuson (2013) conducted a meta-analysis on 84 preschool programs serving disadvantaged youth and found there to be a medium to large effect size across studies on cognitive outcomes, although there was considerable variability in the effect size. Duncan and Magnuson (2013) in their review noted that it was

common for studies to find a fade-out effect, with the cognitive and achievement gains of the preschool program dissipating in elementary school after children return to an environment similar to that of the control group.

PIRLS collected data on duration of preprimary attendance in both PIRLS 2001 and PIRLS 2011. Table 3.3 presents the slightly different response options in PIRLS 2001 when compared with PIRLS 2011 and how they were recoded. Each response option was given a value that corresponds to approximately how many years the child attended preprimary school based on the parents' response. For example, for the response "between 1 and 2 years," the student was given the value of 1.5 years. In order to ensure comparability across cycles, response categories were collapsed so they are equivalent between PIRLS 2001 and PIRLS 2011—for example, "between 2 and 3 years" and "3 years or more" from PIRLS 2011 were collapsed to take on the same value (2.5) as "more than 2 years" from PIRLS 2001.

**Table 3.3: Recoding of Reports on Duration of Preprimary Attendance**

| 2001 response categories | 2011 response categories | Value |
|---|---|---|
| Did not attend | Did not attend | 0 |
| Less than 1 year | 1 year or less | 1 |
| 1 year | | |
| Between 1 and 2 years | Between 1 and 2 years | 1.5 |
| 2 years | 2 years | 2 |
| More than 2 years | Between 2 and 3 years | 2.5 |
| | 3 years or more | |

Another predictor of student achievement is parents like reading. Parents socialize their children to appreciate reading by modeling their interest in reading, and this reading socialization process is crucial to promoting reading in the next generation, both fostering children's motivation to

read as well as their reading achievement (Baker & Scher, 2002; Nagel and Verbood, 2012; Kloosterman, Notten, Tolsma, & Kraaykamp, 2010; Notten and Kraaykamp, 2010). Cross-sectional analysis by Notten and Kraaykamp (2010) found that parental reading predicted future educational attainment. Panel analysis by Nagel and Verbood (2012) also found parental reading habits to be a strong predictor of student reading in high school. These findings align with those reported for PIRLS 2011, where cross-sectional analysis showed a bivariate relationship between parents like reading and reading achievement within each of the 43 PIRLS countries administering this set of items in PIRLS 2011 (Mullis et al., 2012).

Table 3.4 shows the four variables in the parents like reading scale. These four variables were administered unchanged across the two PIRLS cycles. The items solicit information on parental attitude toward reading as well as the general importance of reading in the home. Analyzing the data across countries and cycles, the set of four items had a Cronbach's α reliability coefficient of 0.72. Principal components analysis was also conducted, and it showed unrotated component loadings for the first component above 0.60 for each of the four items, with the first component explaining 54% of the variance in the items. The results showed that the data associated with the items were sufficiently unidimensional to proceed with the scaling.

Like the early literacy activities scale, data associated with the parents like reading items were scaled across countries and cycles through a one-parameter item response theory model, using *Conquest*. Two of the statements (statement a and c) express negative feelings toward reading and were reverse coded prior to scaling. Following the scaling, the Rasch infit statistics were reviewed, and all items had an infit value of less than 1.25, suggesting that the data fit the model. The scale scores were output by *Conquest* on the logit scale and then transformed into a z-score, where the mean across countries and cycles was 0 and the standard deviation was 1.

**Table 3.4: Items Measuring Parent Like Reading**

Stem: Please indicate how much you agree with the following statements about reading.
Response categories: Agree a lot, Agree a little, Disagree a little, Disagree a lot

a) I read only if I have to*
b) I like to spend my spare time reading
c) I read only if I need information*
d) Reading is an important activity in my home

***Reverse Coded**

Source: IEA's Progress in International Reading Literacy Study, PIRLS 2001/PIRLS 2011 Home Questionnaire, Copyright © 2001, 2011 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

### Demographic Variables Used for Creating Subpopulations and as Covariates

Variables indicating the country of the student, the sex of the student, and parents' highest education level were used to create subpopulations and were also employed as covariates in some phases of the analyses. The variable measuring sex of the student is available in the sampling data, and the highest level of education of each parent was computed using information on parental education of each parent, as reported by parents through the PIRLS *Learning to Read Survey*. These two variables were recoded into three internationally comparable levels—College Degree, High School Degree, and No High School Degree, and a new variable was created for the analysis that takes the value of the highest level of the two parents.

According to Verbeek (2008b), researchers should ensure that subpopulation identifiers are relevant—meaning that they are statistically related to the explanatory variables in the model. Ideally, subpopulation identifiers should be related to changes in the explanatory variable of interest, which in this case is early literacy activities, and they should be unrelated to the error term in the model. As choosing subpopulation identifiers is arbitrary, it is important for

researchers to provide a theoretical background for reasons why particular subpopulation identifiers were chosen for the analysis.

It was decided that the subpopulations would be created based on country, student gender, and highest parental education level. In the context of international large-scale assessments, it is difficult to find demographic variables related to changes in early literacy activities yet unrelated to the error term. Because all of these subpopulation identifiers likely have a relationship with the error term, it is assumed that some degree of aggregation bias is present in the analysis.

Country of origin appears like a natural subpopulation identifier as students from the same country have much in common. Their parents also would likely have a lot of similarities in their propensity for engaging children in early literacy activities.

Student sex has been shown to have a relationship with engagement in early literacy activities. Gustafson et al. (2013) analyzing TIMSS/PIRLS 2011 data found that parents engage girls more than boys in early literacy activities before beginning primary school, and the authors found these differences in participation in such activities may mediate some of the gender reading gap. Gustafson et al.'s (2013) findings align with those of Bertrand and Pan (2013), who analyzed panel data from the Early Childhood Longitudinal Study: Kindergarten Cohort (ECLS-K), and found that parents tend to read more to girls than boys. Millard (2003) posits that young boys view literacy activities as feminine and resist participating in such activities, but Millard (2003) also concedes that parents may have lower expectations for boys to excel in literacy activities, leading to parental reluctance to engage them in such activities.

There is also a strong theoretical foundation supporting the hypothesis that parental education levels is relevant for examining differences in early literacy activities. Gustafson et al. (2013) also found that parents with more education report engaging their children more often in early literacy activities, corroborating research by Hart and Risley (1995) and Rowe (2012) who found that parents with higher socioeconomic status also speak more with their children and utilize larger vocabularies.

**Weighting**

Analyses used the Senate weights (SENWGT) included in the PIRLS datasets. The Senate weights provide each country a weight of 500 for each cycle (1000 across PIRLS 2001 and PIRLS 2011). In the aggregated analysis, weights were used in the grouping process to ensure that the means associated with the aggregate groupings were unbiased representations of the student populations composing the subpopulations or countries. In addition, the weights were used in the multilevel analysis in Phase 1 to ensure that countries with larger samples were not overly influential on the results. As recommended by Veerbeek and Nijman (1992), weights were linked to each subpopulation to ensure that smaller subpopulations were not overly influential on the results—this is especially important given the extensive research on this approach that documents small sample bias caused by smaller subpopulations. To compute the weights for the subpopulations, the student-level Senate weights were summed across students during the aggregation process, and this sum became the weight for the subpopulation.

**Missing Data**

Table 3.5 details the extent of missing data in PIRLS 2001 and PIRLS 2011. As can be seen in the table, student sex has a negligible amount of missing values in PIRLS 2001 and no missing values in PIRLS 2011, but the other variables have around 10% of their data missing across the cycles, with the parental education variable slightly above 10%.

**Table 3.5: Amount of Missing Data for Each Explanatory Variable**

| | PIRLS 2001 | | PIRLS 2011 | |
|---|---|---|---|---|
| | Number of Students=91,834 | | Number of Students=97,799 | |
| Variable | Number Missing | Percent Missing | Number Missing | Percent Missing |
| Sex of Student | 275 | 0.3% | 0 | 0.0% |
| Parental Education | 13007 | 14.2% | 12685 | 13.0% |
| Early Literacy Activities | 7814 | 8.5% | 9681 | 9.9% |
| Parents Like Reading | 8209 | 8.9% | 10225 | 10.5% |
| Duration of Preprimary Attendance | 8344 | 9.1% | 10595 | 10.8% |

Missing data can bias parameter estimates. Given that this analysis depends on the assumption that subgroups are comparable over time, to ensure the validity of the inferences drawn from the study it is important that the potential bias linked to missing data be minimized. For these reasons, multiple imputation was used to impute missing values. Along the same lines as the plausible value methodology, multiple imputation uses observed data to predict missing data, and the methodology recognizes the uncertainty in this prediction process by providing numerous plausible estimates for each predicted value, with the variance among the estimates representing the level of uncertainty in the prediction process.

Multiple imputation was conducted using the Marcov Chain Monte Carlo (MCMC) algorithm in SPSS. The MCMC algorithm follows a sequential process in which already imputed data are used to predict subsequent imputations. The imputation analysis was conducted separately for each cycle and for each country, meaning there were 42 separation imputation models. The prediction equation for the multiple imputation analyses included each of the five variables being imputed, as well as a number of auxiliary variables. The auxiliary variables were chosen because they predict PIRLS reading achievement, the values of the variables being imputed, or the missingness of the variables being imputed. To decide which auxiliary variables to include in the prediction equation, stepwise regression was conducted on each of the analysis variables using each cycle's pooled background data across countries to identify the auxiliary variables that predict the value or missingness of the variables in the analysis. This stepwise analysis identified 57 auxiliary variables from the home, student, and school questionnaire data for PIRLS 2001 and 42 variables from these questionnaires for PIRLS 2011. For each cycle, the same auxiliary variables were used in the prediction equation across countries.

Five separate imputation datasets were generated and each was linked with one of the five plausible values measuring reading achievement. All phases of the analysis were conducted using these five imputation datasets following Rubin's (1987) method for pooling parameter estimates across plausible values and estimating the standard error across the imputation datasets. SPSS and Mplus both include a feature for analyzing imputation datasets that uses Rubin's (1987) method.

**Significance Testing**

In 2016, the American Statistical Association released a statement criticizing the research community's overreliance on p-values. The authors of the statement, Wasserstein and Lazar (2016, p. 131) acknowledged that p-values can be helpful for understanding the "statistical incompatibility of the data with the null hypothesis" but warned that conclusions and decisions "should not be based only on whether the p-value passes a specific threshold." Wasserstein and Lazar (2016) went on to clarify that the p-value is not a measure of effect size and is not a good measure of evidence for or against a model.

Given the well-documented issues associated with solely depending on p-values in the statistical decision making process, this dissertation follows a wider interpretative approach when analyzing the results. In examining coefficient estimates, the dissertation examines not only whether the p-value is less than 0.05 but also the size of the coefficient estimates. In aggregated analysis, it is common to see large coefficient estimates accompanied by also large standard errors, and in these situations this dissertation does not dismiss the relationship represented by the coefficient because of the coefficient's nonsignificance. Instead the large relationship between the explanatory and outcome variable is acknowledged as well as the level of uncertainty around the coefficient estimate.

For assessing model fit, the three model fit indices that were used are the $\chi^2$ test of model fit, the root mean square error approximation (RMSEA), and the comparative fit index (CFI). A nonsignificant value for the $\chi^2$ test of model fit ($p>0.05$) is generally considered to be signal of acceptable fit as is an RMSEA value less than .05 and a CFI value above .95. The decision-making process therefore does not just depend on the $\chi^2$ significance test but also an assessment of these other fit indices.

## 3.3 Analysis Overview

Analysis was conducted in seven phases. The purpose of the first four phases is to examine whether regression coefficients measuring the same relationship and using the same data are different across cross-sectional and longitudinal approaches and across different levels of aggregation. Phase 1 and Phase 2 provide cross-sectional results at student- and country-level, respectively. Phase 3 and Phase 4 analyze data longitudinally, with Phase 3 following Gustafsson's (2007) country-level difference-in-differences approach, and Phase 4 conducting difference-in-differences on subpopulation data and thereby demonstrating the subpopulation approach.

Supposing the regression coefficients are different to some extent, a limitation of Phases 1 through 4 is that it is not possible to examine which approach comes close to the true effect of early literacy activities on PIRLS reading achievement—since the true effect is unknown in the PIRLS data.

As such, the case for the subpopulation approach rests on two arguments:

(1) Including subpopulation data provides important information that strengthens the analysis; and

(2) Analyzing subpopulation data provides opportunity for answering research questions about subgroup differences.

Phase 5 and Phase 6 examine the evidence that the subpopulation data add relevant information to the analysis, and Phase 6 and Phase 7 explore new opportunities for analysis of subgroup differences.

**Phase 1: Multilevel Cross-Sectional Analysis on Each Cycle of PIRLS data**

The purpose of Phase 1 is to provide initial estimates of the early literacy activities regression coefficients that can be viewed as baselines for comparisons with the longitudinal models that follow. In this first phase, a three-level random effects model was conducted for each cycle with individual student data being the primary analysis unit. Because students are nested within schools and schools are nested within countries, it is best practice to implement a three-level model. As such, level one is student-level, level two is school-level, and level three is country-level. Because the purpose of the Phase 1 model is to provide a baseline for comparisons with Phases 2 through 4 and those phases only include student characteristics as covariates, only Level 1 covariates are included in the analysis. The basic three-level model is specified as follows:

$$Y_{ijk} = \beta_{000} + \sum_{q=1}^{Q} \beta_{qjk} (X_{qijk}) + \mu_{00k} + r_{0jk} + e_{ijk}, \qquad \textbf{Equation 3.1}$$

$$i = 1, ..., I; \quad j = 1, ..., J \quad k = 1, ..., K \quad q = 1, ..., Q$$

Where $Y_{ijk}$ represents the reading achievement of student $i$ in school $j$ in country $k$ (student $ijk$); $\beta_{000}$ represents the adjusted grand mean across schools and countries; $X_{qijk}$ represents student $ijk$'s value on the student characteristic $q$ variable. Across the five models of Phase 1, the student characteristics include scores on the early literacy activities and parents like reading scales, duration of preprimary attendance, gender, and two binary variables representing highest parental education level—one variable representing highest parental educational level as having a parent who graduated from high school and another representing highest parental educational level as having a parent who graduated from college. The symbol $\beta_{qjk}$ represents the relationship between the student characteristic $q$ and (adjusted) reading achievement in school $j$

of country *k*. The regression coefficients associated with $\beta_{qjk}$ are fixed for all Q variables and therefore not allowed to vary across schools or countries. $e_{ijk}$ is a residual term representing student *ijk*'s deviation from his/her expected score, $r_{0jk}$ is a residual term representing school *j* in country *k*'s deviation from its expected score, $\mu_{00k}$ represents country *k*'s deviation from its expected score. The residual terms $e_{ijk}$, $r_{0jk}$, and $\mu_{00k}$ are assumed to have a mean of 0 and be normally distributed.

Five cross-sectional models were estimated for each cycle. Analyses were conducted in Mplus using Senate weights. Table 3.6 provides the student-level explanatory variables included in each model. The first model included early literacy activities alone at level one and provides a baseline estimate of the relationship between early literacy activities and reading achievement when analyzed at student-level for each cycle. The second model analyzed the relationship between student gender and reading achievement, and the third model analyzed the relationship between early literacy activities and reading achievement while controlling for gender. Evaluating the changes in the coefficient estimates for early literacy activities and gender across the first three models provides an idea of the relationship between these variables, hinting to whether early literacy activities may explain the relationship between gender and reading achievement—the focus of the longitudinal analysis in Phase 6. The fourth model controlled for the covariates parents like reading and duration of preprimary attendance in estimating the relationship between early literacy activities and reading achievement, and, in addition to the covariates from the fourth model, the fifth model also controls for gender and binary variables representing highest parental education. The regression coefficients in the fourth and fifth models provide baselines for comparisons with the longitudinal fixed-effects models that follow

in Phases 3 and 4, which also explicitly control for parents like reading and duration of preschool attendance, and implicitly control for gender and highest parental education, which are considered time-invariant covariates in the fixed-effects approach.

**Table 3.6: Explanatory Variables Included in Each Phase 1 Model**

| | Student-Level Explanatory Variables Included in Each Model |
|---|---|
| Model 1 | Early Literacy Activities |
| Model 2 | Gender |
| Model 3 | Early Literacy Activities, Gender |
| Model 4 | Early Literacy Activities, Parents Like Reading, Duration of Preprimary Attendance |
| Model 5 | Early Literacy Activities, Parents Like Reading, Duration of Preprimary Attendance, Gender, Parent Graduated from High School, Parent Graduated from College |

**Phase 2: Country Cross-Sectional Analysis of Each Cycle of PIRLS Data**

The Phase 2 analyses provide country-level estimates of the relationships between early literacy activities and reading achievement. Using PIRLS Senate weights, each of the early childhood explanatory variables and each of the PIRLS reading plausible values were aggregated to country-level for each of the cycles. Two regression analyses were conducted. For each cycle separately, the first model regresses the average reading achievement for each country on average early literacy activities scores and the second model adds the time-varying covariates average parents like reading and average duration of preprimary attendance to the analysis. The regression models are represented by the following equation:

$$\bar{Y}_c = \beta_0 + \sum_{q=1}^{Q} \beta_q (\bar{X}_{qc}) + e_c \qquad \textbf{Equation 3.2}$$

$$c = 1, \ldots, C; \quad q = 1, \ldots, Q$$

Where $\bar{Y}_c$ represents country $c$'s mean reading achievement, $\beta_0$ is the intercept term, $\bar{X}_{qc}$ is country $c$'s mean on each of the $Q$ predictors—early literacy activities, parents like reading, and

duration of preprimary attendance, $\beta_q$ represents the relationship between predictor $\bar{X}_{qc}$ and a country's (adjusted) average reading achievement, and $e_c$ is the residual term and is assumed to be normally distributed.

**Phase 3: Country Difference-in-Differences Analysis**

In Phase 3, country difference-in-differences analysis was conducted using the fixed-effects approach, as formulated by Gustafsson (2007). As described in Chapter 2, Gustafsson (2007, 2010, 2013) and others in the IEA research community (Liu et al., 2014; Rosén, & Gustafsson, 2014) typically use a fixed-effects approach to estimate difference-in-differences. Phase 3 analysis applied the first-difference approach. As can be seen in Equation 3.3, for each country, the PIRLS 2001 achievement mean was subtracted from the PIRLS 2011 achievement mean and likewise the PIRLS 2001 early literacy activities mean was subtracted from the PIRLS 2011 early literacy activities mean. A difference score was also calculated for the covariates, parents like reading and duration of preprimary attendance. Analysis was conducted in SPSS, with the first model regressing changes in average reading achievement on changes in average early literacy activity scores and the second model also including the covariates average parents like reading scores and average duration of preprimary attendance. The models are represented by the following equation:

$$(\bar{Y}_{2c} - \bar{Y}_{1c}) = \beta_0 + \sum_{q=1}^{Q} \beta_q (\bar{X}_{q2c} - \bar{X}_{q1c}) + e_c \qquad \textbf{Equation 3.3}$$

$$c = 1, \ldots, C;\ q = 1, \ldots, Q$$

Where $\bar{Y}_{2c}$ is the average reading score for country *c* in PIRLS 2011, $\bar{Y}_{1c}$ is the average reading score for country *c* in PIRLS 2001, $\beta_0$ is the intercept term, $\bar{X}_{q2c}$ is the average score on

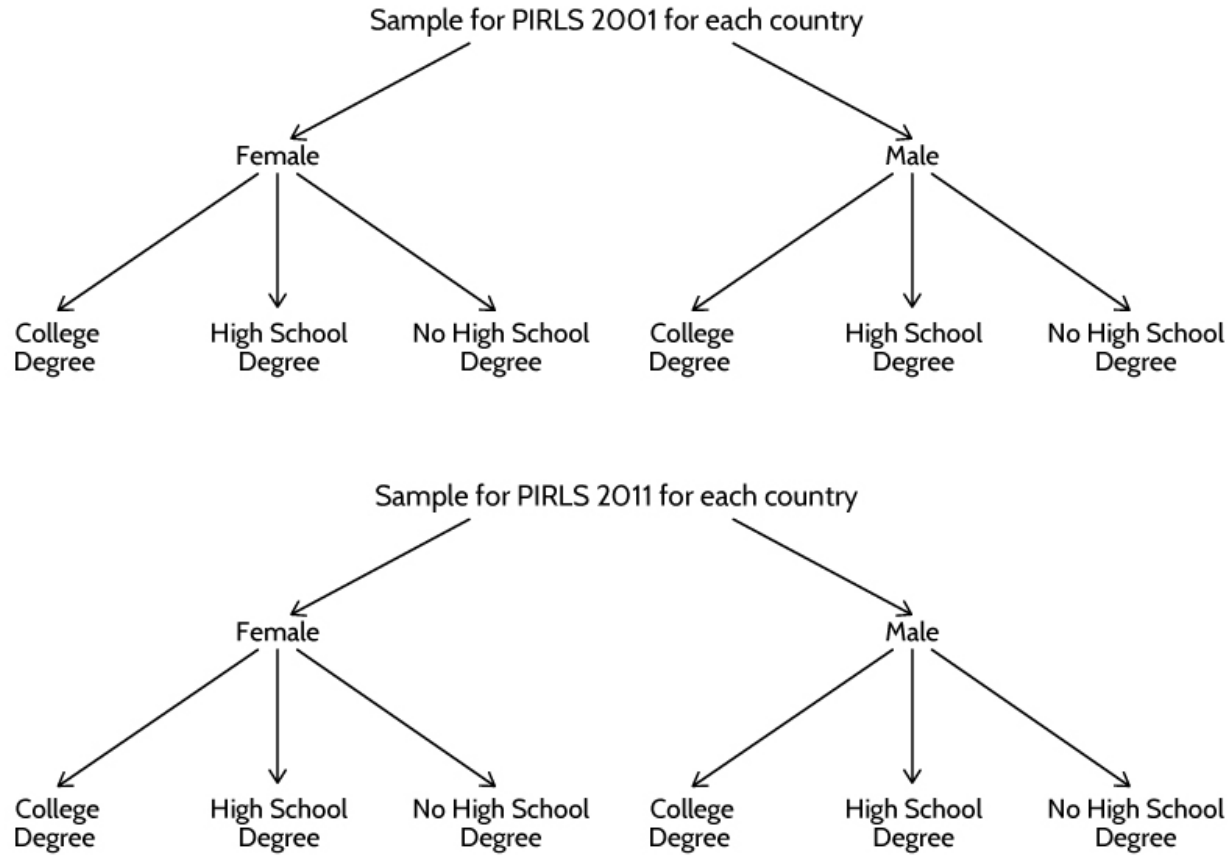
variable $q$ for country $c$ in PIRLS 2011, $\bar{X}_{q1c}$ is the average score on variable $q$ for country $c$ in PIRLS 2001, $\beta_q$ represents the (adjusted) relationship between the average student characteristic difference score and the average reading achievement difference score, $e_c$ represents the residual term associated with country $c$ and is assumed to be normally distributed.

**Phase 4: Subpopulation Difference-in-Differences Analysis**

Phase 4 of the analysis entails conducting fixed-effects difference-in-differences analysis on subpopulation data. First, subpopulations were created, and then analysis was conducted through the fixed-effects approach. Two models were analyzed—the first model examines the relationship between early literacy activities and PIRLS reading achievement, and the second model analyzes this relationship while including the covariates duration of preprimary attendance and parents like reading.

**Creating the Subpopulations.** To create the subpopulations, students were grouped based on the students' country and two demographic characteristics—gender (female, male) and highest parental education (college degree, high school degree, no high school degree). Because there are two gender and three highest parental education categories, each student was classified in one of six subpopulations across the 21 countries—meaning in total there are 126 subpopulations with data collected at two time points for each subpopulation. Figure 3.1 provides an illustration of the classification rules. For example, within each country and for each cycle (e.g., PIRLS 2001), there is one group of students who are female and whose highest level of parental education is a college degree and another group that is composed of male students whose highest level of parental education is a college degree, and so on and so forth.

**Figure 3.1: Classification Rules for Creating Subpopulations within Each Country for Each Cycle**




As described in Chapter 2, there is some debate in the literature about the number of respondents that should comprise each subpopulation. Verbeek and Nijman (1992) recommend at least 100 students in each subpopulation to provide stable estimates of each subpopulation's mean scores. Following classification of students into their subpopulations, the mean sample size of each subpopulation averaged across the two cycles was reviewed. The review showed that each subpopulation on average across the cycles had at least 75 respondents, and all but six of the subpopulations had more than 100 students, on average. Because weighting would be used in the analysis, it was decided to proceed with the analysis using all 126 subpopulations. Due to the weighting, these smaller subpopulations have less influence on the results because they have proportionally smaller weights than larger subpopulations. In addition, outliers were reviewed

for each analysis to flag subpopulations that were overly influential on the results, and therefore if one of these six subpopulations were highly influential they would be flagged through this analysis of outliers.

To prepare for the analysis, for each country the subpopulations with the same demographic characteristics were paired across cycles. For example, within each country the 2001 subpopulation composed of female students whose highest parental education level is a college degree was paired with the 2011 subpopulation of female students whose highest parental education level is a college degree, and the 2001 subpopulation of male students whose highest parental education level is a college degree was paired with the 2011 subpopulation of male students whose highest parental education level is a college degree. The pairing involved restructuring the dataset so that subpopulations with the same characteristics in PIRLS 2001 and PIRLS 2011 would have one entry in the dataset and that entry would include two values on each time-varying variable, one value for PIRLS 2001 and another for PIRLS 2011. This pairing process was followed for all of the subpopulations in the analysis.

**Fixed-Effects Analysis.** After pairing these subpopulations across cycles, the fixed-effects approach to difference-in-differences examined whether subpopulation changes in the early literacy activities scores between PIRLS 2001 and PIRLS 2011 are related to changes in reading achievement. A second analysis examined this relationship after controlling for duration of preprimary attendance and parents like reading. The analyses were conducted in Mplus. Because subpopulations are nested in countries, the Huber-White "sandwich" procedure associated with the TYPE=COMPLEX function was used in these analyses to ensure cluster-robust estimates of standard errors. As can be seen in Equation 3.4, the Phase 4 analysis was conceptually similar to

the Phase 3 analysis but involved subpopulations instead of countries—to note this, the subpopulations are subscripted with *s*:

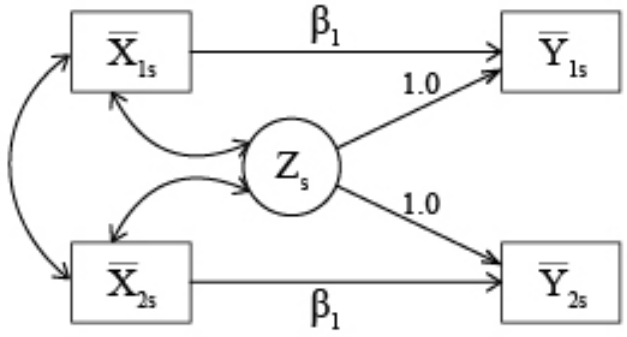
$$(\bar{Y}_{2s} - \bar{Y}_{1s}) = \beta_0 + \sum_{q=1}^{Q} \beta_q (\bar{X}_{q2s} - \bar{X}_{q1s}) + e_s \qquad \textbf{Equation 3.4}$$

$s = 1, \ldots, S; \quad q = 1, \ldots, Q$

In Phase 3, the structural equation modeling approach to fixed-effects analysis was also introduced by conducting parallel fixed-effects analysis on the relationship between early literacy activities and PIRLS reading achievement. Presenting the structural equation modeling approach in parallel with the classic regression-based approach demonstrates that both methods produce identical unstandardized regression coefficients and standard errors. This fixed-effects SEM model is subsequently used as a baseline in deciding whether the random effects model can be implemented in Phase 7 without unduly biasing the coefficient estimates. Figure 3.2 shows the path model for the fixed effects analysis.

**Figure 3.2: Path Model for the Subpopulation Fixed-Effects Analysis**

$$\bar{Y}_{ts} = \beta_{t0} + \beta_1 \bar{X}_{ts} + Z_s + e_{ts}$$




Source: Modified from Allison and Bollen (1997, p.6)

Where $\bar{X}_{1s}$ represents subpopulation *s*'s average early literacy activities score in PIRLS 2001, $\bar{X}_{2s}$ is subpopulation *s*'s average early literacy activities score in PIRLS 2011, $\bar{Y}_{1s}$ represents subpopulation *s*'s average reading achievement score in PIRLS 2001, $\bar{Y}_{2s}$ is subpopulation *s*'s average reading achievement score in PIRLS 2011, $Z_s$ is a latent variable representing the fixed effect for subpopulation *s*—its relationship with $\bar{Y}_{1s}$ and $\bar{Y}_{2s}$ is constrained to be 1 and therefore not estimated in the model, and $\beta_{t0}$ is the intercept term estimated for $\bar{Y}_{ts}$ at each time point.

**Phase 5: Analysis of Within-Country Relationships**

The purpose of Phase 5 is to delve deeper into the subpopulation data to examine the extent to which the subpopulation data within countries adds relevant information to the analysis.

Subpopulations are nested in countries, and therefore the data could be conceptualized as multilevel, with subpopulations at level 1 and countries at level 2. Phase 5 analysis focused on only the subpopulation-level data, and followed four analysis stages:

(1) **Estimating the intraclass correlation coefficient.** As an initial step, the amount of variance at subpopulation- and country-level was quantified by estimating the intraclass correlation coefficient (ICC) for both early literacy activities and PIRLS reading achievement. If all of the variance is at country-level, analyzing subpopulation-level data would not benefit the analysis.

(2) **Dispersion analysis.** Scatterplots of the relationship between changes in average early literacy activities on the x-axis and changes in PIRLS reading achievement on the y-axis were examined to understand the within-country dispersion in the explanatory variable. One of the assumptions of the pseudo-panel approach is that the

subpopulation identifiers are able to create subpopulations that have distinct changes on the explanatory variable. If there are no differences between subpopulations within a country on early literacy activities scores, then the subpopulation-level data provides redundant information. One scatterplot presents all of the countries color-coded for identification purposes, another scatterplot highlights the five countries with the largest standard deviations between the subpopulations on the explanatory variable, and the third scatterplot highlights the five countries with the smallest standard deviations on the explanatory variable.

(3) **Within-country relationship between changes in mean early literacy activities scores and changes in mean PIRLS reading achievement.** A potential argument against the subpopulation approach is that the data are error-ridden at levels of aggregation below country-level. Devereux (2007) argued that subpopulations need around 2000 respondents to avoid small sample bias, but most follow the rule of thumb that 100 respondents per subpopulation is sufficient to mitigate small sample bias. If there is a high-level of error in the subpopulation data, a relationship with achievement would not be found at subpopulation level.

To investigate the within-country relationship between early literacy activities and PIRLS reading achievement, a regression line was estimated using data from the six subpopulations. Because this line was drawn based on only six points it is not reliable for any one country. However, keeping this in mind, it is still informative to explore patterns across countries. A relationship between early literacy activities and PIRLS reading achievement across many of the countries would suggest that the data are not solely random error but is able to capture substantial signal.

(4) **Cross-country analysis of subpopulation-level variance.** A final stage of the analysis examines the subpopulation-level relationship between early literacy activities and PIRLS reading achievement across countries to explore whether a relationship can be found when analyzing solely subpopulation-level data. The purpose of the analysis is to examine whether a relationship exists across countries without taking into account the country-level variance. In this stage, the variance was decomposed by centering subpopulation difference scores on average country difference scores. To center the data, the difference between each subpopulation's average early literacy activities difference score ($\Delta\bar{X}_s$) and the country's average difference score ($\Delta\bar{X}_c$) was computed, and similarly the difference between each subpopulation's average reading achievement difference score ($\Delta\bar{Y}_s$) and the country's average difference score ($\Delta\bar{Y}_c$) was calculated. The country-centered subpopulation difference scores corresponding to reading achievement $(\Delta\bar{Y}_s - \Delta\bar{Y}_c)$ were then regressed on the country-centered difference scores for early literacy activities $(\Delta\bar{X}_s - \Delta\bar{X}_c)$:

$$(\Delta\bar{Y}_s - \Delta\bar{Y}_c) = \beta_0 + \beta_1 (\Delta\bar{X}_s - \Delta\bar{X}_c) + e_s \qquad \textbf{Equation 3.5}$$

$$s = 1, \ldots, S; \quad c = 1, \ldots, C$$

Where $\beta_0$ represents the intercept, $\beta_1$ represents the subpopulation-level relationship between the country-centered difference scores for early literacy activities and the country-centered difference scores for reading achievement, and $e_s$ represents subpopulations *s*'s residual term.

**Phase 6: Comparisons of Fixed-Effects Coefficient Estimates across Groups**

The purpose of Phase 6 analysis is twofold:

(1) To examine whether there is heterogeneity in coefficient estimates across subgroups; and

(2) To demonstrate additional analysis opportunities that can be employed through the subpopulation approach.

Ideally, the subpopulation approach would not only be able to capture additional signal in the data, when compared with the country-level approach, but would also be able to capture subpopulation differences in the data that cannot be captured through the country-level approach. Adapting Allison and Bollen's (1997) structural equation modeling approach to difference-in-differences, Gustafsson and Nilsen (2016) recently implemented multiple group analysis in Mplus to test whether coefficients are significantly different across groups of countries. Analysis in this dissertation uses the multiple group approach to investigate whether the relationship between early literacy activities and reading achievement varies between boys and girls and also to examine whether the relationship varies across highest parental education groups. If the relationship varies across subgroups, the analysis provides evidence that there is analysis-relevant heterogeneity in the data that can be captured at subgroup level.

Two separate multiple group analyses were conducted—the first examining differences in coefficient estimates across boys and girls and the second examining whether there are differences across highest parental education subgroups. For each analysis, a null model was estimated where the coefficient estimates are constrained to be equal across the groups, and then a second model was estimated where the coefficient estimates are allowed to vary across the groups. The null hypothesis is that the groups have the same regression coefficient estimates,

meaning the regression lines are parallel, and the alternative hyphothesis is that boys and girls have different regression coefficients—meaning that the regression lines are not parallel. Differences between the models were evaluated through a $\chi^2$-difference test and a comparison of RMSEA and CFI fit indices, as well as through a comparison of coefficient estimates in the alternative model using a Wald test. In contrast to the fit indices that examine model fit comparing the null and alternative models, the Wald test is implemented by comparing the path coefficients within the alternative model. For the highest parental education groups, where there are three groups, the Wald test acts as an omnibus test—evaluating whether the three coefficient estimates are equal.

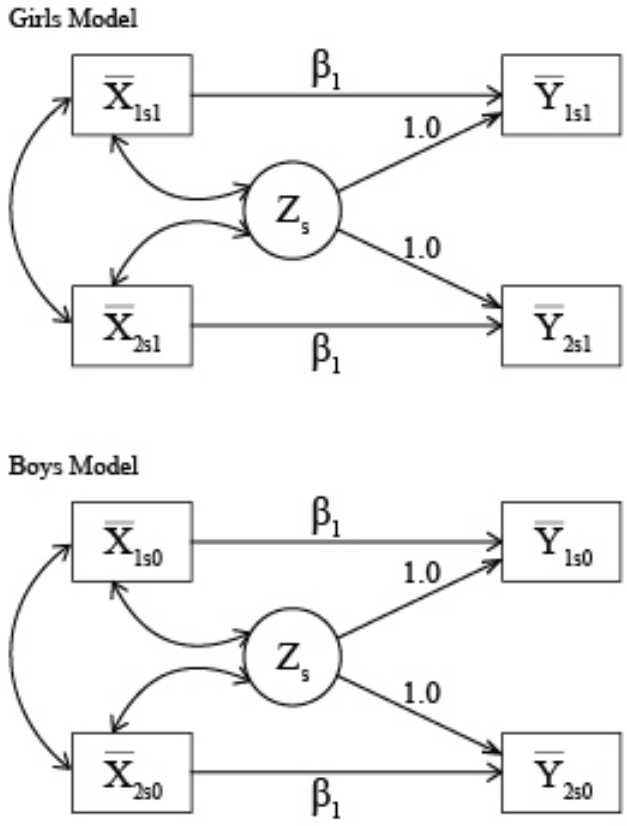
Initial exploratory analysis was conducted using the default Mplus settings, which meant that the intercept at each time point was constrained to be equal across groups for the null and the alternative models. However, constraining the intercept across groups inflated coefficient differences between the groups in the alternative model—differences that were not present when the data were analyzed separately for each group through the fixed-effects approach. Therefore, it was decided that the intercepts would be allowed to vary across groups for the alternative model. By allowing the intercepts to vary, the coefficient estimates in the alternative model equaled the fixed-effects analysis estimates when analysis is conducted separately for each group. Allowing the intercept to vary in the alternative models comes with the cost of fewer degrees of freedom, and consequently the degrees of freedom difference between the null and alternative models increases leading to a less powerful $\chi^2$-difference test.

Figure 3.3 shows the null model. As can be seen in the equation in the figure, the subscript *j* has been added to the fixed-effects models, and one of the diagrams pertains to boys (j=0) and the other diagram pertains to girls (j=1). In the equation and both diagrams for the null

model, $\beta_1$ and $\beta_{t0}$ are not subscripted, meaning that the estimated value is constrained to be equal across both groups.

**Figure 3.3: Null Model for Multiple Group Analysis Where Regression Coefficients are Constrained to be Equal across Boys and Girls**

$$\bar{Y}_{tsj} = \beta_{t0} + \beta_1 \bar{X}_{tsj} + Z_s + e_{tsj}$$



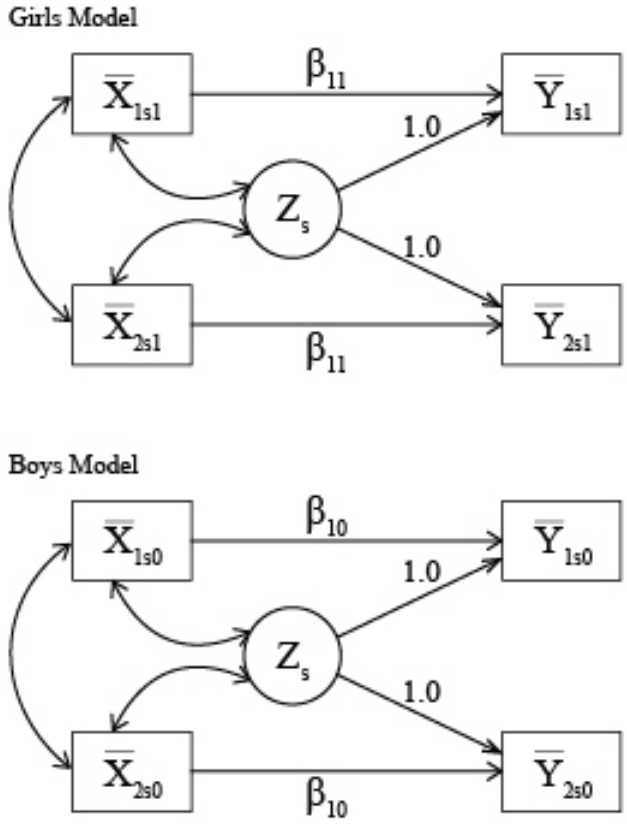

Where $\bar{Y}_{tsj}$ is the average reading achievement of subpopulation *s* composed of *j* gender at time *t* (subpopulation *tsj*), $\beta_{t0}$ is the intercept, $\bar{X}_{tsj}$ is the average early literacy activities score for subpopulation *tsj*, $\beta_1$ is the path coefficient representing the relationship between average reading achievement and early literacy activities across time and groups, $Z_s$ is the fixed effect for

subpopulation *s* representing its time-invariant characteristics, and $e_{tsj}$ is the residual term corresponding to $\bar{Y}_{tsj}$ for subpopulation *tsj*.

Figure 3.4 shows the alternative model. As can be seen in the equation in the figure, the subscript *j* has been added to $\beta_1$ in the equation to become $\beta_{1j,}$ implying that $\beta_1$ is estimated separately across boys and girls. Similarly, the subscript *j* has been added to $\beta_{t0}$ to reflect that the intercept ($\beta_{t0j}$) is allowed to vary across groups.

**Figure 3.4: Alternative Model for Multiple Group Analysis Where Regression Coefficients Vary Across Boys and Girls**

$$\bar{Y}_{tsj} = \beta_{t0j} + \beta_{1j}\bar{X}_{tsj} + Z_s + e_{tsj}$$



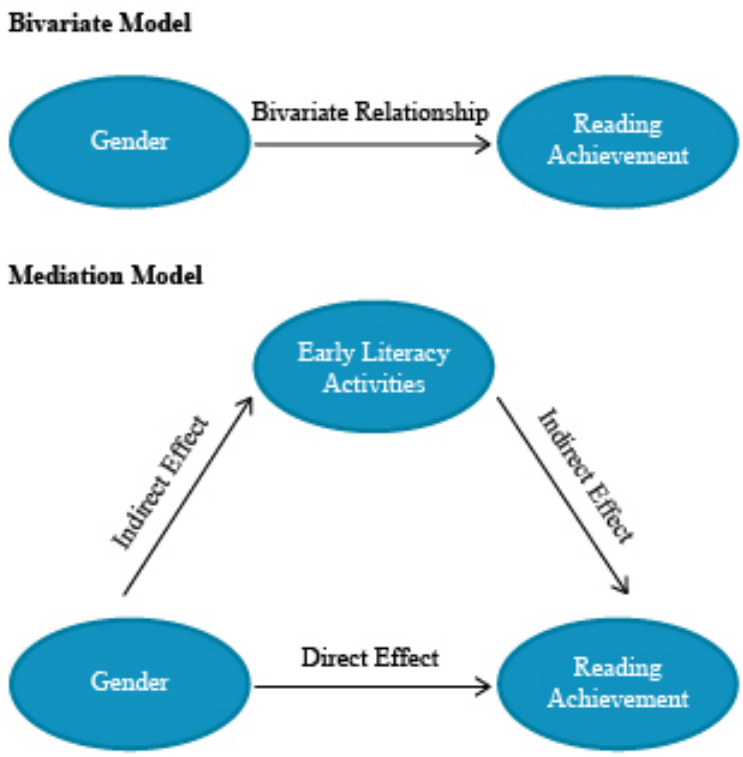
Similar models were conducted to examine differences in coefficient estimates across highest parental education groups. A null model was conducted where the coefficient was constrained to be equal across highest parental education groups, and an alternative model was conducted where the coefficient was allowed to vary for each of the three highest parental education groups.

**Phase 7: Mediation Analysis through a Random-Effects Model**

Another way that subpopulations can be analyzed longitudinally is through more complex models such as mediation models. To illustrate this, the dissertation takes advantage of the structural equation modeling approach to explore to what extent early literacy activities mediates the relationship between gender and reading achievement. Using TIMSS/PIRLS 2011 data, Gustafson et al.'s (2013) analysis suggests that parents' frequency in engaging girls in early literacy activities may mediate girls' advantage on the PIRLS reading assessment. Phase 6 tests whether PIRLS longitudinal data supports this finding.

Mediation models hypothesize a reason for relationships in the data, and test whether the data matches the model based on the correlations in the data. Figure 3.5 shows a conceptual diagram illustrating a bivariate model and a mediation model. The bivariate model illustrates the relationship between gender and reading achievement. In most PIRLS countries, girls have higher reading scores than boys. Girls' reading advantage could have a number of causes. For example, girls may have more innate potential to excel in reading, parents may socialize girls differently than boys, or parents may be more likely to engage girls in early literacy activities, to name just a few. A mediation model hypothesizes an explanation for this relationship and then based on the correlations in the data, the mediation model tests to what extent the relationship from the bivariate model can be explained by the hypothesized mediator.

**Figure 3.5: Conceptual Diagram Illustrating Mediation Modeling**




As can be seen in mediation model in Figure 3.5, the explanation that was tested is that early literacy activities mediates the relationship between gender and reading achievement. In this model, an indirect effect is estimated by regressing early literacy activities on gender and regressing reading achievement on early literacy activities. By regressing early literacy activities on gender, the hypothesis that girls are more likely to participate in early literacy activities is tested. By regressing reading achievement on early literacy activities, the model examines whether reading achievement is related to early literacy activities. Taken together, the model hypothesizes that girls reading achievement advantage at the fourth grade is related, at least partially, to their greater engagement in early literacy activities with their parents before beginning primary school. If the data fits the hypothesized model, it would be expected that the direct effect in the path between gender and reading achievement would decrease and the indirect

effect from gender to reading achievement through early literacy activities would be significant. The indirect effect is calculated by simply multiplying the coefficient of the path from gender to early literacy activities by the coefficient of the path from early literacy activities to reading achievement.

Given the causal focus of the dissertation, it is important to note that mediation models in structural equation modeling make strong assumptions about the directionality of the relationships in the model, and these assumptions cannot be tested within the statistical framework. In this particular case, the directionality of the mediation lies on solid ground since both boys and girls participate in early literacy activities before the fourth grade reading assessment. Nonetheless, an indirect effect through early literacy activities does not necessarily mean that the relationship is causal. Similar to other modeling approaches, one would still need to dismiss all other plausible mediators to fulfill Mill's third condition—a difficult challenge using international large scale assessment data. Ideally, additional covariates would be included in the model to test for alternative explanations for the mediation. However, because this analysis uses cluster-robust standard errors, the number of parameters that can be estimated in the model is limited to 21—number of country clusters included in the analysis.[9] Therefore, it was decided to maintain a more parsimonious mediation model by not including additional covariates.

[9] The mediation model that was implemented (shown in Figure 3.9) estimates 15 parameter and including another time-varying covariate mediator would increase the number of parameters estimated in the model to 26. When the number of parameters exceed the number of clusters, the cluster-robust standard errors may no longer be trustworthy.

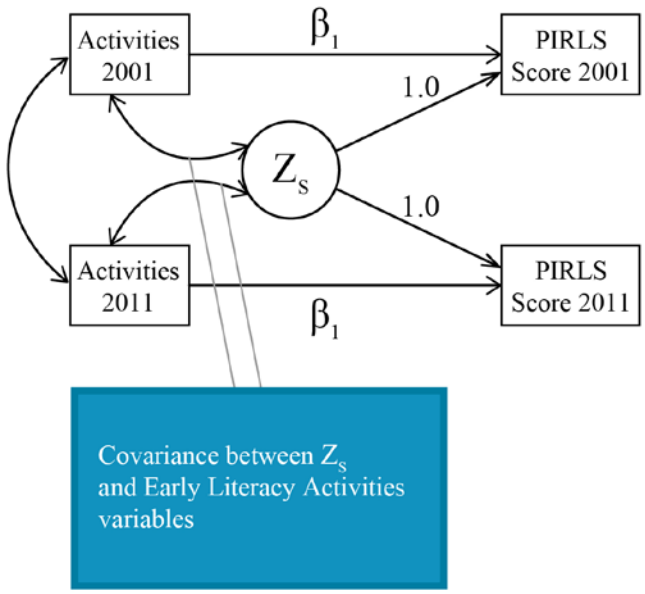
**Transitioning from a Fixed-Effects Model to a Random Effects Model.** Because this analysis would necessitate estimating the relationship between the time-invariant observed covariate gender and reading achievement, it is not possible to estimate this model through the fixed-effects approach—the coefficients associated with the fixed effects would be collinear to the coefficient associated with the time-invariant covariate. As such, to conduct this mediation model, it is first necessary to transition to a random-effects model.

A random-effects model analyzes both cross-sectional relationships in the data as well as longitudinal relationships. As discussed in Chapter 2 (see section 2.2, pp. 19-28), the fixed-effects approach estimates the coefficient of interest solely based on longitudinal relationships—the relationship between changes in the explanatory variable and the outcome variable. In so doing, the model controls for measured and unmeasured time-invariant variables that could bias coefficient estimates.

In the random effects model, in contrast, coefficient estimates are based on both time-varying and time-invariant relationships. As such, it is possible that coefficient estimates could change dramatically when transitioning to a random effects model. Therefore, initial analysis examined whether it is possible to switch to a random-effects model without biasing parameter estimates. In standard approaches, researchers use a Hausman test to examine whether bias is introduced to coefficient estimates when moving from a fixed-effects to a random-effects model. Allison and Bollen (1997) suggest that instead of using the Hausman test to compare coefficient estimates, a more direct approach is to evaluate the covariance between $Z_S$ and early literacy activities in the fixed-effects model as well as to compare fit statistics across the two models.

Figure 3.6 shows the fixed-effects model, pointing out the covariance between the fixed effect $Z_s$ and early literacy activities. This covariance directly examines whether there is a relationship between the fixed effect and the predictor variable. Following from this, in transitioning from the fixed-effects approach to the random-effects approach, this covariance was examined, the fit statistics were compared across the two models, and changes in the magnitude of the coefficient $\beta_1$ were evaluated across the two models.

**Figure 3.6: Path Model for the Fixed-Effects Approach Highlighting the Covariance**

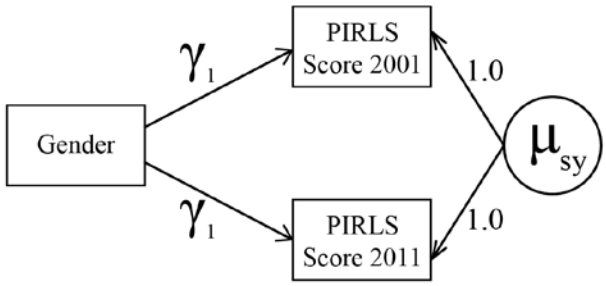



As detailed in Chapter 4, the assumptions were met for transitioning to the random-effects model, meaning minimal relationship between the explanatory variables ($\bar{X}_{ts}$) and the fixed effect, acceptable fit statistics for the random effects model, and mimimal change in the $\beta_1$ coefficient estimates when compared across the models.

**Building the Mediation Model.** As a first step in creating a mediation model, the bivariate relationship between gender and reading achievement from Figure 3.5 was operationalized in the structural equation model shown in Figure 3.7, where the path coefficient ($\gamma_1$) from gender ($W_s$) to reading achievement ($\bar{Y}_{ts}$) is the coefficient of interest. Because the relationship between gender and reading achievement is hypothesized to be time-invariant, and therefore constrained to be equal across time, it can be represented by the single coefficient estimate ($\gamma_1$). Because gender ($W_s$) is a time-invariant observed variable—it maintains the same value over time for each subpopulation, its relationship with reading achievement ($\gamma_1$) would reflect the combined cross-sectional relationship between gender and reading achievement in PIRLS 2001 and PIRLS 2011. Therefore, it is expected that the coefficient estimate for $\gamma_1$ would be similar to estimates from the student-level analyses from Model 2 of the three-level random effects model in Phase 1. $\beta_{t0y}$ represents the intercept corresponding to average reading achievement ($\bar{Y}_{ts}$) at each time point *t*, and $e_{tsy}$ represents the residual term for reading achievement ($\bar{Y}_{ts}$) associated with subpopulation *s* at each time point *t*. $\mu_{sy}$ is a random variable representing subpopulation-specific deviations from the reading achievement model for subpopulation *s.* The *y* subscript is introduced to distinguish the terms related to the model predicting reading achievement ($\bar{Y}_{ts}$), because in subsequent models the mediator early literacy activities $\bar{M}_{ts}$ is added to the model and acts as both a dependent and independent variable.

**Figure 3.7: Path Model Examining the Relationship Between Gender and PIRLS Reading Achievement**

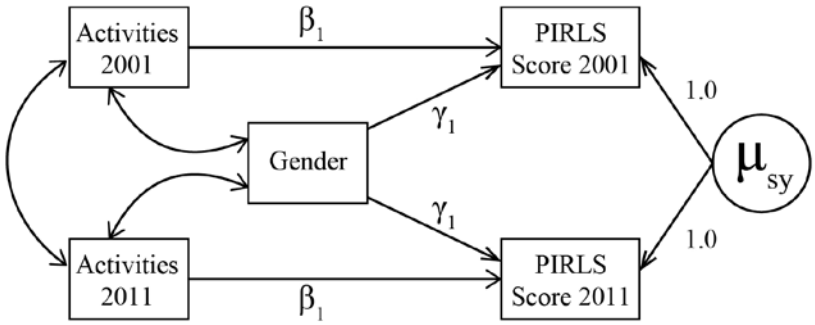
$\bar{Y}_{ts} = \beta_{t0y} + \gamma_1 W_s + \mu_{sy} + e_{tsy}$




As a second step in building the mediation model, early literacy activities was added to the random effects model as a predictor, and this model can be seen in Figure 3.8. In this model early literacy activities ($\overline{M}_{ts}$) and gender both predict PIRLS reading achievement.

**Figure 3.8: Path Model for Predicting PIRLS Reading Achievement with Gender and Early Literacy Activities**

$\bar{Y}_{ts} = \beta_{t0y} + \beta_1 \overline{M}_{ts} + \gamma_1 W_s + \mu_{sy} + e_{tsy}$



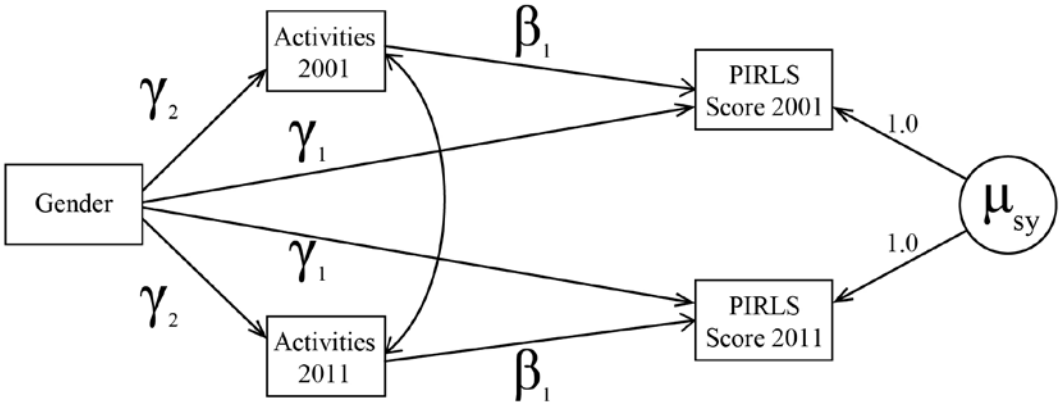

In the above model, $\beta_1$ represents the adjusted relationship between the average early literacy activities score and the average PIRLS reading achievement score, and $\gamma_1$ represents the

adjusted relationship between gender and PIRLS reading achievement. Although the assumptions were fulfilled allowing the transition from a fixed-effects model, $\beta_1$ still measures the combined cross-sectional and longitudinal relationship between early literacy activities and PIRLS reading achievement, and therefore the magnitude of the path coefficient $\beta_1$ can still be affected by covariance between early literacy activities $\bar{M}_{ts}$ and gender $W_s$, and likewise the magnitude of $\gamma_1$ path can also be affected by the covariance between these two predictors. Indeed, if early literacy activities $\bar{M}_{ts}$ mediates the relationship between gender $W_s$ and PIRLS reading achievement ($\bar{Y}_{ts}$), it would be expected that the coefficient $\gamma_1$ would decrease with the inclusion of early literacy activities $\bar{M}_{ts}$ as a predictor of reading achievement.

**Estimating the Mediation Model.** The full operationalization of the mediation model from Figure 3.5 can be seen in Figure 3.9. This mediation model includes the 2001 and 2011 variables measuring average early literacy activities ($\bar{M}_{ts}$) as the mediator of the relationship between gender ($W_s$) and average PIRLS reading achievement ($\bar{Y}_{ts}$).

**Figure 3.9: Path Model for Analyzing the Mediation Effect**

$$\bar{Y}_{ts} = \beta_{t0y} + \beta_1 \bar{M}_{ts} + \gamma_1 W_s + \mu_{sy} + e_{tsy}$$
$$\bar{M}_{ts} = \beta_{t0m} + \gamma_2 W_s + e_{tsm}$$




The coefficient $\gamma_2$ represents whether the propensity to partake in early literacy activities ($\bar{M}_{ts}$) is different across the gender ($W_s$) subgroups, where $e_{tsm}$ is the residual term[10] for subpopulation *s* at each time point *t* associated with early literacy activities ($\bar{M}_{ts}$) and $\beta_{t0m}$ is the intercept for early literacy activities ($\bar{M}_{ts}$) at time *t*. Like the paths between gender and average PIRLS reading achievement, this relationship is hypothesized to be invariant over time and therefore the path coefficients ($\gamma_2$) between gender ($W_s$) and average early literacy activities ($\bar{M}_{ts}$) are constrained to be equal in PIRLS 2001 and PIRLS 2011. The hypothesized indirect effect is equal to the product of $\gamma_2$ and $\beta_1$, with the direct effect of gender on PIRLS reading

[10] As also footnoted on page 26, in the structural equation modeling random effects approach across two time points, it is not necessary to include $\mu_s$ term in the model. The random-effects analysis provides the same coefficient estimates of $\beta_1$ and fit statistics if the dependent variables are allowed to covary. In this case activities 2001 is covarying with activities 2011, but an alternative formulation would be to add a $\mu_{sm}$ to the model to represent this covariance.

achievement represented by $\gamma_1$. Mplus calculates the standard errors for this indirect effect using the delta method (MacKinnon, 2008).

# Chapter 4: Results of Analysis

Analysis was performed across seven phases:

- Phase 1: Multilevel Cross-Sectional Analysis on Each Cycle of PIRLS data
- Phase 2: Country Cross-Sectional Analysis on Each Cycle of PIRLS Data
- Phase 3: Country Difference-in-Differences Analysis
- Phase 4: Subpopulation Difference-in-Differences Analysis
- Phase 5: Analysis of Within-Country Relationships
- Phase 6: Comparisons of Fixed-Effects Coefficient Estimates Across Groups
- Phase 7: Mediation Analysis Through a Random-Effects Model

## 4.1 Phase 1: Multilevel Cross-Sectional Analysis on Each Cycle of PIRLS data

In order to create a baseline for comparisons with subsequent models, multilevel analyses were conducted for each cycle on data pooled across countries. Because students are nested in schools and schools are nested in countries, the three-level models have students at level 1, schools at level 2, and countries at level 3. Preliminary analysis confirmed that it was appropriate to analyze the data across three levels as the PIRLS 2001 intraclass correlation coefficient (ICC) showed that 24% of the variance in PIRLS reading achievement was at school-level and 18% of the variance was at country-level, and in PIRLS 2011 21% of the variance was at school-level and 14% was at country-level.

The analyses for each cycle included a set of five models, with different predictors included in each model. Given the large sample size (N=91,834 in 2001; N=97,799 in 2011), all relationships were found to be statistically significant at the alpha level of 0.05.

The first model examined the relationship between early literacy activities and PIRLS reading achievement. As can be seen in Table 4.1, in PIRLS 2001 there was a regression coefficient of 10.8 (SE=1.06) associated with early literacy activities, meaning a one standard deviation increase in early literacy activities was associated with a nearly 11 point increase in PIRLS reading achievement. For PIRLS 2011, the coefficient associated with early literacy activities was similar at 11.3 (SE=0.95). Across each of the cycles, early literacy activities explained 3% of the student-level variance.

**Table 4.1: Results from the Multilevel Analysis Conducted in Phase 1**

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| **PIRLS 2001** | β (SE) | β (SE) | β (SE) | β (SE) | β (SE) |
| Intercept | 521.6 (7.35)* | 513.2 (7.49)* | 514.8 (7.20)* | 512.9 (7.27)* | 488.0 (7.33)* |
| Early Literacy Activities | 10.8 (1.06)* | | 10.3 (1.04)* | 7.8 (0.91)* | 6.4 (0.83)* |
| Parents Like Reading | | | | 11.1 (0.88)* | 9.1 (0.76)* |
| Duration of Preprimary Attendance | | | | 4.3 (1.03)* | 3.1 (0.95)* |
| Female | | 14.9 (1.28)* | 13.7 (1.20)* | | 14.0 (1.26)* |
| High School | | | | | 19.9 (2.68)* |
| College | | | | | 38.6 (3.76)* |
| Level 1 Variance Explained ($R^2$) | 0.03 | 0.01 | 0.04 | 0.06 | 0.12 |
| **PIRLS 2011** | | | | | |
| Intercept | 528.1 (6.28)* | 523.3 (6.37)* | 523.2 (6.26)* | 515.5 (5.86)* | 493.9 (6.52)* |
| Early Literacy Activities | 11.3 (0.95)* | | 10.9 (0.92)* | 8.1 (0.76)* | 6.9 (0.69)* |
| Parents Like Reading | | | | 11.5 (0.73)* | 9.6 (0.63)* |
| Duration of Preprimary Attendance | | | | 6.6 (1.32)* | 5.3 (1.24)* |
| Female | | 11.4 (1.13)* | 9.8 (1.04)* | | 10.3 (1.06)* |
| High School | | | | | 16.6 (2.74)* |
| College | | | | | 34.6 (3.56)* |
| Level 1 Variance Explained ($R^2$) | 0.03 | 0.01 | 0.04 | 0.07 | 0.12 |

*P=0.05

Model 2 analyzed the relationship between gender and reading achievement, without controlling for other confounding variables. For PIRLS 2001, the results showed a regression coefficient of 14.9 (SE=1.28), meaning girls have an advantage over boys of 15 points, and the results from 2011 show a coefficient of 11.4 (SE=1.13). In both cycles, the gender variable explained 1% of the student-level variance.

Model 3 examined the association between early literacy activities and PIRLS reading achievement after controlling for the effect of gender. In PIRLS 2011, there was a regression coefficient of 10.3 (SE=1.04) associated with early literacy activities and a coefficient of 13.7 (SE=1.20) associated with gender, and in PIRLS 2011 an early literacy activities regression coefficient of 10.9 (SE=0.92) and a gender coefficient of 9.8 (SE=1.04). Because there are multiple predictors in the model, the regression coefficient estimated in Model 3 and subsequent models are partial regression coefficients. The slight decline in the association between gender and PIRLS reading achievement in Model 3, as compared with Model 2, provides evidence of covariance between gender and early literacy activities in their relationship with PIRLS reading achievement—suggesting that early literacy activities may partially mediate the relationship between gender and PIRLS reading achievement. In both PIRLS 2001 and PIRLS 2011, the models explain 4% of the student-level variance.

Model 4 analyzes the relationships between the three early childhood influences and PIRLS reading achievement. With the addition of parents like reading and duration of preprimary attendance, the early literacy activities coefficient declines to 7.8 (SE=0.91) in PIRLS 2001 and 8.1 (SE=0.76) in PIRLS 2011. Parents like reading has significant coefficients of 11.1 (SE=0.88) and 11.5 (SE=0.73) in PIRLS 2001 and PIRLS 2011, respectively, and duration of preprimary attendance has a significant coefficient of 4.3 (SE=1.03) in PIRLS 2001

and a significant coefficient of 6.6 (SE=1.32) in PIRLS 2011. In 2001, the model explained 6% of the student-level variance, and in 2011 the model explained 7% of the student-level variance.

Model 5 includes the home influences with the demographic variables. Adding gender and the parental education variables to the model, the coefficient associated with early literacy activities further declines to 6.4 (SE=0.83) in PIRLS 2001 and 6.9 (SE=0.69) in PIRLS 2011. The parents like reading coefficient declines to 9.1 (SE=0.76) in 2001 and 9.6 (SE=0.63) in 2011, and the duration of preprimary attendance coefficient declines to 3.1 (SE=0.95) in 2001 and 5.3 (SE=1.24) in 2011. The relationship between gender and PIRLS reading achievement remains steady with a regression coefficient of 14.0 (SE=1.26) in 2001 and a coefficient of 10.3 (SE=1.06) in 2011. Having a parent with a high school degree predicted a 19.9 (SE=2.68) point higher score on the PIRLS 2001 assessment and a 16.6 (SE=2.74) point higher score on the PIRLS 2011 assessment, when compared with students whose parents did not have a high school degree. Having a parent with a college degree was associated with a 38.6 (SE=3.76) point higher score in 2001 than that of the students whose parents did not have a high school degree, and a parent with a college degree was associated with a 34.6 (SE=3.56) point higher score in PIRLS 2011. The model explained 12% of the variance in both PIRLS 2001 and PIRLS 2011. It is important to note that the percentage of variance explained at student level by the model is surprisingly small at 12%.

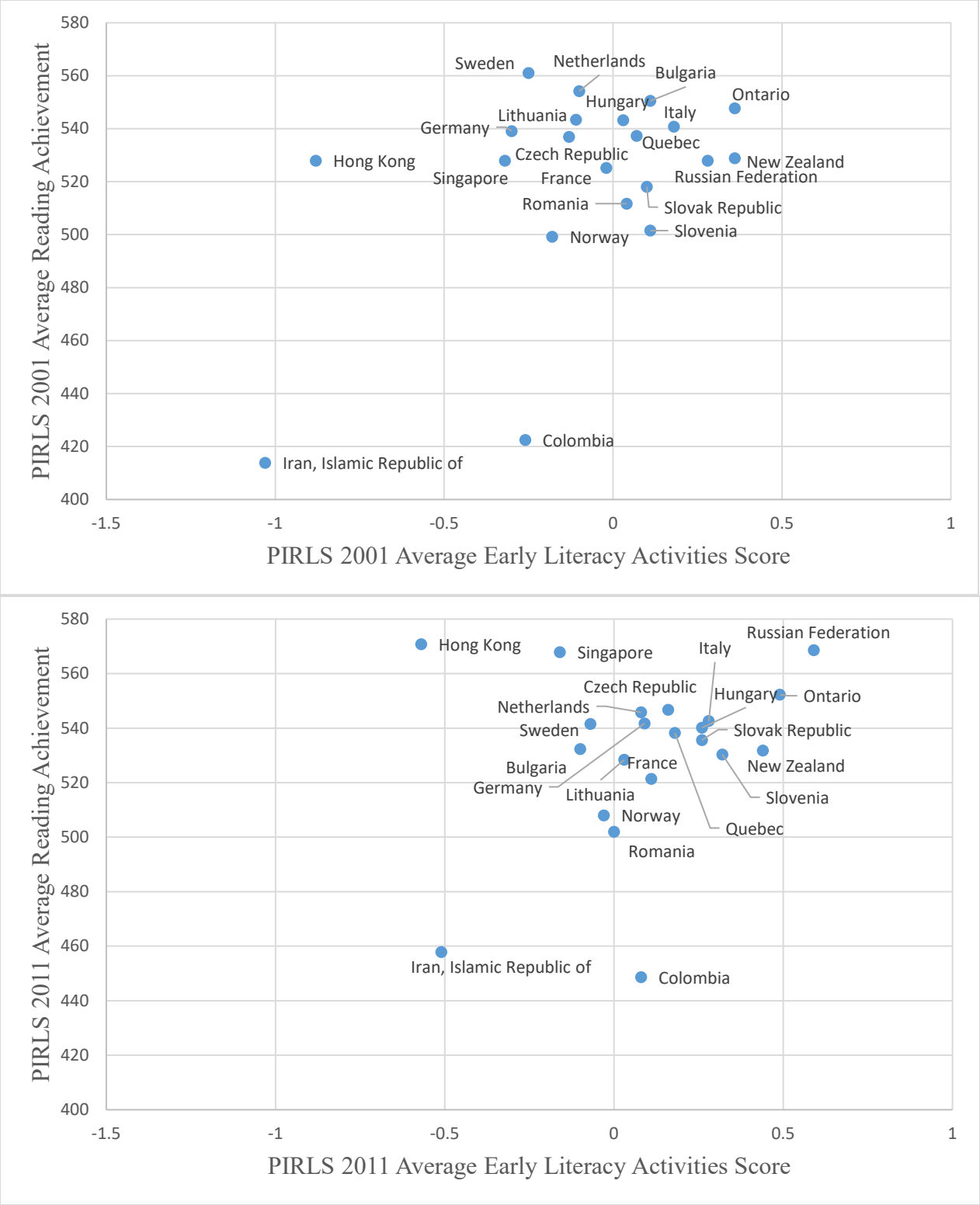
## 4.2 Phase 2: Country Cross-Sectional Analysis on Each Cycle of PIRLS Data

Country-level cross-sectional regression results provide another useful baseline for evaluating the longitudinal difference-in-differences approaches. For this reason, country-level regression analyses were conducted for both the PIRLS 2001 and PIRLS 2011 cycles. Figure 4.1 shows the

country-level relationship between average early literacy activities and average reading achievement for PIRLS 2001 and PIRLS 2011. As can be seen in the graphs, when examined at country-level, there is not a clear visual relationship between early literacy activities and PIRLS reading achievement in either cycle. In PIRLS 2001, a number of high performers like Ontario have high average early literacy activities scores but other high performers like Sweden have lower average early literacy activities scores. Of these countries, Iran has the lowest early literacy activities average as well as the lowest PIRLS reading achievement score. In PIRLS 2011, high performers like Russian Federation and Ontario have high early literacy activities scores, but a number of high achieving countries like Hong Kong and Singapore have lower early literacy activities scores.

Comparing the two graphs, which are on the same scale across cycles, it is noticeable that many countries have shifted to the right in PIRLS 2011—as the mean early literacy activities scores increased across many of the countries between PIRLS 2001 and PIRLS 2011. For example, Iran went from just below -1 in PIRLS 2001 to just below -0.5 in PIRLS 2011. Similarly, the Russian Federation went from an early literacy activities average of 0.3 in PIRLS 2001 to an early literacy activities average of 0.6 in PIRLS 2011. This shift provides evidence that there was enough change in early literacy activities between PIRLS 2001 and PIRLS 2011 to allow for fixed-effects analysis.

**Figure 4.1: Graphs Showing the Country-Level Relationship between Mean Early Literacy Activities and Mean Reading Achievement for Each PIRLS Cycle**



Regression analysis was then conducted. As can be seen in Table 4.2, the regression analysis for Model 6 for PIRLS 2001 showed a significant regression coefficient of 52.7 (SE=21.75) representing the relationship between the average country score on the early literacy activities scale and PIRLS 2001 average reading achievement. With the inclusion of the covariates average parents like reading score and average duration of preprimary attendance in Model 7, the early literacy activities regression coefficient for PIRLS 2001 decreased to a nonsignificant 27.3 (SE=21.46). Parents like reading was associated with a nonsignificant coefficient of 29.4 (SE=24.49) in PIRLS 2001 and duration of preprimary attendance was associated with a significant coefficient of 37.7 (SE=14.30) in PIRLS 2001. For PIRLS 2001, Model 6 explained 24% of the variance and Model 7 explained 49% of the variance.

**Table 4.2: Parameter estimates for Phase 2 Country-Level Cross-Sectional Analysis**

| | Model 6 | Model 7 |
|---|---|---|
| **PIRLS 2001** | β (SE) | β (SE) |
| Intercept | 527.0 (7.71)* | 454.2 (27.25)* |
| Early Literacy Activities | 52.7 (21.75)* | 27.3 (21.46) |
| Parents Like Reading | | 29.4 (24.49) |
| Duration of Preprimary Attendance | | 37.7 (14.30)* |
| Variance Explained ($R^2$) | 0.24 | 0.49 |
| **PIRLS 2011** | | |
| Intercept | 528.0 (7.13)* | 433.6 (35.96)* |
| Early Literacy Activities | 27.7 (24.19) | 18.9 (22.14) |
| Parents Like Reading | | 11.8 (26.46) |
| Duration of Preprimary Attendance | | 46.1 (16.90)* |
| Variance Explained ($R^2$) | 0.06 | 0.36 |

*P=0.05

Before adding covariates, the regression analysis for PIRLS 2011 Model 6 showed a nonsignificant regression coefficient of 27.7 (SE=24.19), representing the relationship between the country-mean scores on the early literacy activities scale and country-mean reading achievement. With the inclusion of the covariates average parents like reading score and average duration of preprimary attendance, the early literacy activities regression coefficient for PIRLS

2011 decreases to 18.9 (SE=22.14), which is also nonsignificant. In PIRLS 2011, parents like reading is associated with a nonsignificant coefficient of 11.8 (SE=26.46) and duration of preprimary attendance is associated with a large significant coefficient of 46.1 (SE=16.90). For PIRLS 2011, Model 6 explained 6% of the variance and Model 7 explained 36% of the variance. It is important to note that there is a large difference in the variance explained by Model 6 in the PIRLS 2001 model and the variance explained in the PIRLS 2011 model.

Given that there are only 21 observations per cycle, an outlying country can have much influence on the estimated regression coefficient. To examine the impact of influential points, the standardized DFBETA (SDFBETA) values from Model 7 were examined for the early literacy activities coefficient. SDFBETA examines the standardized difference between the coefficient estimated when the influential point is included in the analysis $\beta_1$ and the regression coefficient when the influential point is excluded $\beta_{1(-c)}$.

$$\text{SDFBETA} = \beta_1 - \beta_{1(-c)} \qquad \textbf{Equation 4.1}$$

For PIRLS 2001, the countries with the largest SDFBETA absolute values were Iran (SDFBETA=0.80) and Hong Kong (SDFBETA= -0.71). For PIRLS 2011, the countries with the largest SDFBETA absolute values were Hong Kong (SDFBETA= -1.33) and Russian Federation (SDFBETA=0.82). These two influential countries were removed from the analysis, and the analysis was re-conducted with the variables in Model 7.

Table 4.3 shows the results of the analysis after removing these outliers. As can be seen in the table, in PIRLS 2001 there was a slight change in the magnitude of the early literacy activities coefficient after removing the outlier countries, with the coefficient going from 27.3

(SE=21.46) with all of the countries in the analysis to 24.1 (SE=33.21) without outliers, suggesting that the early literacy activities coefficient estimate is relatively stable. However, the standard error increased substantially with the removal of these points, and the variance explained ($R^2$) decreased going from 49% to 21%. In addition to the inflation of the standard error and the large change in the variance explained, the coefficient associated with preprimary attendance decreased from 37.7 (14.30) with outliers to 21.2 (18.90).

**Table 4.3: Evaluating the Effect of Outliers on Model 7 Early Literacy Activities Coefficient Estimates**

| | Model 7 With Outliers | Model 7 Without Outliers |
|---|---|---|
| **PIRLS 2001** | β (SE) | β (SE) |
| Intercept | 454.2 (27.25)* | 484.2 (35.52) |
| Early Literacy Activities | 27.3 (21.46) | 24.1 (33.21) |
| Parents Like Reading | 29.4 (24.49) | 33.2 (25.15) |
| Duration of Preprimary Attendance | 37.7 (14.30)* | 21.2 (18.90) |
| Variance Explained ($R^2$) | 0.49 | 0.21 |
| **PIRLS 2011** | | |
| Intercept | 433.6 (35.96)* | 451.3 (33.97)* |
| Early Literacy Activities | 18.9 (22.14) | 29.1 (27.14) |
| Parents Like Reading | 11.8 (26.46) | 33.5 (25.58) |
| Duration of Preprimary Attendance | 46.1 (16.90)* | 35.7 (16.38)* |
| Variance Explained ($R^2$) | 0.36 | 0.45 |

*P=0.05

The outlier countries were more influential on the coefficient estimates associated with early literacy activities in PIRLS 2011. This regression coefficient increased from 18.9 (SE=21.46) when the outliers were included to 29.1 (SE=27.14) excluding the outliers. The variance explained increased from 36% in the model with all countries to 45% in the model excluding the outliers. It is noteworthy that there was a large increase in the parents like reading coefficient, increasing from 11.8 (SE=26.46) in the model that included all of the countries to 33.5 (SE=25.58) in the model without the outliers. Given the increase in the early literacy

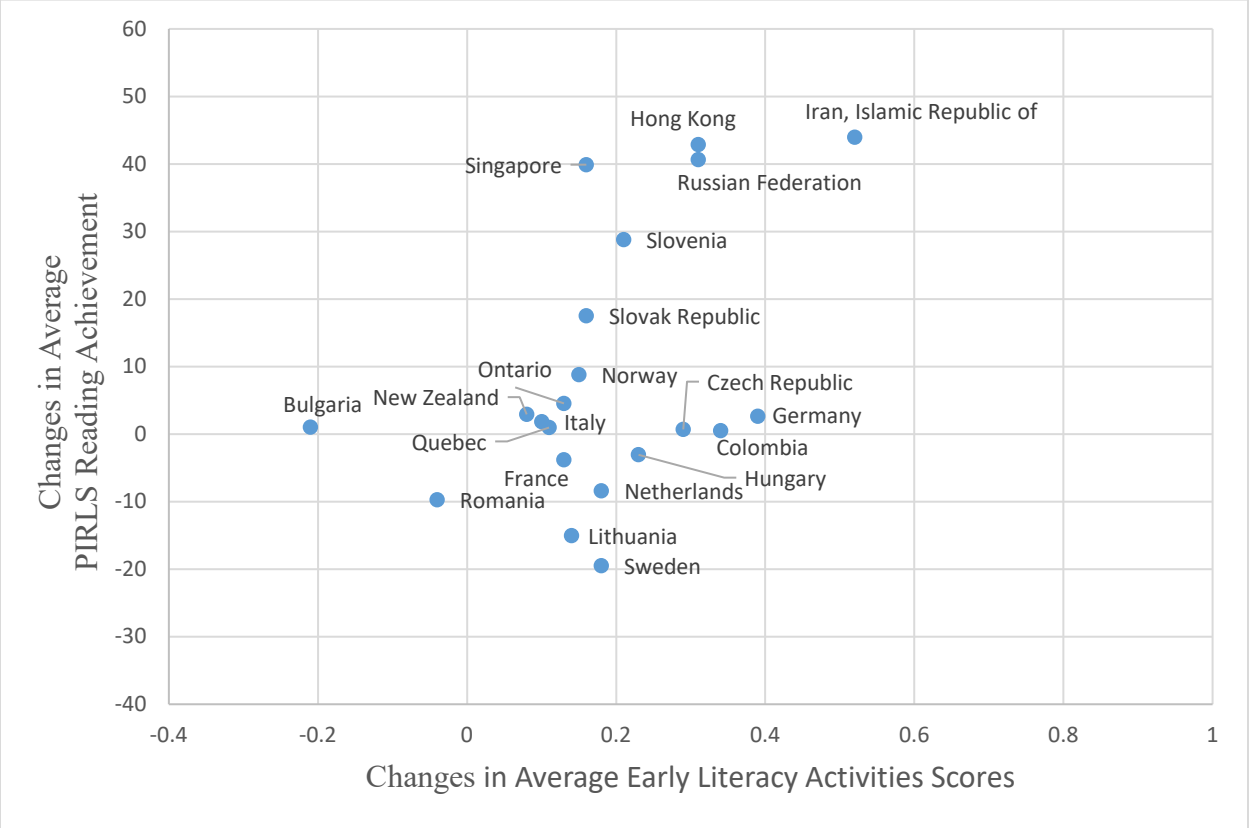
activities coefficient in PIRLS 2011 with the removal of the outliers, it was decided that this should be noted in interpreting the results.

Aggregation can also lead to heteroscedasticity (King, 1997)—differences in residual variance across the distribution of the explanatory variable(s). To evaluate the heteroscedasticity, the standardized residuals from Model 7 were plotted across the values of the average early literacy activities coefficient. The scatterplot is available in Appendix A, and it shows a random pattern suggesting that the assumption of homoscedasticity is fulfilled.

## 4.3 Phase 3: Country Difference-in-Differences Analysis

Country-level longitudinal analysis was then conducted using the first-difference approach. Prior to analysis, the country average difference scores for early literacy activities and PIRLS reading achievement were graphed. As can be seen in Figure 4.2, there seems to be substantial change over time in the average frequency parents engage their children in early literacy activities, with Iran increasing over a half of a standard deviation, and Colombia, Czech Republic, Germany, Hong Kong, and the Russian Federation increasing over a quarter standard deviation. The results provide evidence that there should be sufficient change in early literacy activities to be able to detect a relationship with reading achievement, if one exists.

**Figure 4.2: Scatterplot Displaying Country-Level Changes in Average Early Literacy Activities and Average Reading Achievement Between PIRLS 2001 and PIRLS 2011**




It is notable that a number of countries that made the largest increases in their mean early literacy activities scores such as Iran, Singapore, Hong Kong, and the Russian Federation also had large increases in their mean PIRLS reading achievement. Likewise, Romania declined both in mean early literacy activities scores and PIRLS mean reading achievement. Not all countries, however, fit this pattern.

Figure 4.3 presents the same scatterplot divided into quadrants based on whether the country-mean increased or decreased on the two variables. These quadrants provide a method of classifying countries based upon how they fit the hypothesis that increases in mean early literacy activities scores are associated with increases in mean reading achievement. From this

classification system it is noticeable that six countries do not fit the expected pattern—France, Hungary, Lithuania, the Netherlands, and Sweden increase in activities and decrease in achievement and Bulgaria decreases in activities and slightly increases in achievement.

**Figure 4.3: Scatterplot Showing Country-Level Changes in Average Early Literacy Activities and Average Reading Achievement Between PIRLS 2001 and PIRLS 2011 (Divided into Quadrants)**

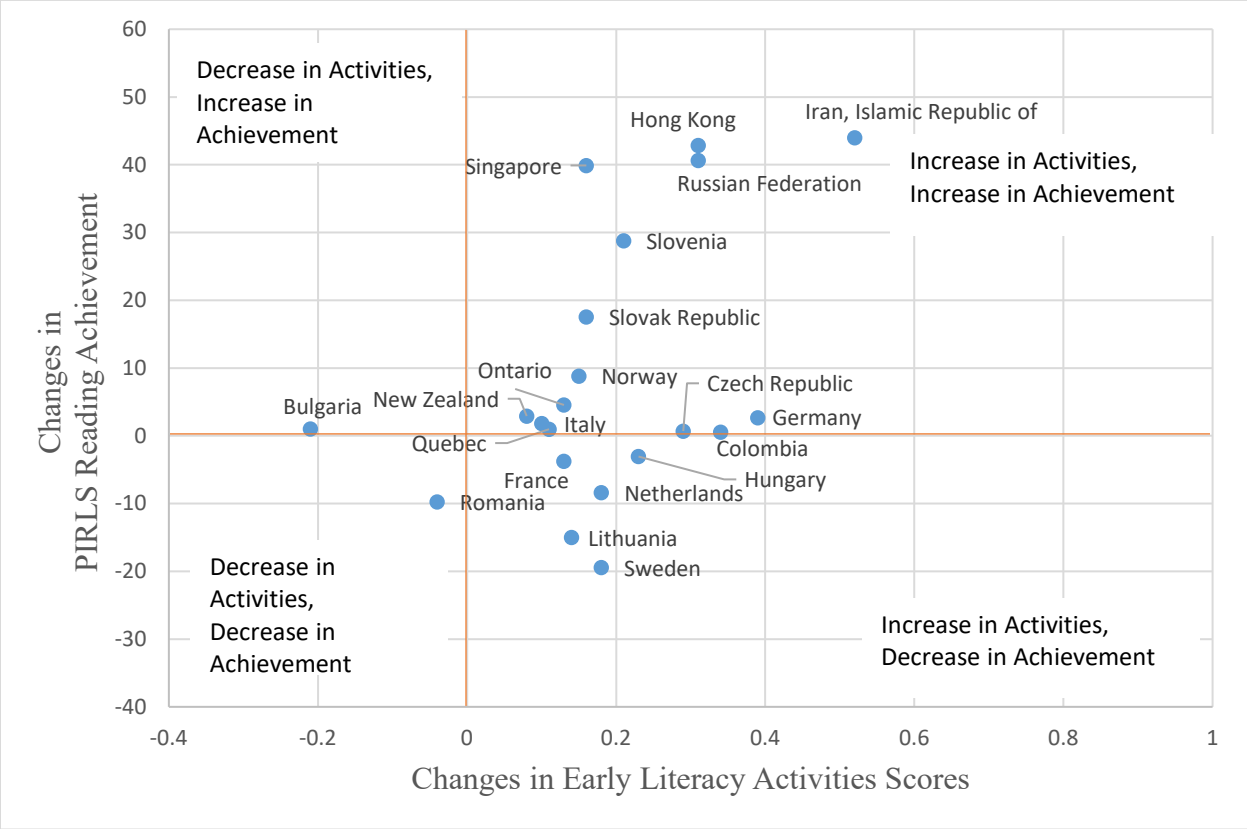



Table 4.4 shows the results of the country-level difference-in-differences analysis. Model 8 includes early literacy activities as the sole predictor of reading achievement. The results show a significant regression coefficient 83.6 (SE=24.01, $p<0.05$) implying that a one point increase across the two cycles in the early literacy activities country-level mean corresponds with about an 84-point increase in mean PIRLS scores. For Model 9, parents like reading and duration of preprimary attendance were added to the model. With the inclusion of these two time-varying

covariates, the results show little change to the early literacy activities coefficient, with a significant coefficient of 85.2 (SE=26.46, $p<0.05$) estimated. The parents like reading coefficient was also large at 19.1 (SE=33.89), but non-significant, and the duration of preprimary attendance coefficient was also non-significant at -6.7 (SE=17.79). Model 8 explained 40% of the variance and Model 9 explained 41% of the variance.

**Table 4.4: Parameter estimates for Phase 3 Country-Level Difference-in-Differences Analyses**

| | Model 8 | Model 9 |
|---|---|---|
| | β (SE) | β (SE) |
| Intercept | -6.7 (5.69) | -1.9 (9.83) |
| Early Literacy Activities | 83.6 (24.01)* | 85.2 (26.46)* |
| Parents Like Reading | | 19.1 (33.89) |
| Preprimary Attendance | | -6.7 (17.79) |
| Variance Explained ($R^2$) | 0.40 | 0.41 |

*P<0.05

Overall, the results of the Phase 3 analyses confirm a large positive relationship between increases in a country's mean early literacy activities score and increases in a country's mean PIRLS reading achievement score.

Another round of analysis was run to examine the influence of outlier countries. Similar to Phase 2, the standardized DFBETA values were used to identify influential countries in Model 9. Results showed that Germany (SDFBETA= -0.70) and Hong Kong (SDFBETA= 0.33) were the most influential countries in terms of changing the magnitude of the regression coefficient associated with early literacy activities.

Table 4.5 shows the results of the analysis after excluding these two countries. Comparing the results of Model 9 with outliers to the results without outliers, it can be seen that the coefficient representing the relationship between average early literacy activities and average PIRLS reading achievement increased to 94.1 (SE=27.99) after omitting these two countries. The

variance explained increased to 0.49, meaning the model with these countries omitted explains 49% of the variance in PIRLS reading achievement. It was concluded that the influence of these countries was minimal given the magnitude of the coefficient estimates. The model was also reviewed for heteroscedasticity (scatterplot available in Appendix A) and the variance appeared to be random across the early literacy activities distribution.

**Table 4.5: Evaluating the Effect of Outliers on Model 9 Early Literacy Activities Coefficient Estimates**

| | Model 9 With Outliers | Model 9 Without Outliers |
|---|---|---|
| | β (SE) | β (SE) |
| Intercept | -1.9 (9.83) | -2.0 (9.74) |
| Early Literacy Activities | 85.2 (26.46)* | 94.1 (27.99) |
| Parents Like Reading | 19.1 (33.89) | 21.9 (34.27) |
| Preprimary Attendance | -6.7 (17.79) | -8.5 (18.82) |
| Variance Explained ($R^2$) | 0.41 | 0.49 |

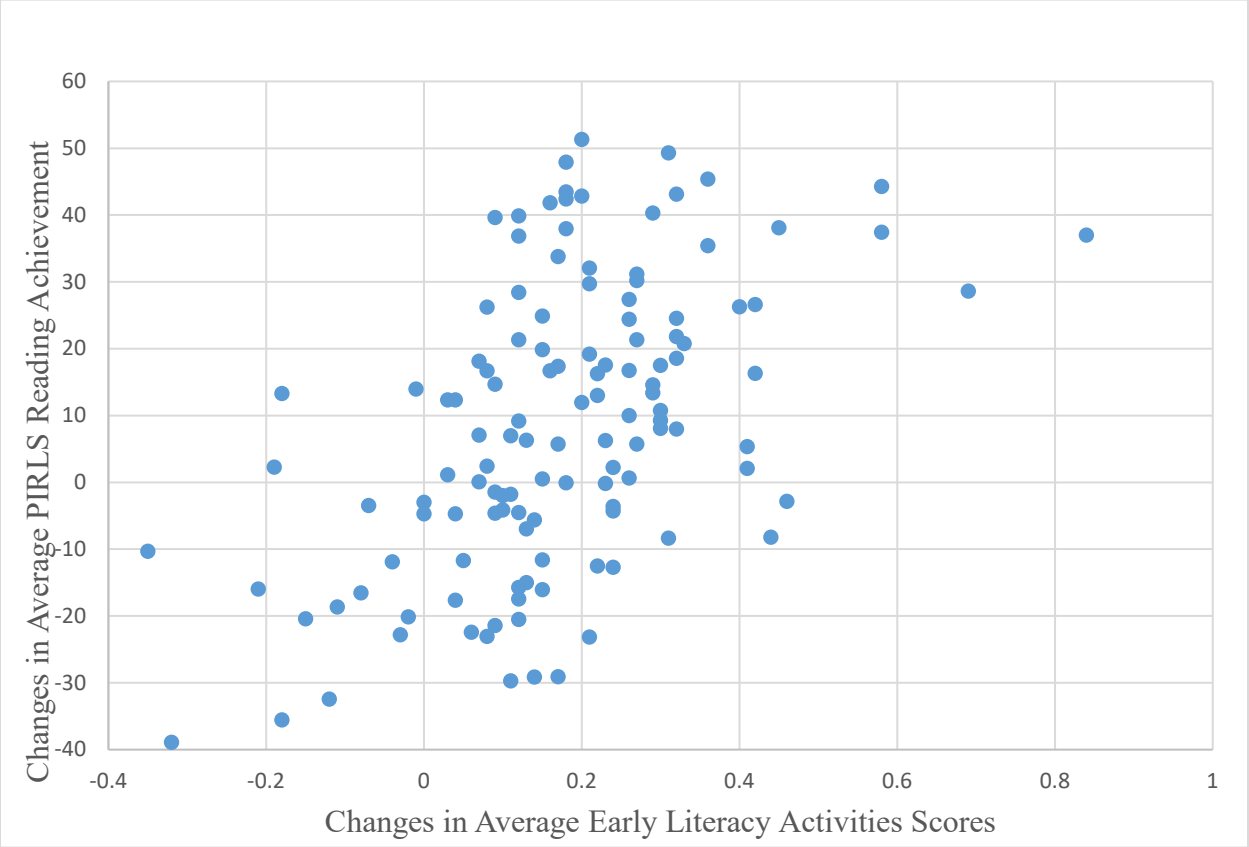
*P<0.05

## 4.4 Phase 4: Subpopulation Difference-in-Differences Analysis

Following the country-level analysis, weighted student data were aggregated to subpopulation-level. In total 126 subpopulations were created across the 21 countries—six subpopulations per country.

Figure 4.4 shows the relationship between changes in mean early literacy activities scores and changes in mean PIRLS reading achievement scores among subpopulations. As can be seen in the scatterplot, there is a positive association between early literacy activities and PIRLS reading achievement, with substantial variability present in the data.

**Figure 4.4: Subpopulation Relationship Between Changes in Average Early Literacy Activities Scores and Changes in Average Reading Achievement**




Subpopulation difference-in-differences analysis was then conducted employing the first-difference approach. The results are shown in Table 4.6. Model 10 shows a significant coefficient of 66.3 (SE=12.58) when the early literacy activities variable alone is entered into the model. The significant coefficient implies that a one point increase on the early literacy activities scale for a subpopulation mean corresponds with an increase of 66 points in average PIRLS reading achievement. Comparing the results in Model 10 with subpopulation data and Model 8 with country data, the coefficient associated with mean early literacy activities changes by nearly 20 points by measuring it at a different level of aggregation, going from 84 points when measured though country difference-in-differences to 66 points when measured through subpopulation difference-in-differences. The variance explained also decreases from 40% in Model 8 to 27% in Model 10. Because the true individual-level relationship between early

literacy activities and PIRLS reading achievement is unknown, it is unclear which model comes closer to the relationship between the variables at individual level.

**Table 4.6: Parameter estimates for the Phase 4 Subpopulation Approach**

| | Model 10 | Model 11 |
|---|---|---|
| | β (SE) | β (SE) |
| Intercept | -4.1 (3.86) | -2.1 (6.41) |
| Early Literacy Activities | 66.3 (12.58)* | 64.8 (14.38)* |
| Parents Like Reading | | 8.8 (17.24) |
| Length of Preprimary Attendance | | -1.0 (17.35) |
| Variance Explained ($R^2$) | 0.27 | 0.28 |

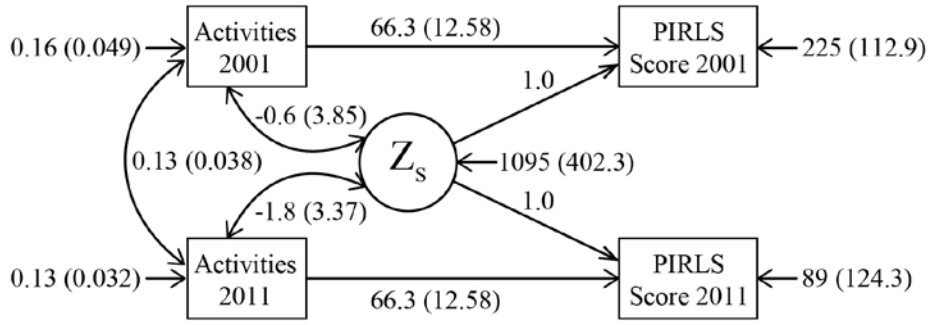
*P<0.05

Model 11 shows only a small shift in the mean early literacy activities coefficient (64.8 SE=14.38) following the inclusion of the time-varying covariates. Again, the average early literacy activities coefficient was nearly 20 points smaller than the country-level model, with the coefficient at country-level in Model 9 being 85.2 (SE=26.46). The coefficient describing the relationship between average parents like reading and average PIRLS reading achievement was nonsignificant at 8.8 (17.24) and the relationship between average duration of preprimary attendance and average PIRLS reading achievement was nonsignificant at -1.0 (SE=17.35).

The outliers were reviewed to examine whether there were influential points in the data. The standardized DFBETA values were all quite small with the largest having value of 0.03 and the smallest having a value of -0.03. Given the small size of these values, it was decided that the early literacy activities regression coefficient was not overly affected by outliers. The data were also examined for heteroscedasticity by plotting the standardized residuals by the changes in early literacy activities scores, as shown in Appendix A, and the results show a random pattern for the residuals across the early literacy activities distribution.

To provide a baseline for estimating fit statistics for Phase 7, the Model 10 fixed-effects model was also estimated through the structural equation modeling approach in Mplus with early literacy activities as the sole predictor. The results are shown in the path model in Figure 4.5. As can be seen in the figure, the unstandardized path coefficient estimating the relationship between changes in average early literacy activities and changes in average reading achievement is 66.3 (SE=12.58)—identical to the regression coefficient estimated through the first-difference approach. Overall, the fixed-effects approach to structural equation modeling shows excellent fit with a $\chi^2$ model fit of 0.486 (df=1, p>0.05), an RMSEA of 0.00, and a CFI of 1.00. The fixed-effects results show a nonsignificant covariance of 0.6 (SE=3.85) between $Z_s$ and the 2001 early literacy activities variable and a nonsignificant covariance of -1.8 (SE=3.37) between $Z_s$ and 2011 early literacy activities.

**Figure 4.5: Path Model with the Subpopulation Fixed-Effects Model Results (Model 10)**



## 4.5 Phase 5: Analysis of Within-Country Relationships

Phase 5 examines within-country relationships using the subpopulation data. Because within-country variability is lost through aggregation to country-level, this section evaluates to what extent this within-country variability is relevant for understanding the relationship between early literacy activities and PIRLS reading achievement. In the subpopulation approach, the within-country variation refers to the differences between the six subpopulations within each country.

**Estimating the Intraclass Correlation Coefficient.** Because subpopulations are nested within countries, a key statistic is the intraclass correlation coefficient, which describes the percentage of a variable's variance at different levels of aggregation. Following from this, the intraclass correlation coefficient was estimated across countries for both changes in average PIRLS reading achievement and changes in average early literacy activities. The results for changes in average PIRLS reading achievement showed an intraclass correlation coefficient of .75, meaning that 75% of the variance in changes in average PIRLS reading achievement is at country level and the other 25% of the variance is at subpopulation level. The intraclass correlation coefficient was also calculated for changes in early literacy activities and the results showed that 57% of the variance in mean changes in early literacy activities was at country-level and 43% of the variance in mean changes in early literacy activities was at subpopulation level.

**Dispersion Analysis.** Given that there is substantial variance at subpopulation-level in changes in mean early literacy activities and changes in mean reading achievement, the next stage of the Phase 5 analysis graphically examined the relationships between subpopulations within countries. The subpopulation scatterplot from Figure 4.4 was modified to color-code the subpopulations by the countries to which they pertain. Figure 4.6 shows the modified scatterplot. The scatterplot provides a general picture of the dispersion in changes in average reading

achievement and changes in average early literacy activities that is present in each country. However, given that 21 unique colors were used in the plot, some of the colors are quite similar and it can be difficult to decipher the identity of the countries. For this reason, additional scatterplots were generated, with one focusing on the five countries with the largest standard deviation in changes in early literacy activities means between the country's subpopulations and another focusing on the five countries with the smallest standard deviation between the changes in subpopulation means.

Figure 4.7 highlights the five countries with the largest dispersion in changes in average early literacy activities scores—Iran, Romania, Russian Federation, New Zealand, and Slovak Republic. As can be seen in the figure, the Russian Federation had a wide spread on average early literacy activities but little difference in PIRLS average reading achievement across the groups. For Iran, New Zealand, Romania, and Slovenia, there appears to be clear pattern where the changes in the mean PIRLS reading achievement for a subpopulation increases with the changes in the mean early literacy activities scores.

Figure 4.8 shows the five countries with the smallest standard deviation between the subpopulations—Czech Republic, Germany, Italy, Ontario, and Quebec. Although the Czech Republic, Germany, and Quebec had only small differences in changes in average early literacy activities between subpopulations, they had a wider spread between the subpopulations on changes in average reading achievement. Italy had little variation between the subpopulations on either the early literacy activities scores or the subpopulation scores.

**Figure 4.6: Subpopulation Relationship Between Changes in Average Early Literacy Activities Scores and Changes in Average Reading Achievement (Color Coded)**

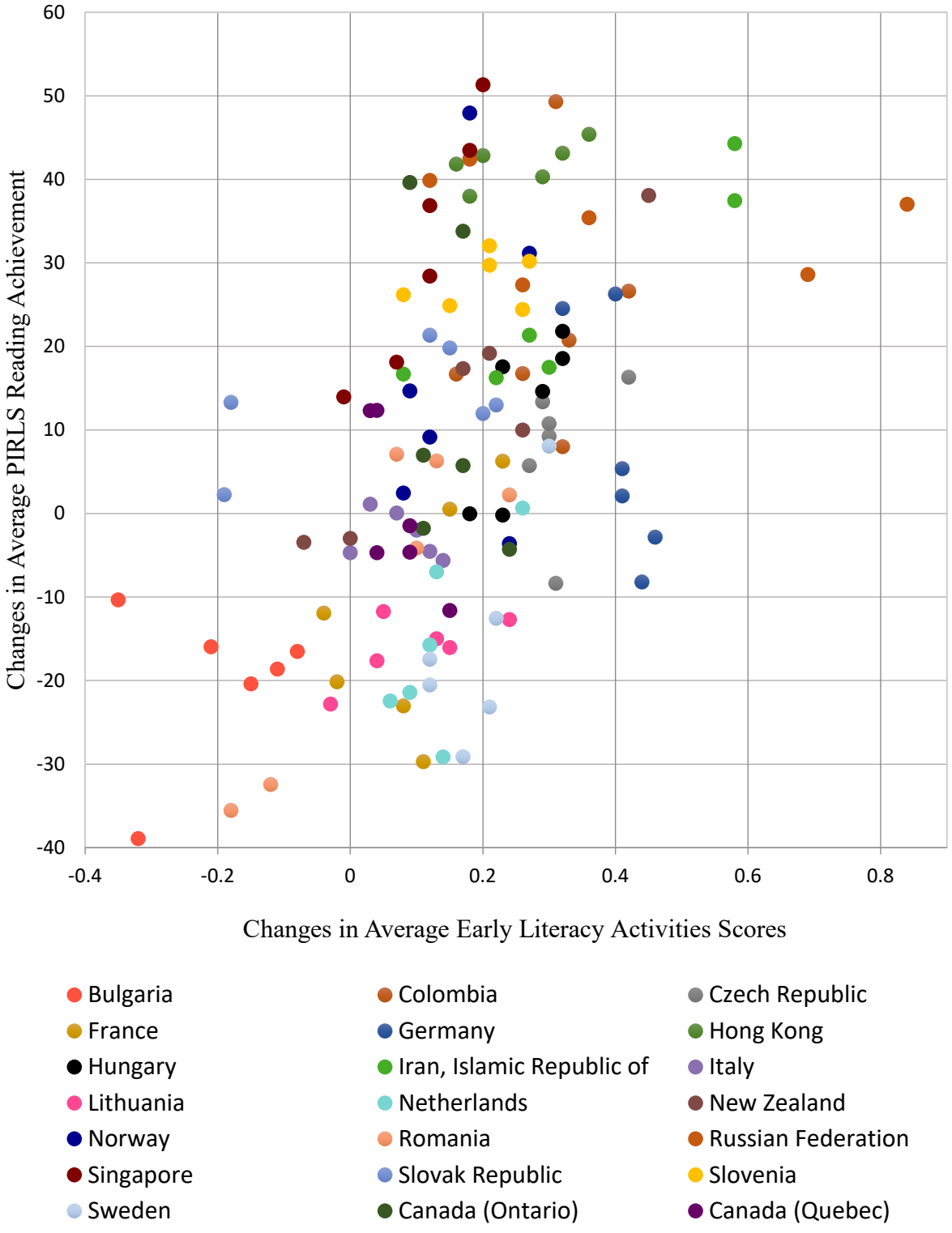


**Figure 4.7: Highlighting the Five Countries with the Largest Standard Deviation Between the Subpopulations in Changes in Average Early Literacy Activities Scores**

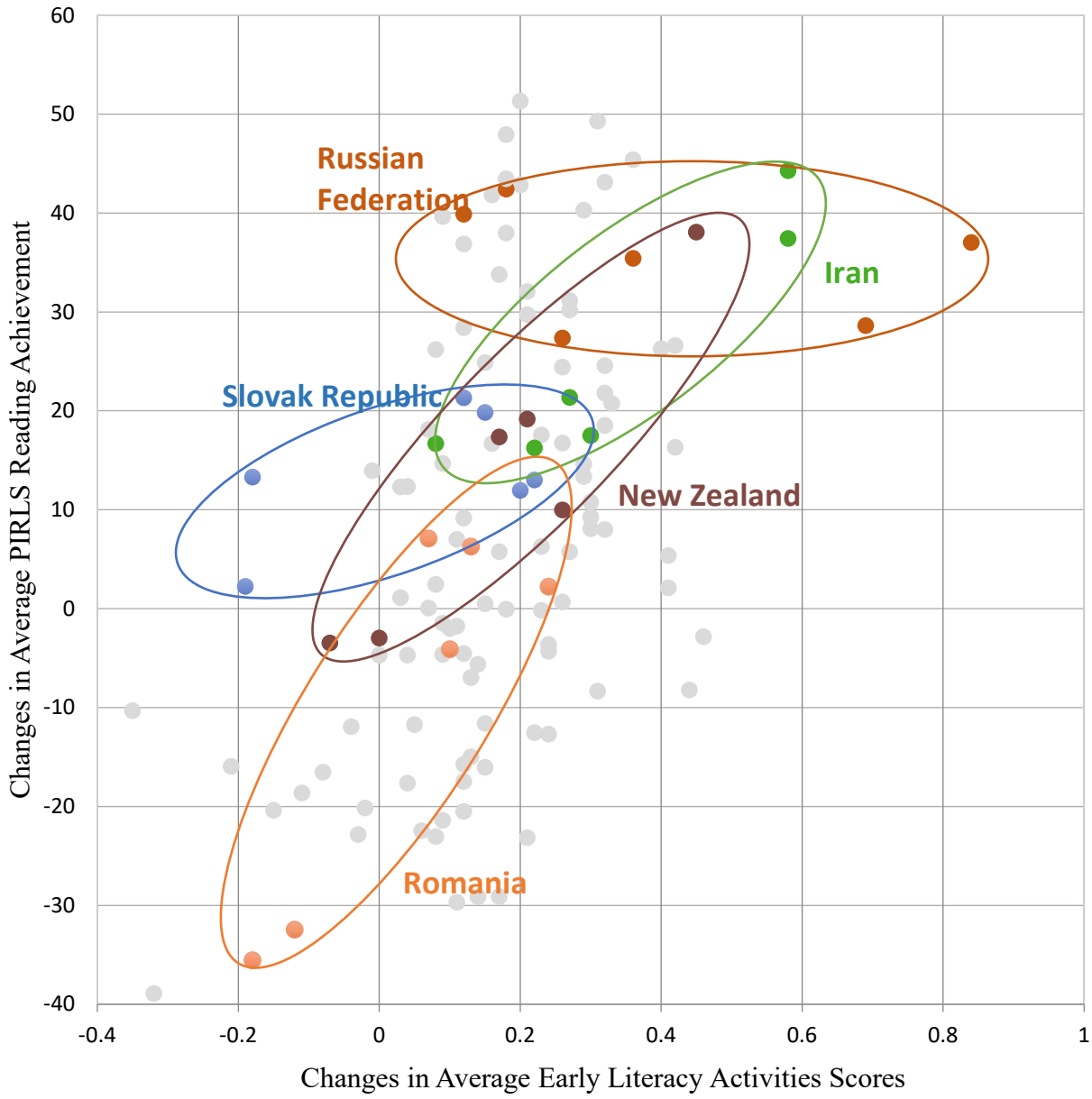
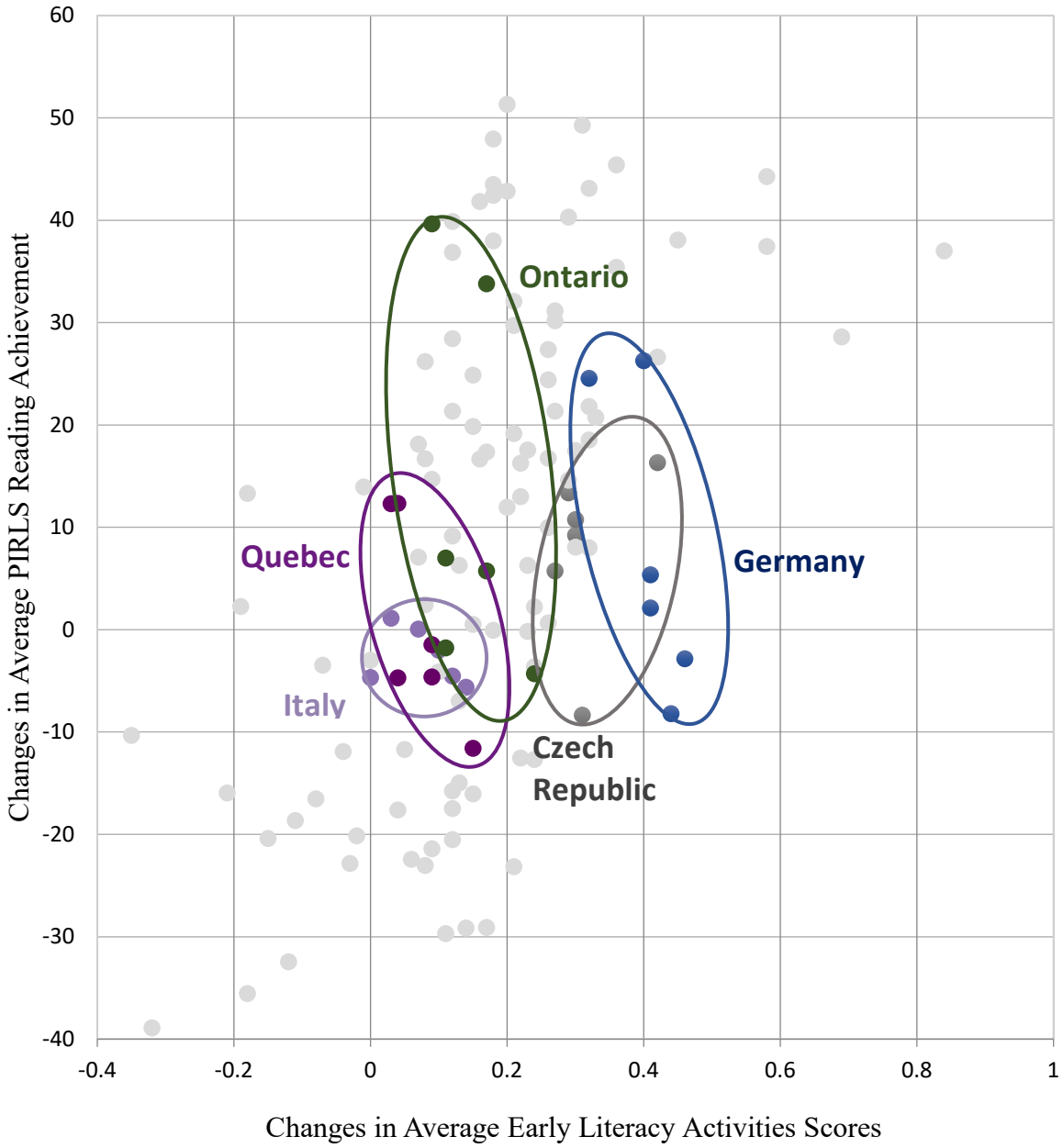


**Figure 4.8: Highlighting the Five Countries with the Smallest Standard Deviation Between the Subpopulations in Changes in Average Early Literacy Activities Scores**



**Within-country relationship between changes in early literacy activities and changes in PIRLS reading achievement.** Additional exploratory analysis was conducted by plotting a regression line for each country estimating the relationship between changes in average early literacy activities and changes in average PIRLS reading achievement. The regression coefficient was estimated using PIRLS Senate weights, and because of this the regression line is pulled toward the subpopulations representing more students. Also, because there are only six subpopulations per country, it should be noted that the coefficients pertaining to each country are measured with a large amount of error. Keeping this in mind, it is possible to gain insights from examining patterns across the countries.

Figure 4.9 shows the 9 countries that had the sharpest positive relationship (regression coefficients above 0.3) between changes in early literacy activities means and changes in PIRLS reading achievement means—Singapore, Romania, Hungary, Iran, New Zealand, Colombia, France, Netherlands, and Sweden. It is notable that Iran, New Zealand, and Romania are three of the five countries with the largest standard deviations included in Figure 4.7.

Figure 4.10 shows the 6 countries that have a slight positive relationship (regression coefficients between 0 and 0.3) between PIRLS reading achievement and early literacy activities scores—Hong Kong, Slovenia, Slovak Republic, Czech Republic, Bulgaria, and Lithuania. Slovak Republic was one of the five countries with the largest standard deviation between the subpopulations in Figure 4.7 and Czech Republic was one of the five countries with the smallest differences between the subpopulations in Figure 4.8.

Figure 4.11 provides the scatterplots for the six countries with negative coefficients—Russian Federation, Germany, Ontario, Norway, Italy, and Quebec. Four of these countries—

Germany, Italy, Ontario, and Quebec, were also countries with the least dispersion among the subpopulations, and the Russian Federation is the country with the largest dispersion among the subpopulations.

**Figure 4.9: Relationship Between Changes in Average PIRLS Reading Achievement and Changes in Average Early Literacy Activities Scores for Countries with Coefficients over 0.3**

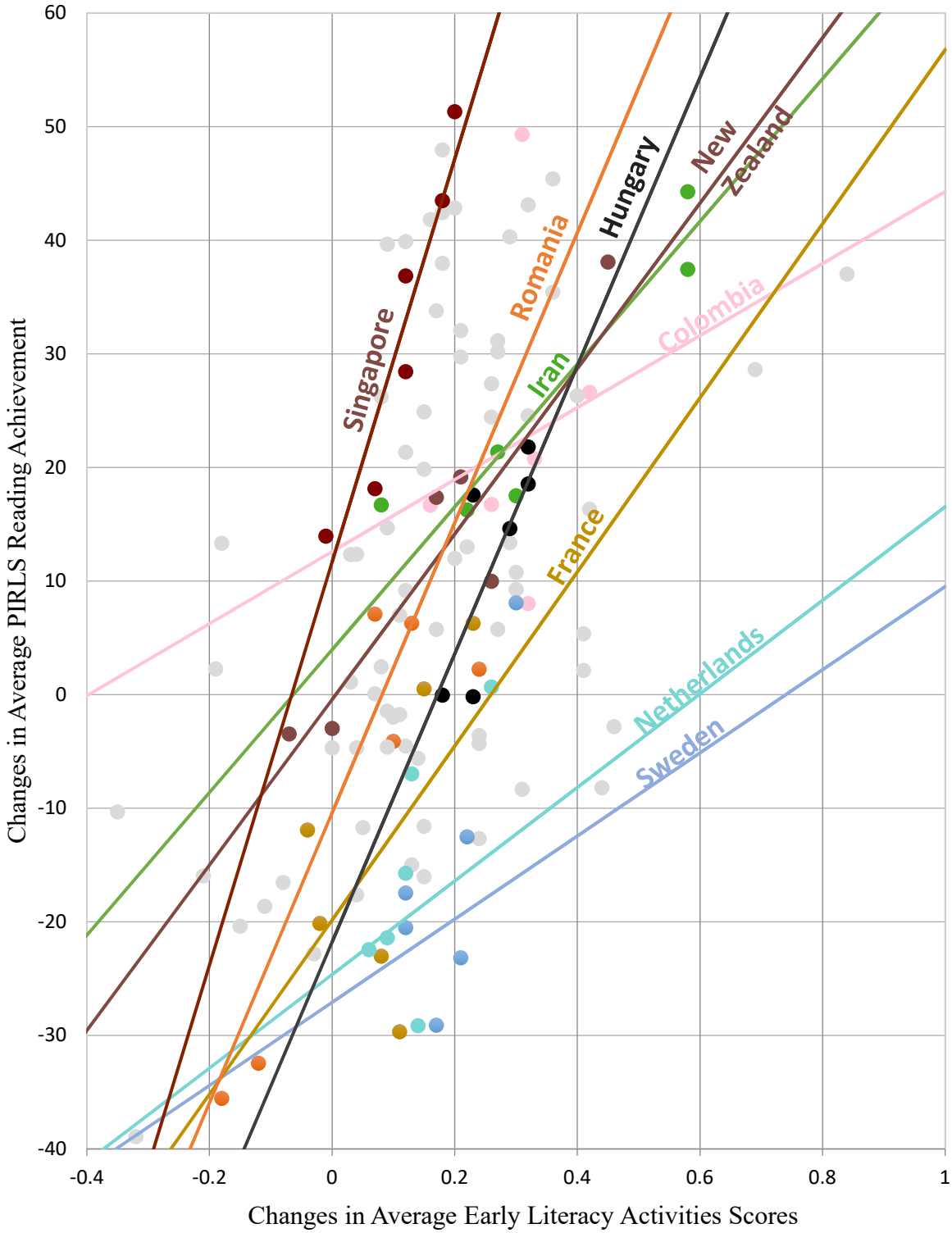


**Figure 4.10: Relationship Between Changes in Average PIRLS Reading Achievement and Changes in Average Early Literacy Activities Scores for Countries with Coefficients Between 0 and 0.3**

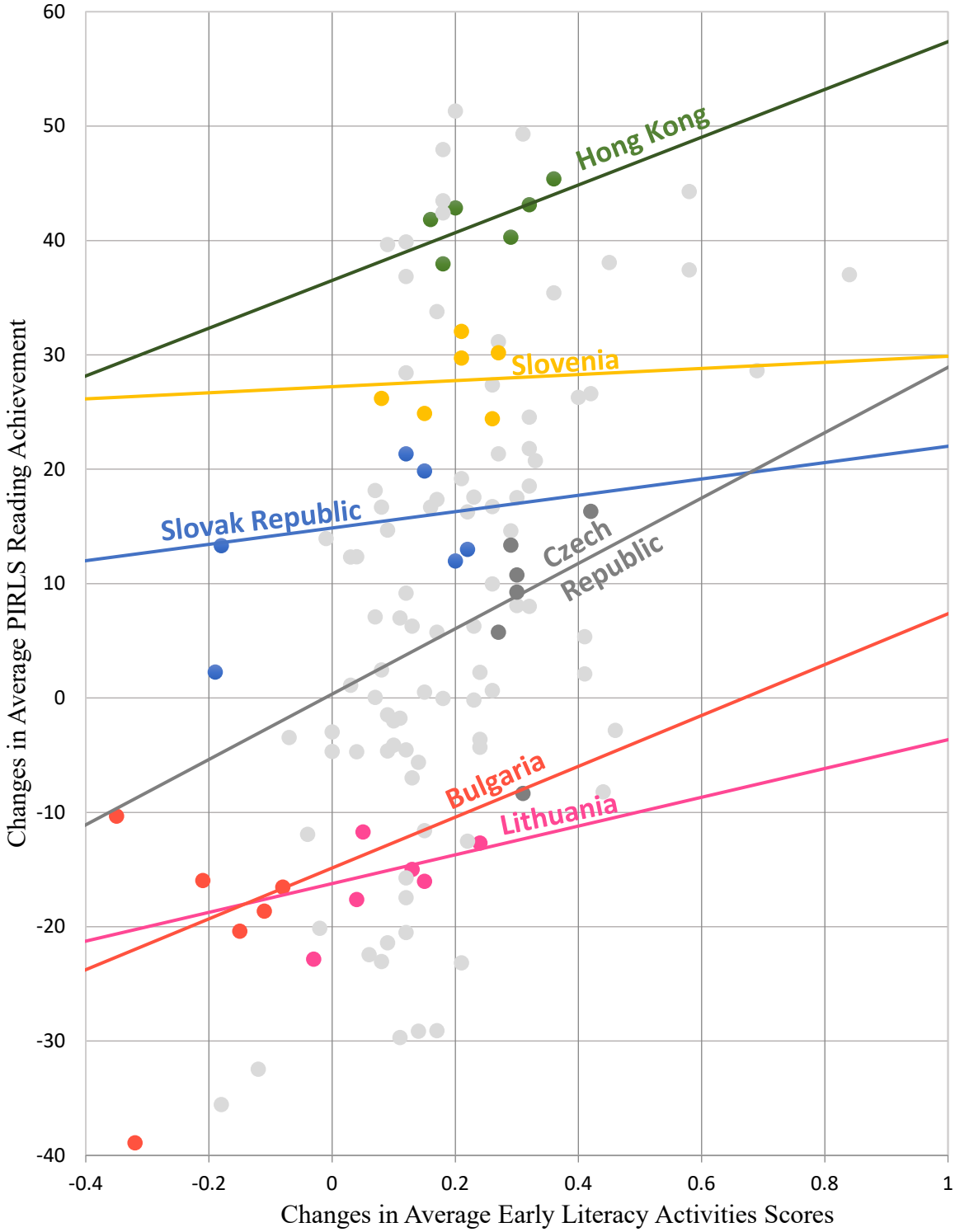
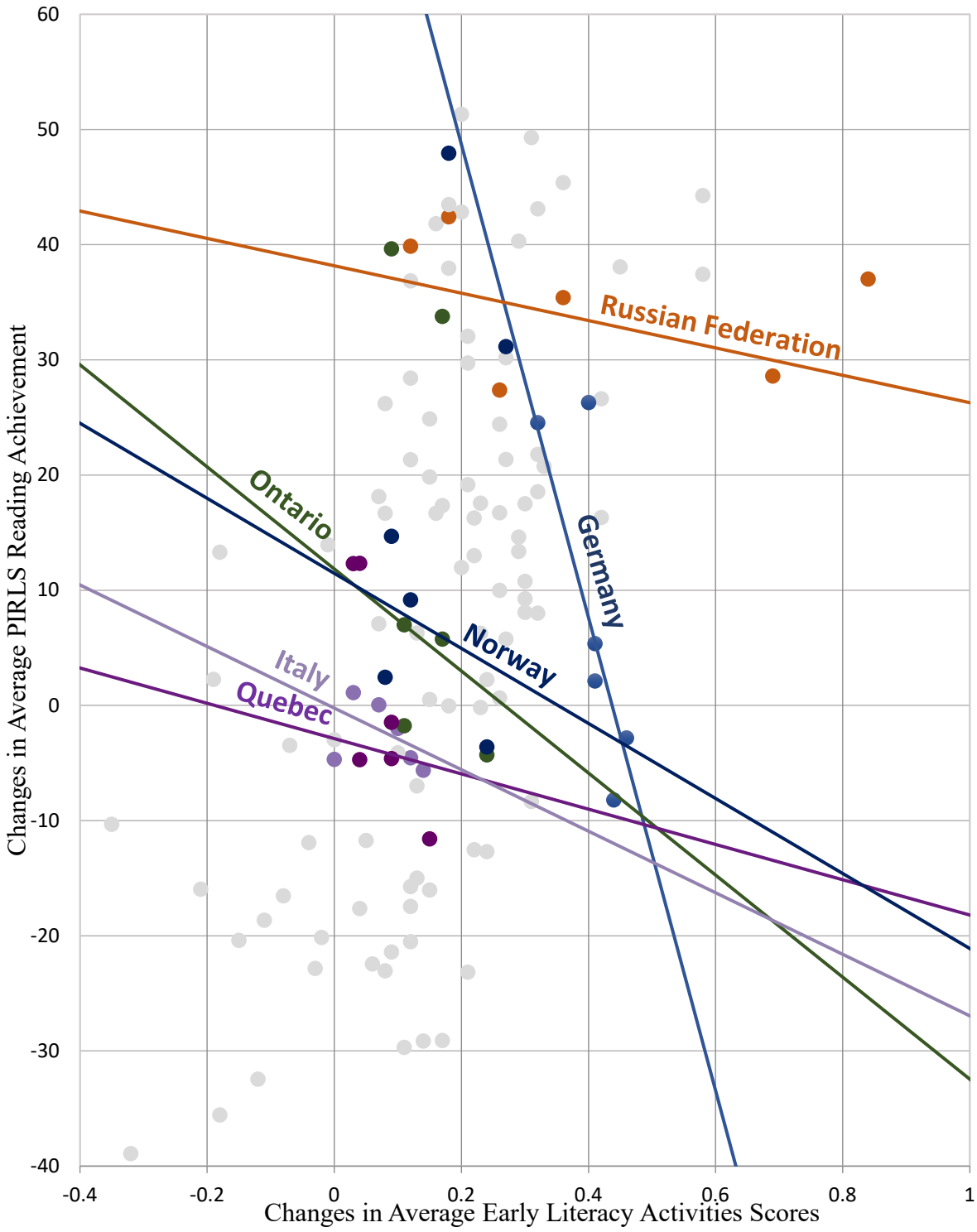


**Figure 4.11: Relationship Between Changes in Average PIRLS Reading Achievement and Changes in Average Early Literacy Activities Scores for Countries with Negative Coefficients**

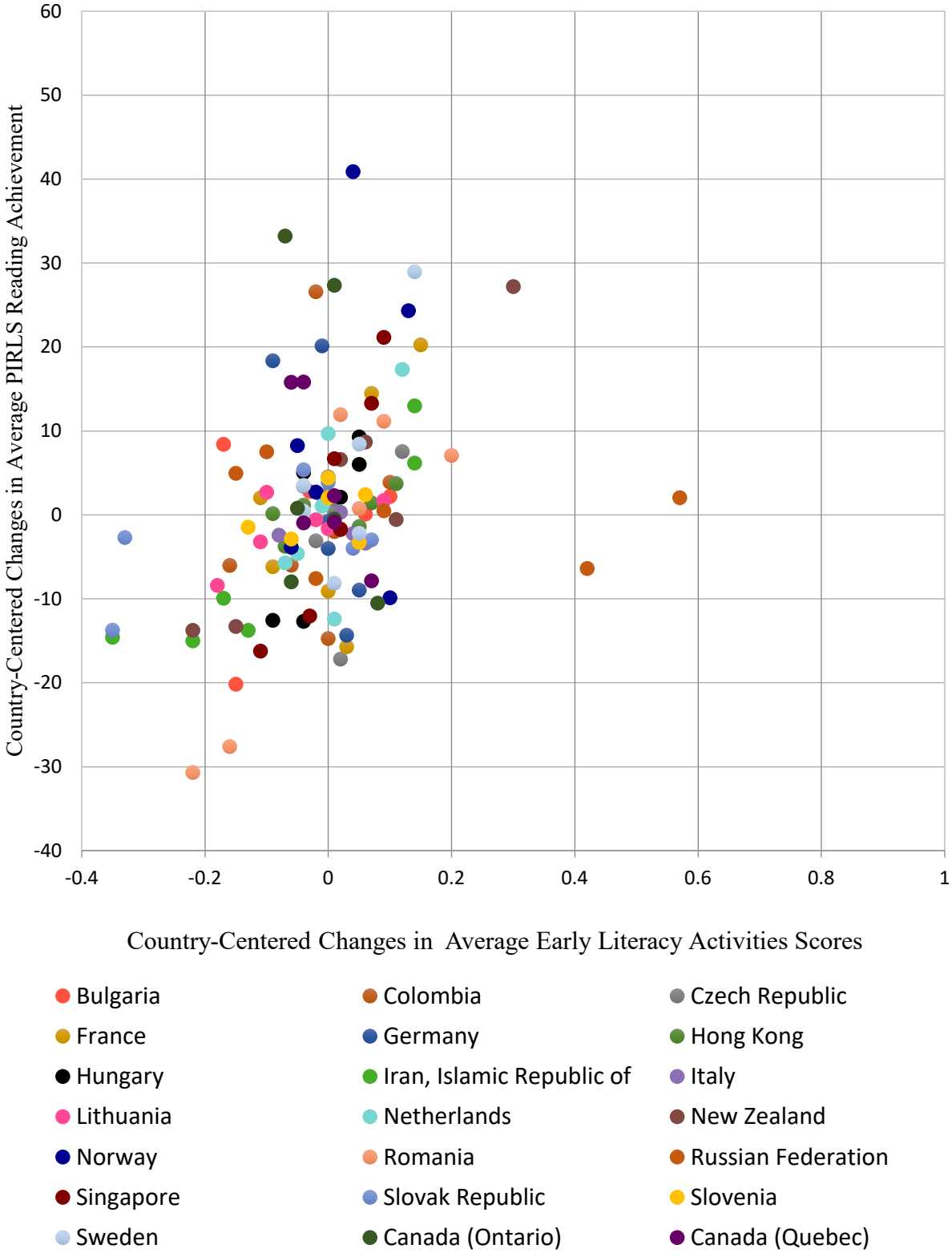


**Cross-country analysis of subpopulation-level variance.** In order to examine the subpopulation-level relationships between early literacy activities and reading achievement, the variance in early literacy activities was decomposed by centering the average difference score for early literacy activities for each subpopulation on the average early literacy activities difference score for the country as a whole. Likewise, the average difference score for PIRLS reading achievement for each subpopulation was centered on the average reading achievement difference score for the country. As such, Phase 5 analyzes solely the subpopulation-level variance.

Following the centering, a scatterplot was created to illustrate the relationship between country-centered changes in early literacy activities and country-centered changes in PIRLS reading achievement. Figure 4.12 displays the scatterplot. Overall, there appears to be a positive relationship among subpopulations, with the subpopulations with greater increases in early literacy activities relative to the country-mean score having more increases in average reading achievement relative to the country-mean score. There, however, appears to be a lot of variability in this relationship.

Analysis was then conducted on these country-centered difference scores, regressing the reading achievement scores on the early literacy activities scores. The results showed a significant regression coefficient of 39.0 (SE=17.90), suggesting that an increase of one point in the subpopulation score on the early literacy activities scale relative to the country mean was associated with a 39 point increase in PIRLS reading achievement. The model explained 16% of the variance in PIRLS reading achievement. The results confirm that a relationship between changes in early literacy activities and PIRLS reading achievement can be found when analyzing solely subpopulation-level variance.

**Figure 4.12: Subpopulation-Level Relationship Between Country-Centered Changes in Average PIRLS Reading Achievement and Country-Centered Changes in Average Early Literacy Activities Scores**



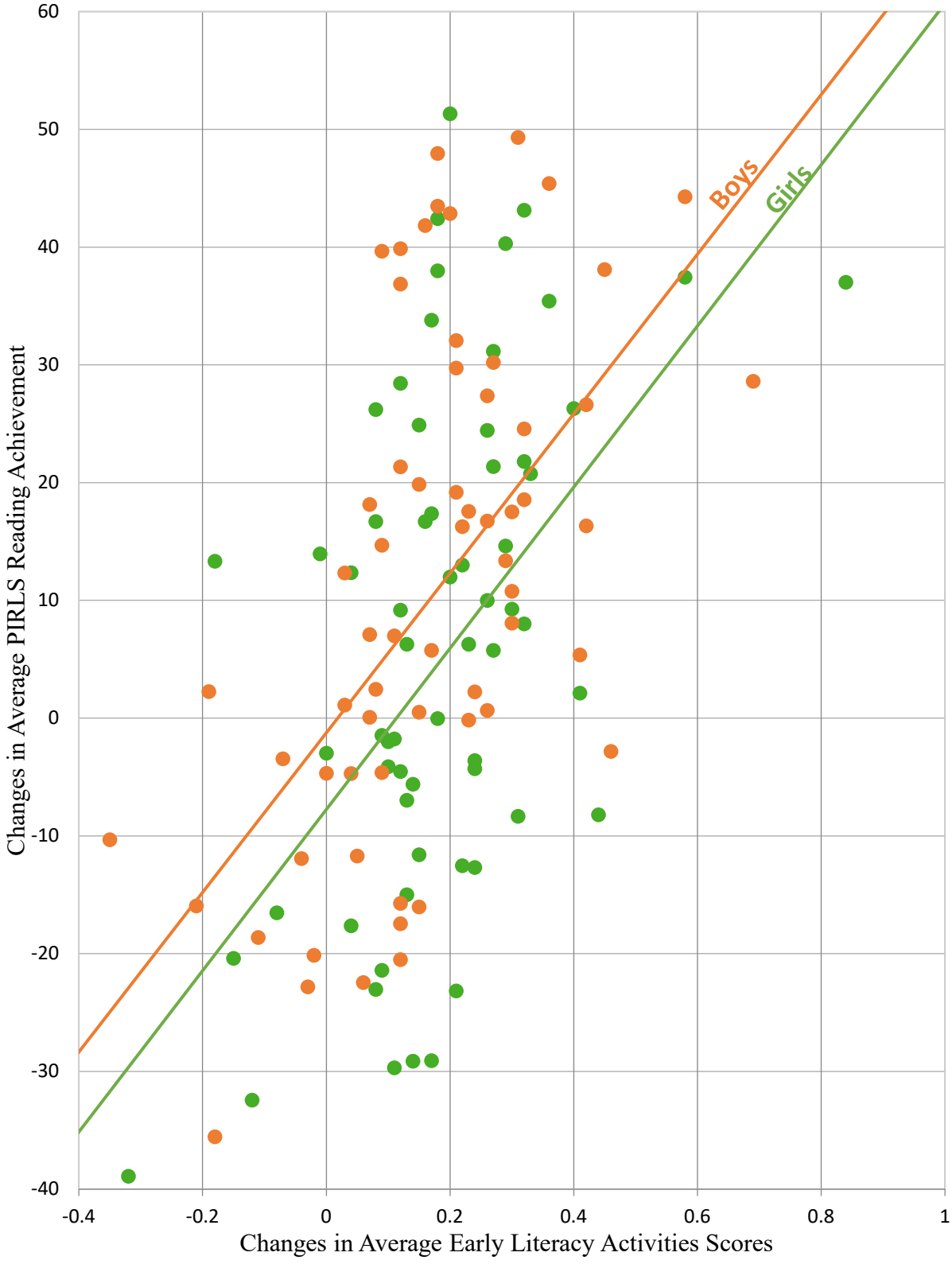
No overly influential points were identified in the outlier analysis, as all SDFBETA values were between -0.15 and 0.15. The data were also reviewed for heteroscedasticity and the data appear to have a random pattern of residual variance across the distribution of early literacy activities (the scatterplot is available in Appendix A).

## 4.6 Phase 6: Comparisons of Fixed-Effect Coefficient Estimates across Groups

Building on the subpopulation model from Phase 4, Phase 6 examines whether the relationship between changes in average early literacy activities scores and changes in average PIRLS reading achievement scores varies between the subpopulations composed of boys and the subpopulations composed of girls, and whether the relationship varies among the highest parental education groups.

Initial exploratory analysis was conducted by examining a scatterplot displaying the relationship between changes in average early literacy activities scores and changes in average PIRLS reading achievement scores by the gender of each subpopulation. The scatterplot can be seen in Figure 4.13. The orange line represents the relationship between changes in average early literacy activities scores and PIRLS reading achievement for subpopulations with boys and the green line represents the relationship for subpopulations with girls. As can be seen in the plot, the two lines are approximately parallel suggesting that the male and female subpopulations have a similar relationship.

**Figure 4.13: Relationship Between Changes in Average Early Literacy Activities and Changes in Average Reading Achievement by Gender of Subpopulation**



It is important to keep in mind that the relationship represented by the intercept of the lines signifies changes in average reading achievement between PIRLS 2001 and PIRLS 2011. Therefore, the line pertaining to the boys' subpopulations is slightly above the line corresponding with the girls' subpopulations because the average student achievement of boys' subpopulations increased relative to that of girls between PIRLS 2001 and PIRLS 2011.

Following this exploratory analysis, a structural equation model was employed to conduct fixed-effects analysis using the multiple group approach. The results for the null model, where the relationship between changes in the mean early literacy activities score and changes in the mean reading achievement was fixed across boys and girls, had an unstandardized coefficient of 66.3 (12.56) and excellent fit with a $\chi^2$ model fit of 3.880 (df=4, p>0.05), an RMSEA of 0.00, and a CFI of 1.00. The alternative model, where the relationship between average early literacy activities scores and average reading achievement scores was allowed to vary across boys and girls, showed an unstandardized coefficient of 68.4 (SE=14.17) for girls and an unstandardized coefficient estimate of 67.7 (SE=12.92) for boys. The alternative model also showed strong fit with a $\chi^2$ model fit of 0.496 (df=2, p>0.05), an RMSEA of 0.00, and a CFI of 1.00. A $\chi^2$ difference test was conducted to examine whether the unconstrained model, with unique coefficients estimated for boys and girls, had significantly better fit than the constrained model. The results showed a nonsignificant $\chi^2$ difference of 3.384 (df=2, p>0.05). The Wald Test of parameter constraints also showed nonsignificant results ($\chi^2$=0.005, df=1, p>0.05), and the RMSEA and CFI fit statistics were identical across the two models. Because the model tests and fit statistics showed the two models to be similar, and the boys and girls coefficient estimates were almost equal in the alternative model, it was concluded that the relationship between

changes in mean early literacy activities and changes in mean reading achievement was similar across the male and female subpopulations (the path models for the analysis are available in Appendix A).

Figure 4.14 shows a scatterplot examining group differences in changes in average PIRLS reading achievement and changes in average early literacy activities between parental education groups. As can be seen in the plot, the subgroup with no high school education has the steepest slope and the college education group has the least steep slope. The crossing of the regression lines hints to a moderation effect.

To formally evaluate this relationship, a structural equation model was employed using the Mplus multiple group function. The null model, where the relationship between early literacy activities and reading achievement was constrained to be equal across parental education groups, had an unstandardized coefficient of 68.2 (10.53) and good fit with a $\chi^2$ model fit of 4.468 (df=7, p>0.05), an RMSEA of 0.00, and a CFI of 1. For the alternative model, the relationship between early literacy activities and reading achievement was allowed to vary across the parental education groups and showed an unstandardized coefficient of 74.8 (SE=12.76) for those whose highest parental education was less than a high school degree, a coefficient estimate of 62.4 (SE=20.64) for those whose parents highest education level was completing high school, and a coefficient of 34.8 (SE=29.84) for those whose highest parental education level was a college degree. The alternative model showed excellent fit with a $\chi^2$ model fit of 2.256 (df=3, p>0.05), an RMSEA of 0.00, and a CFI of 1.00. A $\chi^2$ difference test was conducted to examine whether the unconstrained model, with unique coefficients estimated for each of the highest parental education groups, had significantly better fit than the constrained model. The results showed a nonsignificant $\chi^2$ difference of 2.212 (df=4, p>0.05), suggesting that the two models have similar

fit, and the RMSEA and CFI both show excellent fit in the null and alternative models. The Wald Test, which compared the slopes across the three groups, was nonsignificant ($\chi^2$=3.369, df=3, p<0.05). Although the differences in coefficient estimates across the three groups seems sizable, the fit statistics and Wald Test suggest the coefficients to be similar (The path models for the analysis are available in Appendix A).

Given these perplexing findings, further analysis was conducted to examine why the standard error was so large for the college parental education group—as the sample sizes are similar across the three education groups. In addition to being a function of sample size, the standard error of the regression weight in simple regression is a function of residual variance and the sum of the squared deviations of the explanatory variable. Figure 4.15 shows the scatterplot for the no high school group and Figure 4.16 shows the scatterplot for the college group. Comparing the figures, it can be seen that there is a visually apparent association between early literacy activities and PIRLS reading achievement for the no high school group and the association is not apparent for the college group. The residual variance also appears to be larger for the college group compared with the no high school group. Finally, there is less dispersion in the early literacy activities explanatory variable in the college group (all values between -.2 and .5) than the no high school group (values spread between -0.4 and 1). As such, it can be concluded that the high standard error for the college group can be attributed to a high residual variance and a relatively low sum of the squared deviations for the early literacy activites explanatory variable.

**Figure 4.14: Relationship Between Changes in Average Early Literacy Activities and Changes in Average Reading Achievement for Highest Parental Education Groups**

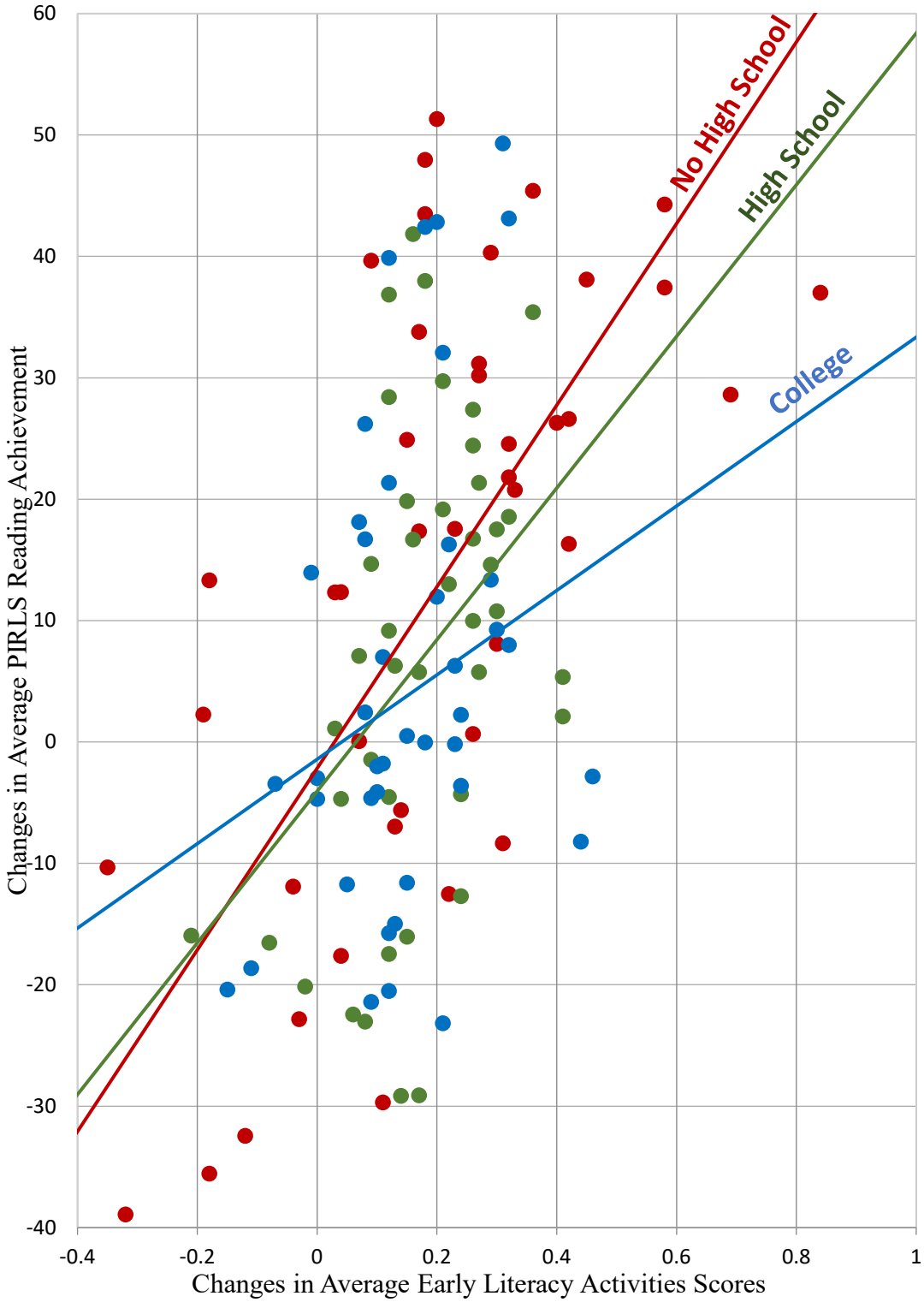


**Figure 4.15: Relationship Between Changes in Average Early Literacy Activities and Changes in Average Reading Achievement for the No High School Parental Education Group**

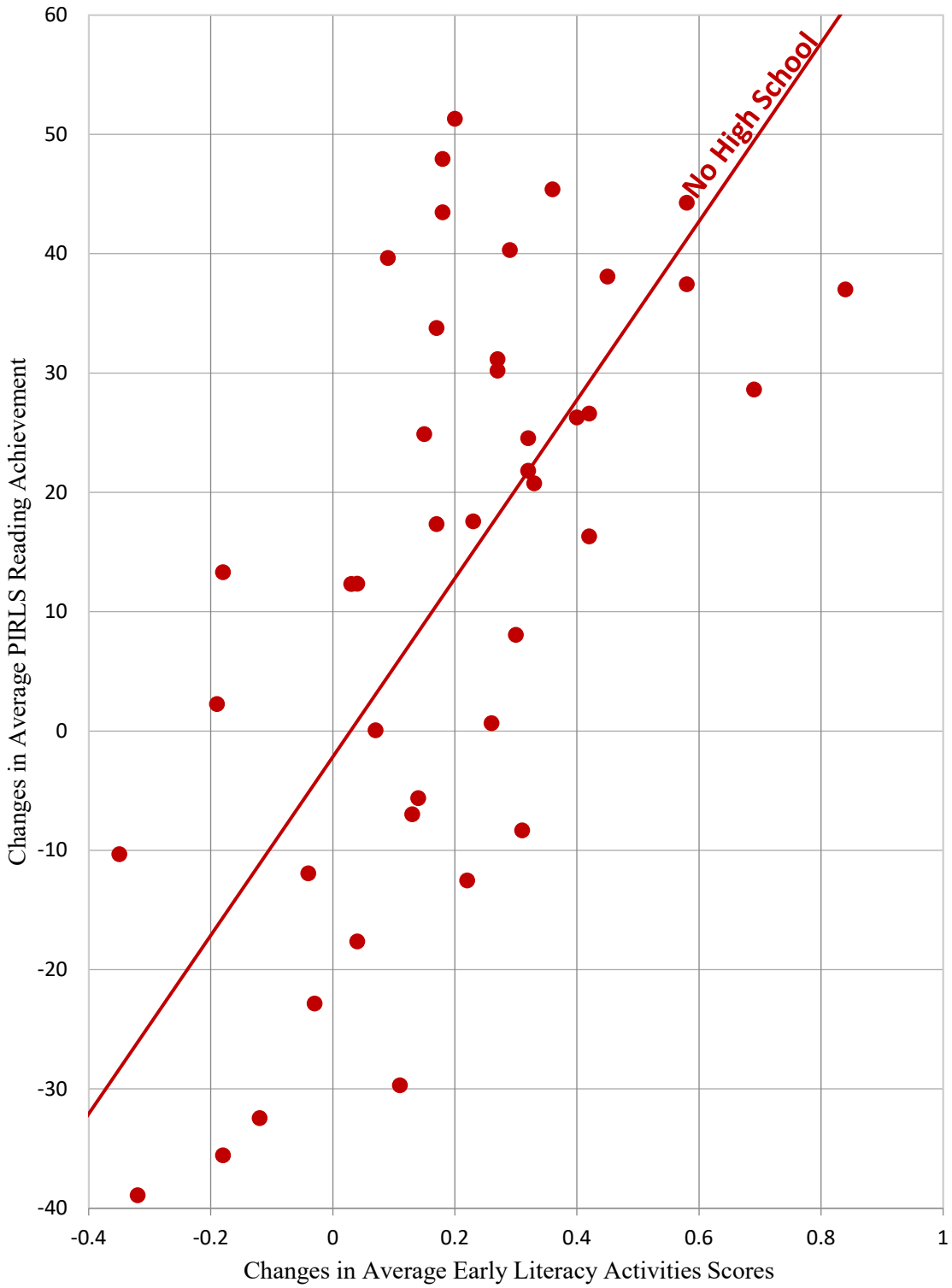
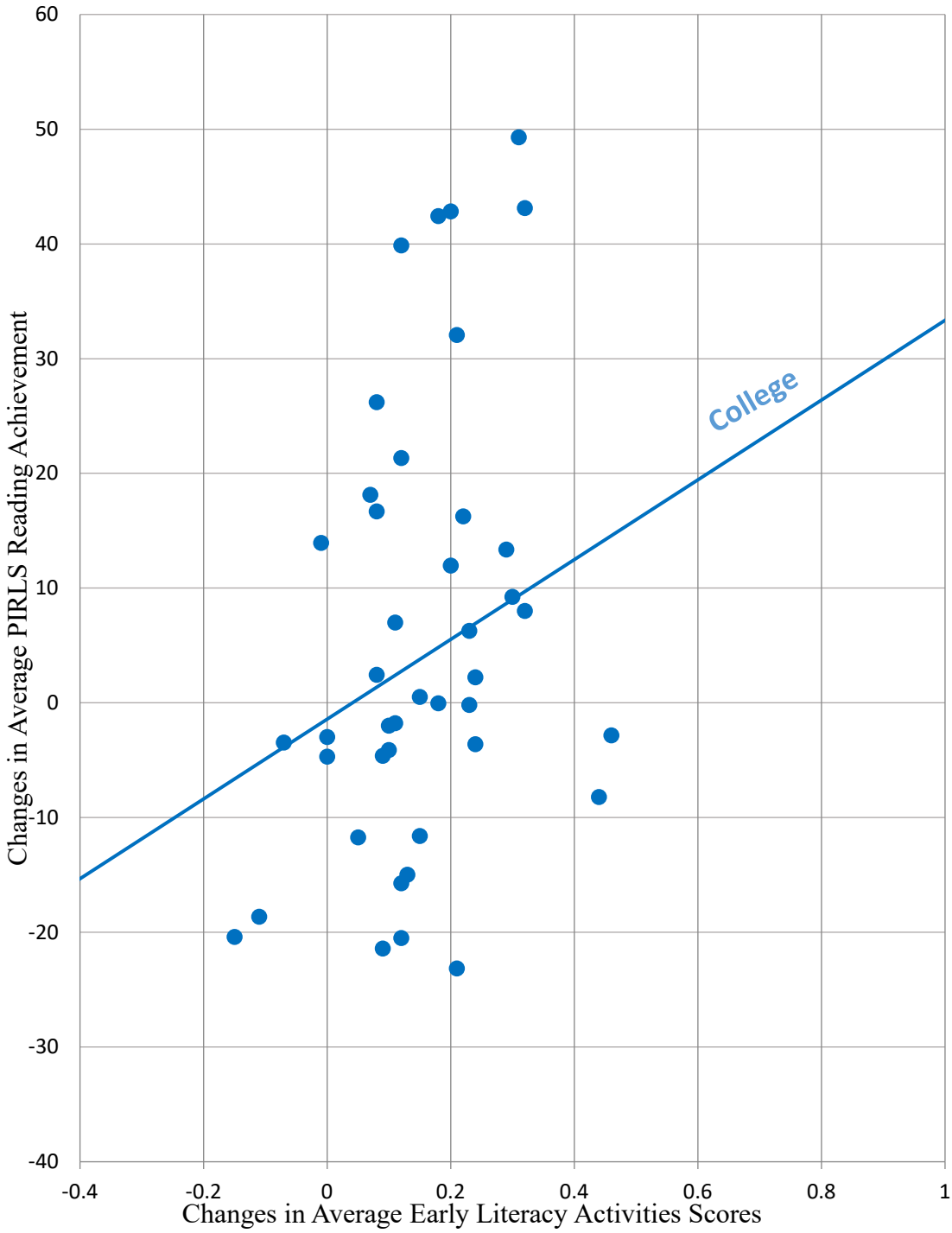


**Figure 4.16: Relationship Between Changes in Average Early Literacy Activities and Changes in Average Reading Achievement for the College Parental Education Group**



## 4.7 Phase 7: Mediation Analysis through a Random-Effects Model

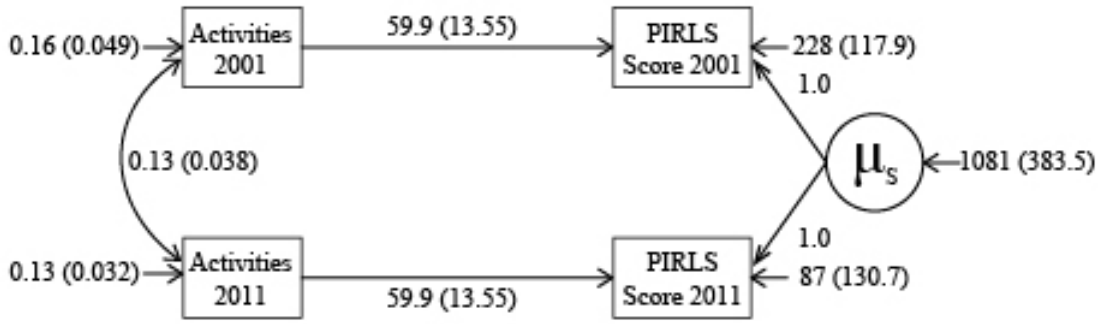
Another advantage of the proposed subpopulation approach is that it can be implemented to conduct subgroup mediation analysis using the structural equation modeling approach.

**Moving to a Random-Effects Model.** In order to move from a fixed-effects approach to a random-effects approach, it is recommended that tests be conducted to ensure that the random-effects approach does not introduce undue bias to the coefficient estimates. Although this is usually done through a Hausman test, Bollen and Brand (2010) contend that this can be done in the structural equation modeling approach through an examination of the significance of the path between $Z_s$ and the explanatory variable and by comparing the fit statistics of the random-effects and the fixed-effects models. From Figure 4.5, it can be seen that the covariance between $Z_s$ and the early literacy activities variables for each of the time points appears to be small in magnitude and nonsignificant, suggesting that little bias would be introduced in moving to a random-effects model.

A random-effects model was then estimated by constraining the covariance parameters between the latent variable and early literacy activities to 0. Because in the random effects model the latent variable represents random effects instead of fixed effects, it is referred to as $\mu_s$ instead of $Z_s$. Figure 4.17 shows the results of this analysis. The coefficient representing the path from early literacy activities to reading achievement was estimated to be 59.9 (SE=13.55)—a slight decrease from the coefficient estimated through the fixed-effects model but well within the estimated standard error. The random-effects model showed excellent fit, with a $\chi^2$ model fit of 2.713 (df=3, p>0.05), an RMSEA of 0.00, and a CFI of 1.00. A $\chi^2$ difference test was conducted comparing the $\chi^2$ model fit across the fixed-effects and random-effects models. The $\chi^2$ difference

of 2.227 was found to be nonsignificant (df=2, p>0.05) implying similar fit across the two models. Given the nonsignificant covariance between $Z_s$ and the early literacy activities variables, the strong fit of the random-effects model, and the fact that the differences in coefficient estimates were small given the standard error, it was decided it would be plausible to move to the more parsimonious random-effects model.

**Figure 4.17: Path Model with the Subpopulation Random-Effects Model Results**

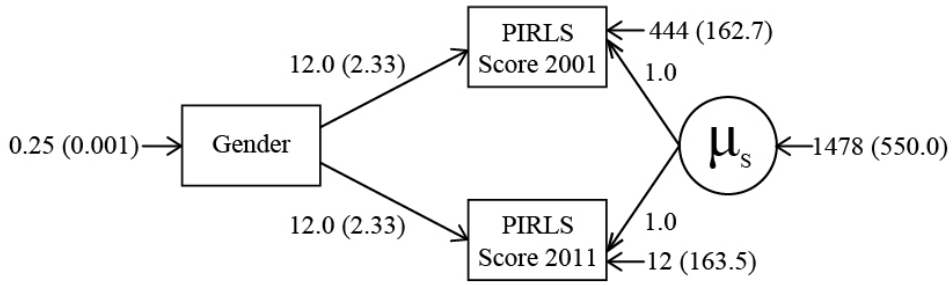



**Mediation Analysis.** In the random-effects model, a number of analysis options become possible, including mediation analysis. The first step in examining whether the relationship between student gender and reading achievement is mediated by early literacy activities is to examine the relationship between gender and reading achievement across the two PIRLS cycles without any other predictors in the model, by creating a model where gender predicts PIRLS reading achievement. This model assumed that the relationship between gender and reading achievement was fixed across time and therefore the path between gender and PIRLS 2001 reading achievement was constrained to be equal to the path between gender and PIRLS 2011 reading achievement.

As can be seen in Figure 4.18, the coefficient representing the dichotomous variable gender was 12.0 (SE=2.33), signifying that average girls subpopulation scores were 12 points

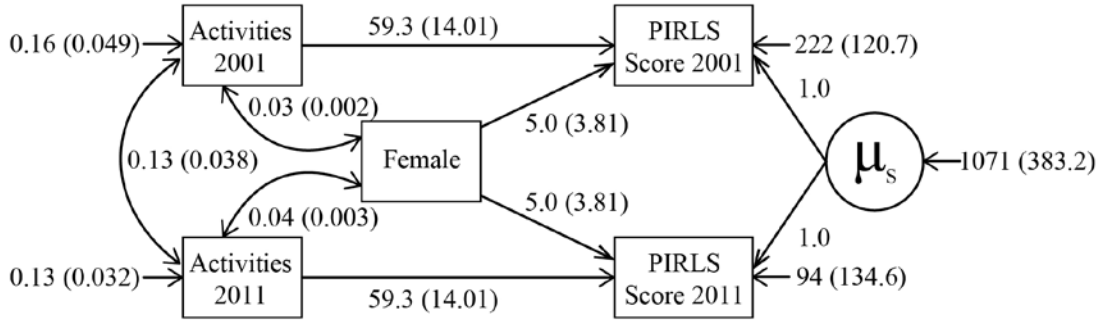
higher than the average boys subpopulation scores on the PIRLS assessment across the two assessment cycles. The 12-point difference between boys and girls is comparable to the 15-point and 11-point differences estimated for PIRLS 2001 and PIRLS 2011, respectively, in Model 2 of Phase 1—providing evidence that the subpopulation approach is able to capture much of the heterogeneity between boys and girls present in the data. The model, however, showed poor fit with a significant $\chi^2$ model fit of 8.986 (df=1, $p<0.05$) and a large RMSEA of 0.25. The CFI was acceptable at 0.99.

**Figure 4.18: Path Model Representing the Subpopulation Relationship Between Gender and PIRLS Reading Achievement**




The early literacy activities variables were then added to the model. Figure 4.19 shows the results. The early literacy activities coefficient of 59.3 (SE=14.01) showed little change from the random effects model with the introduction of the time-invariant covariate gender. However, the relationship between gender and reading achievement decreased dramatically to become nonsignificant at 5.0 (SE=3.81). The data showed strong fit to the model with a nonsignificant $\chi^2$ of 4.85 (df=4, $p>0.05$), an RMSEA of 0.04, and a CFI of 0.99.

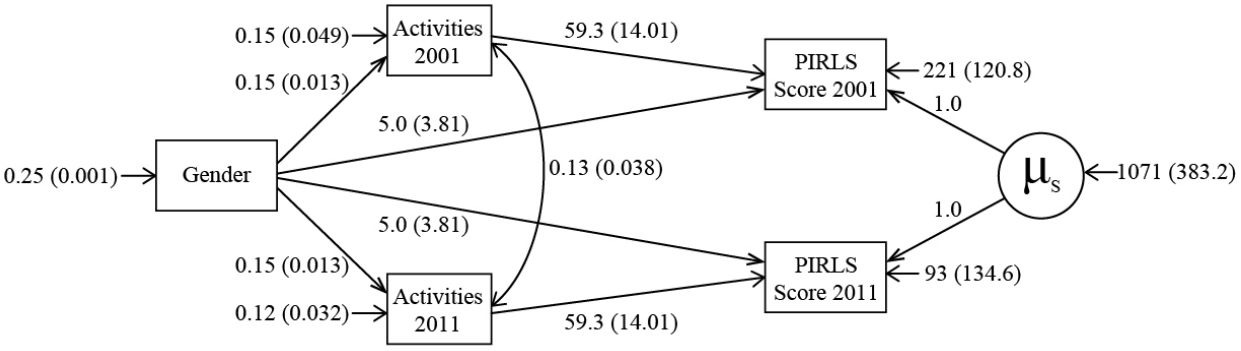
**Figure 4.19: Path Model Representing the Subpopulation Relationship Between PIRLS Reading Achievement and the Explanatory Variables Gender and Early Literacy Activities**




The dramatic decline of the path coefficient associated with gender suggests sizable covariance between the early literacy activities variables and gender. Figure 4.20 shows that there is significant covariance between gender and early literacy activities in both 2001 (0.30, SE=0.002) and 2011 (0.37, SE=0.003)—providing evidence that a mediation effect may be present in the data.

Building on the mediation research of Gustafsson et al. (2013), the full mediation model tests whether there is an indirect effect with early literacy activities mediating the positive relationship between girls advantage in PIRLS reading achievement. Figure 4.20 shows the mediation model. The significant coefficient associated with early literacy activities remains unchanged at 59.3 (SE=14.01), and the coefficient associated with gender remained nonsignificant at 5.0 (SE=3.81). The difference between this model and the previous model in Figure 4.19 is that it also estimates a regression path from gender to early literacy activities. The path coefficient associated with this additional path is 0.15 (SE=0.013), implying that girls have scores on the early literacy activities scale that are 0.15 of a standard deviation higher than boys. The model showed strong fit with a nonsignificant $\chi^2$ of 6.362 (df=5, p>0.05), an RMSEA of 0.04, and a CFI of 0.99.

**Figure 4.20: Path Model Representing Early Literacy Activities Mediating Relationship between Gender and PIRLS Reading Achievement**




The advantage of such a mediation model is that it becomes possible to estimate to what extent the difference in average reading achievement between boy and girl subpopulations is mediated by average early literacy activities scores. The results show a total effect of gender on reading achievement of 13.9 (SE=3.09) that includes a significant indirect effect of 8.8 (SE=2.26) through early literacy activities. Following from this, the model shows that early literacy activities explains over 60% of the gender differences in PIRLS reading achievement.

# Chapter 5: Discussion

The contribution of this dissertation is extending the country-level difference-in-differences approach, which has been applied extensively in the international large-scale assessment context (Gustafsson, 2013; Gustafsson & Nilsen, 2016; Hanushek et al., 2013; Liu et al., 2014; Rosén & Gustafsson, 2014; Rosén & Gustafsson, 2016), to analyze data at lower levels of aggregation. Because the country-level approach is the established approach, this dissertation has framed the subpopulation approach as an extension of this understood methodology.

Another way to frame both the country and subpopulation approaches would be as pseudo-panel approaches, with the difference between the two methodologies being that the country-level approach groups students solely based on the country identifier and the subpopulation-approach includes the country identifier and additional demographic variables related to the subpopulations of interest.

The primary question examined in this dissertation is whether there is benefit to aggregating to subpopulations below country level. Based on the literature review in Chapter 2 and the results from Chapter 4, the advantages are twofold:

- Including subpopulation-level data provides important information that strengthens the analysis;
- The subpopulation approach provides opportunities for analysis of subgroup differences.

## 5.1 Comparing Cross-Sectional and Longitudinal Approaches across Levels of Aggregation

For researchers seeking to make strong inferences about observed relationships using individual data, it is generally understood that longitudinal analysis is superior to cross-sectional analysis because longitudinal analysis allows the researcher to focus analysis on individual-level changes over time, with fixed-effects analysis providing the opportunity to control for time-invariant characteristics such as demographic variables or prior ability. Given the benefits of longitudinal analysis, Gustafsson (2007) proposed taking advantage of the repeated cross-sectional design of international large-scale assessment data to analyze the data longitudinally. He contended that treating international large-scale assessment data as a country-level longitudinal model would aid causal interpretation by better controlling for omitted variables—lessening the number of rival explanations for the key relationships in the data.

Because the fixed-effects approach controls for time-invariant influences, both measured and unmeasured, it is expected that it provides different regression coefficient estimates when compared with cross-sectional analyses. In Gustafsson's (2007) analysis of class size, he finds a sizeable contrast between the difference-in-differences results and those found in cross-sectional studies, with the difference-in-differences results aligning with those found in studies that use random assignment—providing evidence that the country-level longitudinal approach may come closer to estimating the true individual-level relationships than cross-sectional approaches.

A shortcoming of the country-level approach is that it examines relationships at high-levels of aggregations, and a threat to the validity of inferences drawn through Gustafsson's (2007) approach is that the country-level relationships may not be the same if the data were analyzed at lower-levels of aggregation.

For these reasons, initial analyses in Phase 1 through Phase 4 provide regression coefficient estimates associated with the relationship between early literacy activities and PIRLS reading achievement across cross-sectional and longitudinal approaches and across different levels of aggregation. Because the true individual-level relationship between early literacy activities and PIRLS reading achievement is unknown, it is not possible to ascertain the approach that comes closer to the true relationship. Nevertheless, a comparison of coefficient estimates can provide insights into similarities and differences in estimates across the approaches.

Phase 1 analyzed PIRLS data for both the 2001 and 2011 cycles using a multilevel model. From the Phase 1 analysis, there is evidence of a positive relationship across countries between the frequency of participation in early literacy activities and PIRLS reading achievement, even after controlling for student gender, parental education, parents like reading, and the duration of preprimary attendance. With these controls, the results show that an increase of one standard deviation across students on the early literacy activities scale predicts a 6.4 (SE=0.83) point reading achievement increase in PIRLS 2001 and a 6.9 (SE=0.69) point increase in PIRLS 2011. This multilevel model shows that high achieving students tend to have been engaged more by their parents in early literacy activities at a young age than their lower achieving peers.

Phase 2 analysis, which examined the country-level relationship between mean early literacy activities scores and mean PIRLS reading achievement, produced a nonsignificant coefficient of 27.3 (SE=21.46) in PIRLS 2001 and 18.9 (SE=22.14) in PIRLS 2011, after controlling for parents like reading and the duration of preprimary attendance. For the PIRLS 2011 analysis a few outlier countries were influential, and after removing these countries the coefficient increased to 29.1 (27.14). Although the coefficients across the cycles were

nonsignificant, it must be understood that the design was underpowered due to the small number of countries included in the analysis. The magnitude of the coefficients (27.3 and 18.9/29.1) reflect the tendency of higher achieving countries to have higher average scores on the early literacy activities scale.

In Phase 3, the country-level difference-in-differences approach was implemented using fixed-effects regression, and the results showed a significant unstandardized coefficient estimate of 85.2 (SE=26.46) for early literacy activities after controlling for the time-varying covariates parents like reading and duration of preprimary attendance. The results suggest that a one point increase in a country's early literacy activities average score is associated with an 85-point increase in a country's mean reading achievement.

Phase 4 demonstrated the proposed subpopulation approach. The Phase 4 results, based on the subpopulation approach to difference-in-differences, estimated a regression coefficient of 64.8 (SE=14.38) for early literacy activities when controlling for parents like reading and duration of preprimary attendance—meaning a one point increase in a subpopulation's early literacy activities mean score is associated with a 65 point increase in mean reading achievement.

In contrast to Gustafsson's (2007) example analysis where he found differences in directionality of the relationships when comparing difference-in-differences results to cross-sectional analysis results, the directionality of the estimated relationships in the present analysis is consistent across the four phases. As such, the results from the longitudinal approaches in this particular example complement the evidence for the efficacy of early literacy activities found in cross-sectional analyses.

Nevertheless, a takeaway from these four analyses is that the estimated regression coefficients do vary depending on the approach. The Phase 1 and Phase 2 cross-sectional analyses produce regression coefficients of less than 10 and around 20-30 points, respectively, and the longitudinal approaches in Phase 3 and Phase 4 produce larger regression coefficients of 85 and 65, respectively. The variations provide evidence that longitudinal approaches produce different results when compared with cross-sectional analysis approaches.

Comparing regression coefficients in Phase 1 and Phase 2, it is noticeable that the magnitude of the regression coefficients associated with early literacy activities are higher in the aggregated country-level model in Phase 2. The early literacy activities regression coefficient is also higher for country-level difference-in-difference in Phase 3 than it is in the subpopulation approach in Phase 4. The differences in coefficients across the levels align with the findings of Robinson (1960), who found that coefficient estimates vary drastically across levels of aggregation.

Another notable difference between the results from Phases 1 through 4 are the changes in the standard errors associated with the early literacy activities coefficient. Because standard errors are a function of sample size, the difference in the magnitude of the standard errors between the models can be attributed to differences in the number of primary units included in the analysis. For Phase 1, the primary unit of analysis is student data and in 2001 there were 91,834 students in the sample and in 2011 there were 97,799 students. In Phase 2 and Phase 3, analysis was conducted at country-level and this analysis across the 21 countries had the largest standard errors. For Phase 4, the primary unit of analysis are the 126 subpopulations, and although cluster-robust standard errors were implemented in the analysis, the standard errors were about half the size of the standard errors in Phase 2 and Phase 3.

## 5.2 Capturing Within-Country Relationships

After demonstrating the subpopulation approach in Phase 4, Phase 5 delved deeper into the subpopulation data to examine to what extent it provides additional vantage points on the relationship between early literacy activities and PIRLS reading achievement. Preliminary analysis examined the intraclass correlation coefficient, estimating the variance across subpopulation- and country-levels for early literacy activities and PIRLS reading achievement. The results showed that 43% of the variance in early literacy activities and 25% of the variance in reading achievement were at subpopulation level.

Although the intraclass correlation coefficient shows that there is variance in both the explanatory and outcome variables at subpopulation level, this variance could be error variance. Since Deaton (1985), econometricians have acknowledged that the difference scores could be error-ridden due to the sample size of each of the subpopulations, and they have debated the number of respondents needed per subpopulation to provide stable difference score estimates. Veerbeek and Nijman (1992) suggested that 100 respondents were needed in each subpopulation, and this rule of thumb has been widely applied. However, Devereux (2007) suggested that 2000 respondents may be necessary to avoid small sample bias. In the case that the explanatory variable difference scores have a lot of sampling error, the relationship between the explanatory variable and outcome variable would be attenuated, and this attenuation biases downwards the estimates of the regression coefficients. This is a plausible explanation for the decrease in the regression coefficient from country difference-in-differences in Phase 3 to the subpopulation difference-in-differences in Phase 4.

If the variance at lower levels were primarily error variance, random patterns would be expected in the data. For this reason, Phase 5 examined the within-country data to gain insights

into whether the relationship between early literacy activities and PIRLS reading achievement is present at lower levels of aggregation.

Scatterplots were then created to explore the within-country dispersion between subpopulations. Further descriptive analysis graphed the regression lines for each of the 21 countries across three scatterplots—with Figure 4.9 showing the nine countries with a steep positive relationship between changes in subpopulation means on early literacy activities and PIRLS reading achievement, Figure 4.10 the six countries having a slightly positive relationship, and Figure 4.11 the six countries that have a negative relationship. Interestingly, four of the five countries with little dispersion in early literacy activities also show a negative relationship between early literacy activities and PIRLS reading achievement.

It is well documented that differential changes in the explanatory variable among subpopulations are necessary to see relationships in the data (Verbeek, 2008b). In these five countries, the very small differences between the subpopulations in early literacy activities could be due solely to random error. If these five countries with the least dispersion between the subpopulations are removed from the analysis, the regression lines associated with 14 of the remaining 16 countries would have a positive slope—the mean reading achievement of a subpopulation increases as the mean early literacy scores increase.

Another way the subpopulation data were examined was by decomposing the variance and analyzing whether there is a subpopulation-level relationship between mean early literacy activities and mean PIRLS reading achievement across the pooled set of countries. The results showed a significant coefficient of 39.0 (17.58) associated with early literacy activities. This coefficient estimate from Phase 5 is smaller than what was found in Phase 3 and Phase 4 analysis

but leads to the same conclusion—increases in average early literacy activities predicts increases in average PIRLS reading achievement.

In summary, the Phase 5 analysis was conducted to examine whether the relationships at subpopulation level were relevant for analysis, with concerns that the data may be noisy due to the fact that the subpopulations are composed of smaller samples than the countries. The results showed that when there was substantial dispersion in early literacy activities, it was possible to see a within-country relationship between early literacy activities and PIRLS reading achievement for most of the countries as well as a relationship when analyzing the pooled within-country data across countries. The analysis provides evidence that similar relationships to those found a country-level can be captured at subpopulation level.

Finding a similar relationship at country-level to those at subpopulation-level, however, raises the question of whether subpopulation analysis produces redundant information to the country-level approach. In this particular example, the results are not substantially different across levels of aggregation for the longitudinal approaches and the conclusions would be the same—increases in mean early literacy activities are associated with increases in mean reading achievement. Because country-level analysis is easier to perform and less complicated to explain, the question becomes: Is subpopulation-level analysis worth the complications?

This perspective, however, misses the point that there are unique stories to tell within the subpopulation data. For example, referring back to the Figure 4.3—the scatterplot of country-level changes in early literacy activities by country-level changes in PIRLS reading achievement divided into quadrants, six countries do not fit the expected pattern. France, Hungary, Lithuania, the Netherlands, and Sweden show a country-level increase in activities and decrease in

achievement, and Bulgaria shows a country-level decrease in activities and slight increase in achievement. Following from this, one may conclude that increasing early literacy activities is associated with increases achievement in some countries but not others. However, delving into the subpopulation data, it is noticeable that all six of these countries have positive within-country regression lines in Figures 4.9 and 4.10—suggesting that within each of the countries, the subpopulations that had the largest increase in average early literacy activities scores also had the largest increases in PIRLS reading achievement.

This example provides evidence that there may be much to learn from subpopulation level data even if the pooled results seem to confirm the country-level analysis. Future analysis implementing this technique could examine the relationship between other educational practices and student achievement to see to what extent the subpopulation- and country-level relationships align.

## 5.3 Examining Differential Relationships across Subgroups

The purpose of the Phase 6 analysis was twofold:

(1) To examine whether the relationship between early literacy activities and PIRLS reading achievement, as represented by the regression coefficients, is heterogenous across subgroups; and

(2) To demonstrate an opportunity for analyzing differences in coefficient estimates using the Mplus multiple group approach.

Phase 6 applied the multiple-group, fixed-effects approach to examine whether there is heterogeneity in coefficient estimates among gender and highest parental education groups using

data pooled across the 21 countries. The multiple group analysis technique applied to difference-in-differences was first implemented by Gustafsson and Nilsen (2016), when they examined whether there were differences in regression coefficient across OECD and non-OECD countries on 21 measures of instructional quality. The results of the 21 analyses did not show any significant differences between OECD and non-OECD countries.

Gustafsson and Nilsen (2016) analyzed a sample of 38 countries. Given that the subpopulation analysis involves 126 units, the subpopulation analysis should have more power to detect statistically significant differences than Gustafsson and Nilsen's (2016) model.

The results of the Phase 6 analysis showed a similar relationship between changes in mean early literacy activities and changes PIRLS reading achievement across male and female subgroups. However, sizeable differences in coefficient estimates were found when comparing parental education groups, with a coefficient estimate of 74.8 (SE=12.76) associated with subpopulations whose parents do not have a high school degree, a coefficient estimate of 62.4 (SE=20.64) for subpopulations whose parents highest education level was high school graduate, and a coefficient of 34.8 (SE=29.84) for subpopulations whose highest parental education level was college graduate. Large standard errors were associated with the coefficient estimates—especially for the college parental education group, and the model fit comparisons and the Wald test did not reject the null hypothesis of equal regression coefficients. The large standard errors associated with the coefficient estimate for the college parental education group appear to be related to the large amount of residual variation for this group and the relatively small sum of the squared deviations for this group in early literacy activities.

It is debatable how to interpret the results from the Phase 6 multiple group analysis. Given the emerging consensus that p-values should not be the only criteria for drawing conclusions, it could be argued that the differences in coefficient estimates provide some evidence of heterogeneity in the data, although it is unclear to what extent the differences in coefficient estimates are due to chance alone.

## 5.4 Mediation Modeling Applications

Phase 7 illustrates an opportunity provided through the subpopulation approach for mediation modeling of subgroup differences. Mediation modeling with time-invariant covariates necessitates leaving the fixed-effects approach, and therefore initial analysis was conducted to confirm it was possible to adopt a random-effects approach without unduly biasing coefficient estimates.

Mediation analysis was then conducted and the results of the subpopulation analysis provided evidence that girls, on average, were more likely than boys to engage in early literacy activities with their parents, and that their greater likelihood to partake in such activities explains over 60% of the girls' advantage on the PIRLS reading assessment internationally, confirming the research of Gustafsson et al. (2013). This longitudinal model, however, explained much more of the reading gap than Gustafsson et al.'s (2013) cross-sectional study. The link between differences in early literacy activities across the genders and later differences in PIRLS reading achievement has now been well documented using PIRLS data across both cross-sectional and longitudinal approaches. Although others like Bertrand and Pan (2013) and Millard (2003) have also found that girls participate more in early literacy activities, these two studies do not provide empirical evidence that links these differences in literacy activities to later differences in reading

achievement. As such, these PIRLS results are unique as they provide an interesting explanation for the reading achievement gap and should be followed up with additional longitudinal analysis.

It is important to keep in mind that the Phase 7 analysis applied the random effects approach and therefore the results reflect both longitudinal and cross-sectional relationships in the data. Although there is extensive literature on the perils of inducing bias when switching to random-effects analysis (Allison 2005, 2009), it is possible to switch to a random effects model without biasing coefficient estimates when certain assumptions are supported. In this case, the assumptions were fulfilled, but it is more often the case that the random-effects assumptions are not supported, and therefore switching to a random-effects model to conduct such mediation analysis may bias the coefficient estimates.

## 5.5 Difference-in-Differences and Causal Inference

Taken together, the results from the early literacy activities analysis could be used to make the argument for justifying programs that encourage parents to engage both girls and boys (and especially boys) more in early literacy activities. The analyses from Phase 1-5 triangulate evidence that early literacy activities are positively related to student reading achievement at the fourth grade across cross-sectional and longitudinal approaches and across different levels of aggregation. The evidence from Phase 7 suggests that parents engage boys less frequently in such activities, and this could contribute to differences in reading achievement at the fourth grade. Likewise, the evidence in Phase 6 suggests that when boys and girls are engaged in early literacy activities their engagement tends to have a similar relationship with reading achievement.

Nevertheless, causal claims should be tempered because the models do not control for time-varying omitted variables. Classical difference-in-differences allows for causal inference when the common trend assumption is met, because showing common trends provides strong evidence that other time-varying factors are not influencing the outcome variable. From Figure 2.8 from Chapter 2, it can be seen that the common trend assumption is not supported across countries when examining PIRLS data since the country trend lines do not appear to be parallel.

Without fulfilling the common trend assumption, it becomes difficult to dismiss the influence of time-varying covariates as innumerable factors could lead to increases or decreases in student achievement across countries. One step in the right direction would be to include a number of measured covariates in the analysis. An advantage of the subpopulation approach over the country-level approach in this regard is that the subpopulation approach offers a larger sample size and therefore more degrees of freedom for including such time-varying covariates.

Nevertheless, for analysis across countries, it remains nearly impossible using either the country or subpopulation groupings to control for all plausible influences by including measured covariates in the model. One possible extension on the subpopulation approach would be to implement sensitivity analysis (Montgomery, Richards, & Braun, 1986; Rosenbaum & Rubin, 1983). In sensitivity analysis, a hypothetical omitted variable is generated that is related to both the explanatory variable of interest for the analysis and the outcome variable. By generating this theoretical variable, which has varying correlations with both the explanatory and outcome variables, and inserting this hypothetical variable into the analysis model as a covariate, researchers are able to approximate the relationships the omitted variable must have with the explanatory and outcome variables to change the direction, magnitude, and/or significance of the coefficient estimate of interest.

Another possibility for strengthening causal inference through these longitudinal approaches is to focus analysis on countries that tend to have parallel-trend lines. Similar to economists assuming that the economies of adjacent states have common trends, it may be possible to assume that common trends are expected for countries that are from the same geographic region and have shown common trend in the past. Referring back to Figure 2.8, Singapore and Hong Kong have a very similar trend line from PIRLS 2001 through PIRLS 2011. If Singapore were to have introduced a new education policy in 2012, one could assume common trend with Hong Kong and examine the efficacy of the policy by comparing the deviations in the trend line between PIRLS 2011 and PIRLS 2016.

It should be kept in mind that although the Phase 7 mediation analysis provides additional perspectives on the relationships in the data, it provides a similar level of causal evidence (or lack their of) to the country and subpopulation difference-in-differences approaches in Phase 3 and Phase 4. The advantage of the mediation modeling approach in Phase 7 is that it provides an explanation for girls' advantage on the PIRLS reading assessment. Because there was no random assignment, however, this explanation is still primarily dependent on the theory that more engagement in early literacy activities caused higher reading achievement. To draw causal conclusions it would still be necessary to dismiss other possible explanations for the relationship between increased participation in early literacy activities and increased reading achievement.

From a causal perspective, it should be noted that more could be done in the Phase 7 model to strengthen causal inferences by controlling for omitted variables, such as adding time-varying covariates. More demographic characteristics could also be controlled for in the random effects approach. Because the current analysis used the sandwich estimator in Mplus to estimate cluster-robust standard errors, the number of parameters estimated could not exceed the number

of clusters—equal to the 21 countries in the analysis—and therefore there was a limit to the number of covariates that could be added to the analysis. In more recent cycles of TIMSS and PIRLS, participation has reached around 50 countries at the fourth grade, opening up the possibility for creating even more complex models through the subpopulation approach and controlling for additional covariates.

## 5.6 Limitations and Future Research

This dissertation connects Gustafsson's (2007) country-level difference-in-differences approach with Deaton's (1985) pseudo-panel approach and thereby provides a new approach for analyzing international comparative assessment data longitudinally. In connecting these two methodologies, this dissertation opens up a number of interesting questions that should be kept in mind when employing this approach and provides interesting opportunities for future methodological research.

(1) **How can aggregation bias be mitigated?** Similar to the country-level approach, the subpopulation results could still suffer from aggregation bias especially because the variables used in the aggregation process (e.g., country, gender, and parental education) are likely related to changes in reading achievement after controlling for the effects of early literacy activities. In the context of international large-scale assessments it is difficult to find instrumental variables—variables related to the explanatory variable and not related to the error term, which can be used as subpopulation identifiers. One area for future research is to examine what other TIMSS and PIRLS variables can be used as subpopulation identifiers. In addition, future research could examine ways in which covariates could be employed in the

aggregation process in order to decrease the relationship between the subpopulation identifier and the error term (King, 1997).

(2) **What is the optimal sample size of the subpopulations?** As described by Baltagi (1995), deciding how fine-grained to make the subpopulations is a tradeoff between capturing the analysis-relevant heterogeneity in the data and ensuring that the sample size for each subpopulation is sufficient for stable difference score estimates. The problem with small subpopulations, as argued in the econometrics literature, is that the difference score estimates become less stable with smaller groups.

The results of the Phase 3 and Phase 4 analyses show that the magnitude of the coefficient estimates for the subpopulation analysis is smaller than for the country analysis. It is unclear whether the smaller coefficient estimate for the subpopulation approach is converging to the true student-level relationship or whether the coefficient is smaller due to attenuation linked to error in the difference score estimates for the explanatory variable.

Future analysis should explore to what extent dividing countries into subpopulations aids analysis and to what extent this process attenuates relationships in the data. Phase 5 of this dissertation has provided some evidence that signal of a relationship between early literacy activities and reading achievement exists at subpopulation level. Further analysis could examine how relationships change as the size of the subpopulations decrease, as well as methods for quantifying the error in the subpopulation data such as Deaton's errors-in-variables estimator. Research could also use actual panel data from national large scale assessments to examine to what extent the pseudo-panel

results at varying levels of aggregation are able to close in on the results of longitudinal analysis of student-level data.

(3) **Can coefficient estimates be stabilized through empirical Bayes?** Absent in the econometrics debate around the optimal sample size for subpopulations seems to be the possibility of applying empirical Bayes to stabilize the difference score for small subpopulations. Through the use of empirical Bayes methodology, it may be possible to condition the subpopulations difference score estimates using the subpopulation identifiers and in so doing improve the reliability of the difference scores. By stabilizing the difference score estimates, it may be possible to create much more fine-grained subpopulations while minimizing the attenuation of the relationships due to error in the estimation of the explanatory variable.

In addition to these three areas that are closely related to the focus of this dissertation, the example analysis expanded upon Gustafsson and Nielson's (2016) structural equation modeling approach to difference-in-differences to include mediation analysis through a random effects model. As the structural equation modeling approach to difference-in-differences is still relatively unknown, future work should document the lessons learned and opportunities available for implementing this approach using international large-scale assessment data.

## 5.7 Conclusions

Since TIMSS began in 1995, enormous efforts have been made to maintain a trend line for achievement data at both the fourth and eighth grades, with TIMSS 2015 marking 20 years of trends. Similarly, PIRLS 2016 marks 15 years of trend with the upcoming release of the *PIRLS International Results in Reading* report in December 2017. Over this time, the TIMSS and

PIRLS datasets have offered researchers innumerable opportunities to explore relationships in the data and, specifically, predictors of student achievement. Nevertheless, analyses are generally conducted on one cycle of data, overlooking the possibilities for taking advantage of the repeated cross-sectional design of the studies.

The subpopulation approach provides a new methodology for researchers to examine international large-scale assessment data. When contrasted with cross-sectional analysis approaches, the subpopulation approach like country-level difference-in-differences provides the opportunity to control for factors that do not change over time—strengthening the causal argument. The pseudo-panel approach together with Hanushek and Wößmann's (2006) vertical difference-in-difference approach, Gustafsson's (2007) horizontal approach, and De Simone's (2013) imputed regression pseudo-panel methodology provide educational researchers with a useful toolbox of possibilities for pseudo-longitudinal data analysis using international large-scale assessment data.

## References


Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS.* Cary, N.C.: SAS Institute.

Allison, P. D. (2009). *Fixed effects regression models.* Thousand Oaks, CA: Sage Publications.

Allison, P. D., & Bollen K. A. (1997). Change Score, Fixed Effects and Random Component Models: A Structural Equation Approach. Paper presented at the Annual Meeting of the American Sociological Association, Toronto, Ontario, Canada, August 9-13, 1997.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'Metrics.* Princeton, NJ: Princeton University Press.

Baker, L. & Scher, D. (2002). Beginning readers' motivation for reading in relation to parental beliefs and home reading experiences. *Reading Psychology, 23*(4), 239–269.

Baltagi, B. H. (1995). *Econometric analysis of panel data*. West Sussex, England: John Wiley & Sons Ltd.

Bertrand, M. & Pan, J. (2013). The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, *5*(1), 32-64.

Blundell, R., Megir, C., & Neves, P. (1993). Labor supply and intertemporal substitution. *Journal of Econometrics*, *59*, 137-160.

Bollen, K. A., & Brand, J. E. (2008). Fixed and Random Effects in Panel Data Using Structural Equations Models. (*California Center for Population Research. On-Line Working Paper Series No. PWP-CCPR-2008-003*). Los Angeles, CA: University of California – Los Angeles.

Bollen, K. A. & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces*, *89*(1), 1-34.

Browning, M., Deaton, A., & Irish, M. (1985). A profitable approach to labor supply and commodity demands over the life cycle. *Econometrica, 53*(3), 503-544.

Card, D. (1992). Using regional variation in wages to measure the effects of the federal minimum wage. *Industrial and Labor Relations Review, 46*(1), 22-37.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review, 84*(4), 772-793.

Caro, D. H. (2015). Causal mediation in educational research: An illustration using international assessment data. *Journal of Research on Educational Effectiveness, 8*(4), 577-597.

Chmielewski, A., & Dhuey, E. (2017). *The analysis of international large-scale assessments to address causal questions in education policy*. Paper commissioned by the National Academy of Education. Retrieved from https://naeducation.org/ilsa-workshop-agenda-june-september/

Choi, A., Gil, M., Mediavilla, M., & Valbuena, J. (2016a). *Double toil and trouble: Grade retention and academic performance* (IEB Working Paper 2016/07). Barcelona: Institut d'Economia de Barcelona.

Choi, A., Gil, M., Mediavilla, M., & Valbuena, J. (2016b). *The evolution of educational inequalities in Spain: Dynamic evidence from repeated cross-sections* (IEB Working Paper 2016/25). Barcelona: Institut d'Economia de Barcelona.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Deaton, A. (1985). Panel data from time-series of cross-sections. *Journal of Econometrics, 30*, pp. 109-126.

Deaton, A. (1997). *The analysis of household surveys: A microeconometric approach to development policy.* Baltimore: John Hopkins University Press.

De Simone, G. (2013). Render unto primary the things which are primary's: Inherited and fresh learning divides in Italian lower secondary education. *Economics of Education Review*, *35*, 12-23.

Devereux, P. J. (2007). Small sample bias in synthetic cohort models of labor supply, *Journal of Applied Econometrics*, *22*(4), 839-848.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Prospectives*, *27*(2), 109-132.

Foy, P., & Drucker, K. T. (2013). *PIRLS 2011 user guide for the international database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Foy, P., Brossman, B., & Galia, J. (2012). Scaling the TIMSS and PIRLS 2011 achievement data. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011* (pp.1-28). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/methods/pdf/TP11_Scaling_Achievement.pdf

Ginsberg, A. & Smith, M.S. (2016). *Do randomized controlled trials meet the "gold standard"? A study of the usefulness of RCTs in the What Works Clearinghouse.* Retrieved from American Enterprise Institute website: https://www.carnegiefoundation.org/wp-content/uploads/2016/03/Do-randomized-controlled-trials-meet-the-gold-standard.pdf

Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Gustafsson, J.-E. (2007). Understanding causal influences on educational achievement through differences over time within countries. In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 37-63). Washington D.C.: The Brookings Institute.

Gustafsson, J.-E. (2010). Longitudinal designs. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 77-101). New York, NY: Routledge.

Gustafsson, J.-E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, *24*(3), 275-295.

Gustafsson, J.-E. (2016). Longitudinal analysis and the potential for causal interpretations [pdf document]. Retrieved from https://naeducation.org/wp-content/uploads/2017/02/Gustafsson-June-17-2016-ILSA-Slides.pdf

Gustafsson, J.-E., Hansen, K. Y., & Rosén, M. (2013). Effects of home background on student achievement in reading, mathematics, and science at the fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Gustafsson, J.-E. & Nilsen, T. (2016). The impact of school climate and teacher quality on mathematics achievement: A difference-in-differences approach. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher quality, instructional quality, student outcomes* (pp. 81-95). Amsterdam, The Netherlands: IEA.

Hannan, M. T., & Burnstein, L. (1974). Estimation from grouped observations. *American Sociological Review*, *39*(3), 374-392.

Hanushek, E. A., Link, S., & Wößmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics*, *104*, 212-232.

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, *116* (510), C63-C76.

Hanushek, E. A., & Wößmann, L. (2017). School resources and student achievement: A review of cross-country economic research. In M. Rosén, K. Y. Hansen, & U. Wolff (Eds.), *Cognitive abilities and educational outcomes: A festschrift in honour of Jan-Eric Gustafsson* (pp. 149-171). Cham, Switzerland: Springer International Publishing.

Hart, B. & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, Maryland: Paul H. Brookes Publishing.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Husén, T. (1967). *International Study of Achievement in Mathematics. A Comparison of Twelve Countries. Volume I.* New York, NY: John Wiley & Sons.

Joncas, M. & Foy, P. (2012). Sample Design in TIMSS and PIRLS. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011* (pp.1-21). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf

Keppel, G., & Wickens, T. D. (2004). *Design and analysis. A researcher's handbook. Fourth edition.* Upper Saddle River, NJ: Pearson Prentice Hall.

King, G. (1997). *A solution to the ecological inference problem. Reconstructing individual behavior from aggregate data.* Princeton, NJ: Princeton University Press.

Kloosterman, R., Notten, N., Tolsma, J., & Kraaykamp, G. (2010). The effects of parental reading socialization and early school involvement on children's academic performance: A panel study of primary school pupils in the Netherlands. *European Sociological Review, 27*(3), 291–306.

Liu, H., Bellens, K., Gielen, S., Van Damme, J., & Onghena, P. (2014). A country level longitudinal study on the effect of student age, class size, and socio-economic status – Based on PIRLS 2001, PIRLS 2006, and PIRLS 2011. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 223-242). Muenster, Germany: Waxman.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Lawrence Erlbaum Associates.

Martin, M. O., Mullis, I. V. S., Foy, P. (2015). Assessment design for PIRLS, PIRLS Literacy, and ePIRLS in 2016. In I. V. S. Mullis & M. O. Martin (Eds.), *The PIRLS 2016 assessment framework, 2nd Edition* (pp. 55-69). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V. S., Foy, P., Arora, A. (2012). Creating and interpreting the TIMSS and PIRLS 2011 context questionnaire scales. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011* (pp.1-11). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/methods/pdf/TP11_Context_Q_Scales.pdf

Martin, M. O., Mullis, I. V. S., Foy, P., Hooper, M. (2016). *TIMSS 2015 International Results in Science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Melhuish, E. C., Phan, M. B., Sylva, K., Sammons, P., Siraj-Blatchford, I., & Taggert, B. (2008). Effects of the home learning environment and preschool center experience upon literacy and numeracy development in early primary school. *Journal of Social Issues, 64*(1), 95–114.

Millard, E. (2003). Gender and early childhood literacy. In N. Hall, J. Larson, & J. Marsh, *Handbook of Early Childhood Literacy*. London: Sage Publications.

Moffitt, R. (1993). Identification and estimation of dynamic models with a time series of repeated cross sections. *Journal of Econometrics*, *59*, 99-123.

Montgomery, M. R., Richards, T., and Braun, H. (1986). Child health, breast-feeding, and survival in Malaysia: A random-effects logit approach. *Journal of the American Statistical Association*, *81*(394), 297-309.

Morgan, S. L., and Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research. Second Edition.* New York, NY: Cambridge University Press.

Mullis, I. V. S., & Martin, M. O. (2015). Introduction. In I. V. S. Mullis & M. O. Martin (Eds.), *The PIRLS 2016 assessment framework, 2nd Edition* (pp. 3-9). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. (2012). *PIRLS 2011 international results in reading.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., Hooper, M. (2016) *TIMSS 2015 International Results in Mathematics.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Nagel, I., and Verbood, M. (2012). Reading behavior from adolescence to early adulthood: A panel study of the impact of family and education on reading fiction books. *Acta Sociologica, 55*(4), 351-365.

Nelson, C. A., & Sheridan, M. A. (2011). Lessons from neuroscience research for understanding causal links between family and neighborhood characteristics and educational outcomes. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 27-46). New York: Russell Sage Foundation.

Notten, N. & Kraaykamp, G. (2010). Parental media socialization and educational attainment: Resource or disadvantage? *Research in Social Stratification and Mobility*, *28*(4), 453-464.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* New York, NY: Cambridge University Press.

Punter, A., Glas, C. A. W., & Meelissen, M. R. M. (2016). *Psychometric framework for modeling parental involvement and reading literacy.* Amsterdam, The Netherlands: IEA.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods. Second edition.* Thousand Oaks, CA. Sage Publications.

Robinson, J. P. (2014). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski, *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 521-545). Boca Raton, FL: Chapman & Hall/CRC Press.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*(3), 351-357.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, 45*(2), 212-218.

Rosén, M. & Gustafsson, J.-E. (2014). Has the increased access to computers at home caused reading achievement to decrease in Sweden? In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 207-222). Muenster, Germany: Waxmann.

Rosén, M. & Gustafsson, J.-E. (2016). Is computer availability at home causally related to reading achievement in grade 4? A longitudinal difference in differences approach to IEA data from 1991 to 2006. *Large-scale Assessments in Education*, *4*(5), 1-19.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, *83*(5), 1762-1774

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Rutkowski, L. (2016a). *A look at the most pressing design issues in international large-scale assessments.* Paper commissioned by the National Academy of Education. Retrieved from https://naeducation.org/ilsa-workshop-agenda-june-september/

Rutkowski, L. (2016b). Introduction to special issue on quasi-causal methods. *Large-scale Assessments in Education, 4(*8), 1-6.

Rutkowski, D. & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-scale Assessments in Education, 4*(6), 1-18.

Schlotter, M., Scherdt, G., & Wößmann, L. (2014). Econometric methods for causal evaluation of education policies and practices: A non-technical guide. In R. Striethold, W. Bos, J.-E. Gustafsson, & M. Rosén. *Educational policy evaluation through international comparative assessments* (pp. 95-126). Muenster: Waxmann.

Sénéchal, M. & LeFevre, J.-A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development*, *73*(2), 445–460.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Belmont, CA: Wadsworth Publishing.

Strietholt, R., Gustafsson, J.-E. Rosén, M., & Bos, W. (2014). Outcomes and Causal Inference in International Comparative Assessments. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 1-18). Muenster, Germany: Waxman.

UNESCO Institute for Statistics. (2012). *ISCED: International standard classification of education.* Retrieved from UNESCO website: http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx

Verbeek, M. (2008a). *A guide to modern econometrics. Third Edition.* West Sussex, England: John Wiley & Sons Ltd.

Verbeek, M. (2008b). Pseudo-panels and repeated cross-sections. In L. Mátyás & P. Sevestre (Eds.). *The econometrics of panel data: Fundamentals and recent developments in theory and practice* (pp. 369-383). Berlin: Springer-Verlag.

Verbeek, M. & Nijman, T. (1992). Can cohort data be treated genuine panel data? *Empirical Economics, 17*(1), 9-23.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143-2152.

Wilson, S. E., and Butler, D. M. (2007). A lot more to do: The sensitivity of time-series cross-section analyses to simple alternative specifications. *Political Analysis*, *15*(2), 101-123.

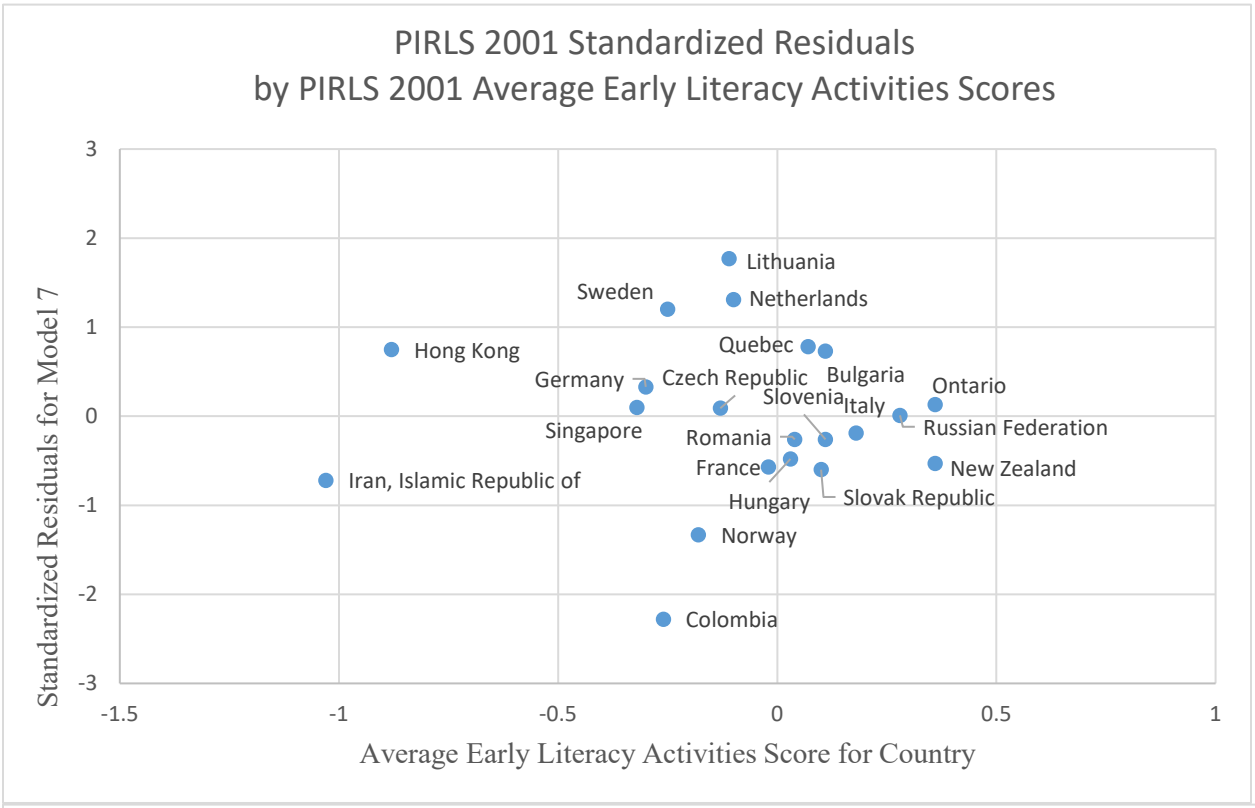
Woolbridge, J. W. (2010). *Econometric analysis of cross section and panel data. Second edition.* Cambridge, MA: MIT Press.

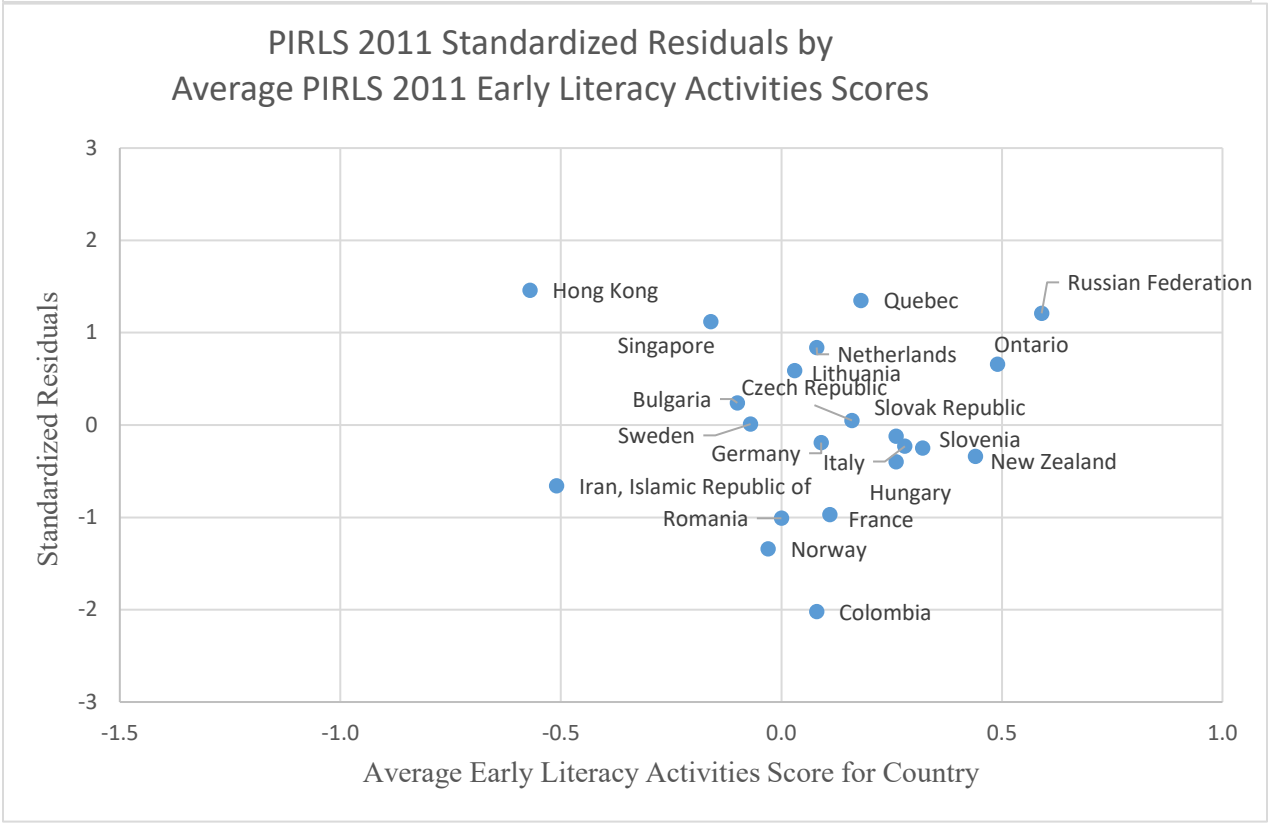
# Appendix A: Additional Analysis Details

## Phase 2: Analysis of Heteroscedasticity

Figure A.1 displays the Model 7 plots for PIRLS 2001 and PIRLS 2011. When heteroscedasticity is present, a cone shaped pattern would be expected where the variance in the residuals becomes larger at the tails of the distribution of the explanatory variable. The 2001 scatterplot shows that the largest residuals are around the center of the distribution, and the pattern looks relatively random, meaning that the analysis seems to fulfill the regression assumption of homoscedasticity. In the corresponding plot for PIRLS 2011, the variance related to the standardized residuals appears to be relatively uniform across the early literacy activities distribution, and no cone shape patterns exist in the data. The 2011 results also imply that the regression assumption of homoscedasticity is fulfilled.

**Figure A.1: Scatterplots for Evaluating Country-Level Heteroscedasticity in PIRLS 2001 and PIRLS 2011 for Model 7**

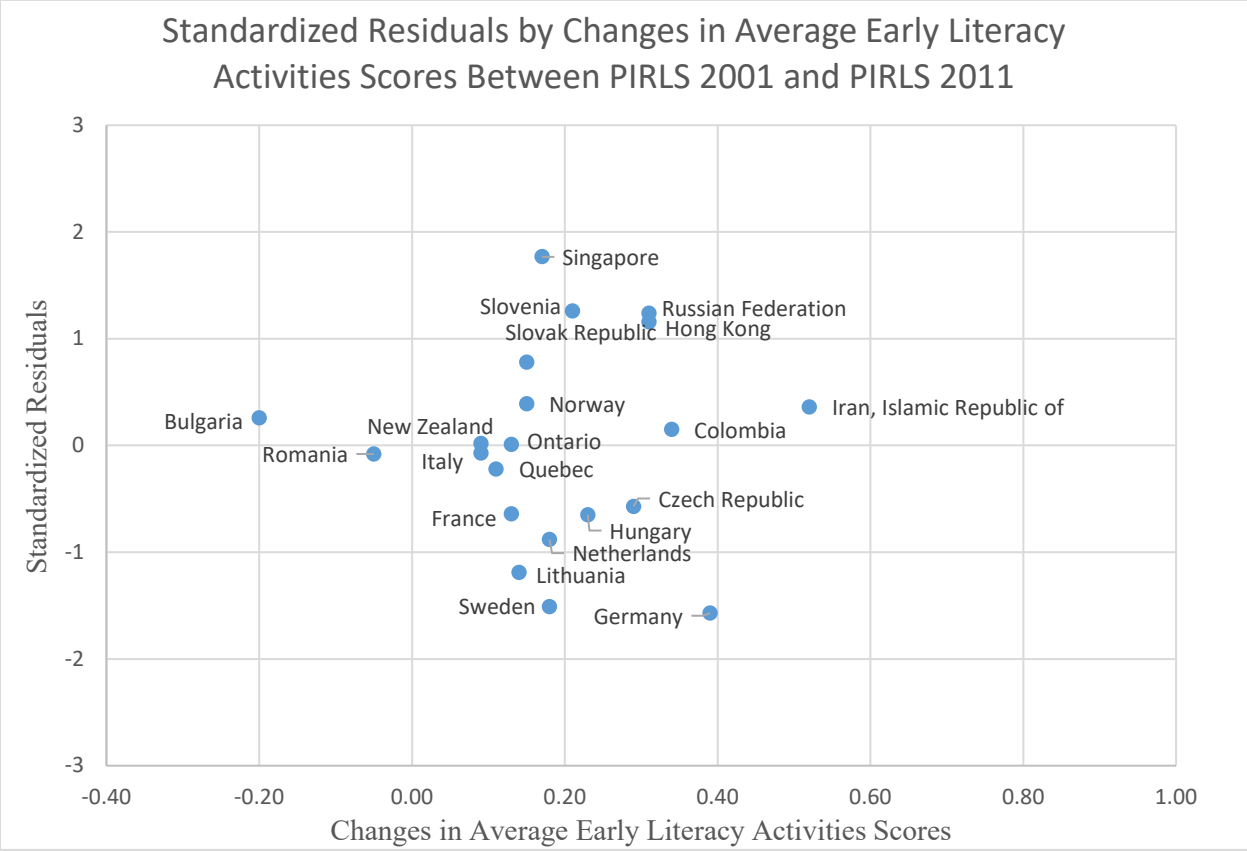


## Phase 3: Analysis of Heteroscedasticity

For Phase 3, further analysis was conducted to check for heteroscedasticity by plotting the standardized residual from Model 9 against the changes in average early literacy activities scores. As can be seen in Figure A.2, the variability seems to be approximately random across the distribution of early literacy activities, suggesting that heteroscedasticity is not a concern.

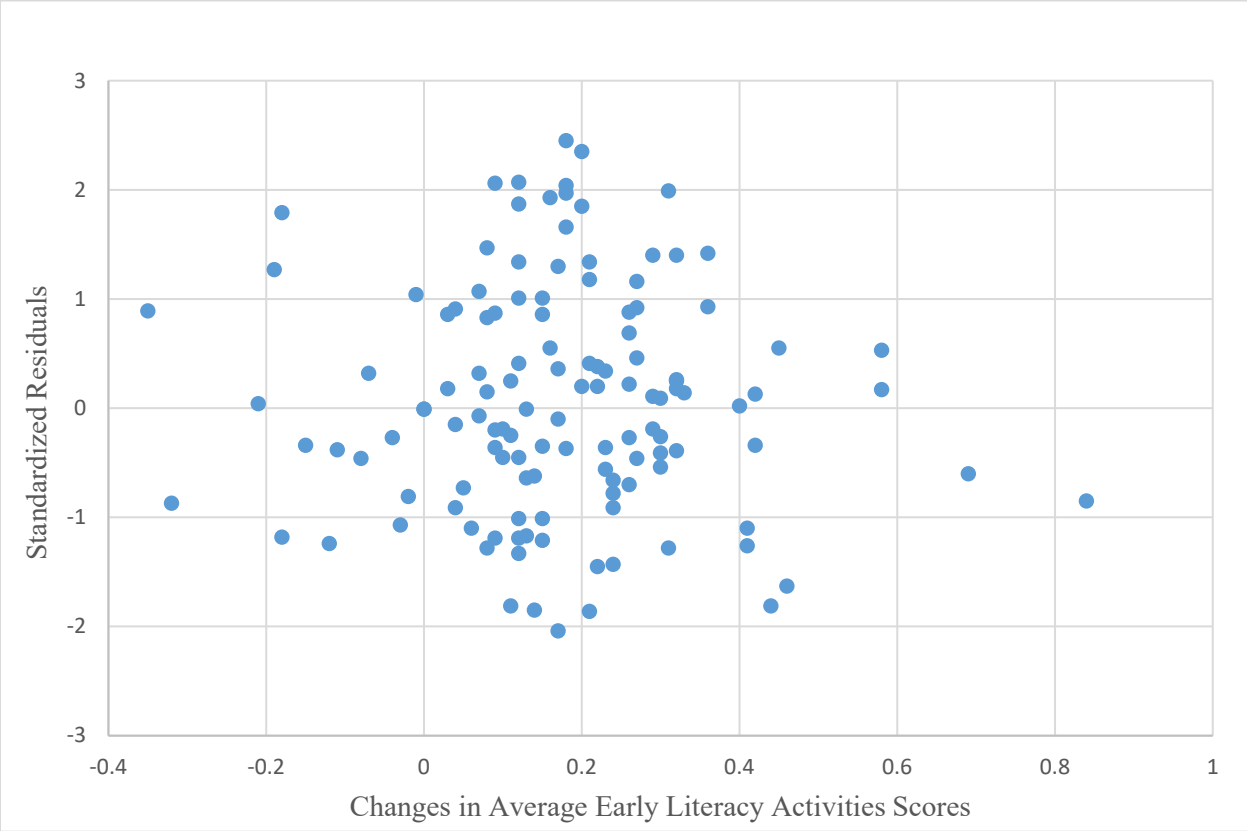
**Figure A.2: Scatterplots for Evaluating Country-Level Heteroscedasticity for Model 9**



## Phase 4: Analysis of Heteroscedasticity

The data were examined for heteroscedasticity by plotting the standardized residuals by the changes in early literacy activities scores, as shown in Figure A.3. The results show a random pattern to the residuals across most of the early literacy activities distribution. Although there appears to be less variance above 0.05 on the early literacy activities scale, this could be linked to the fact there are few countries, and it does not appear to warrant transformation.

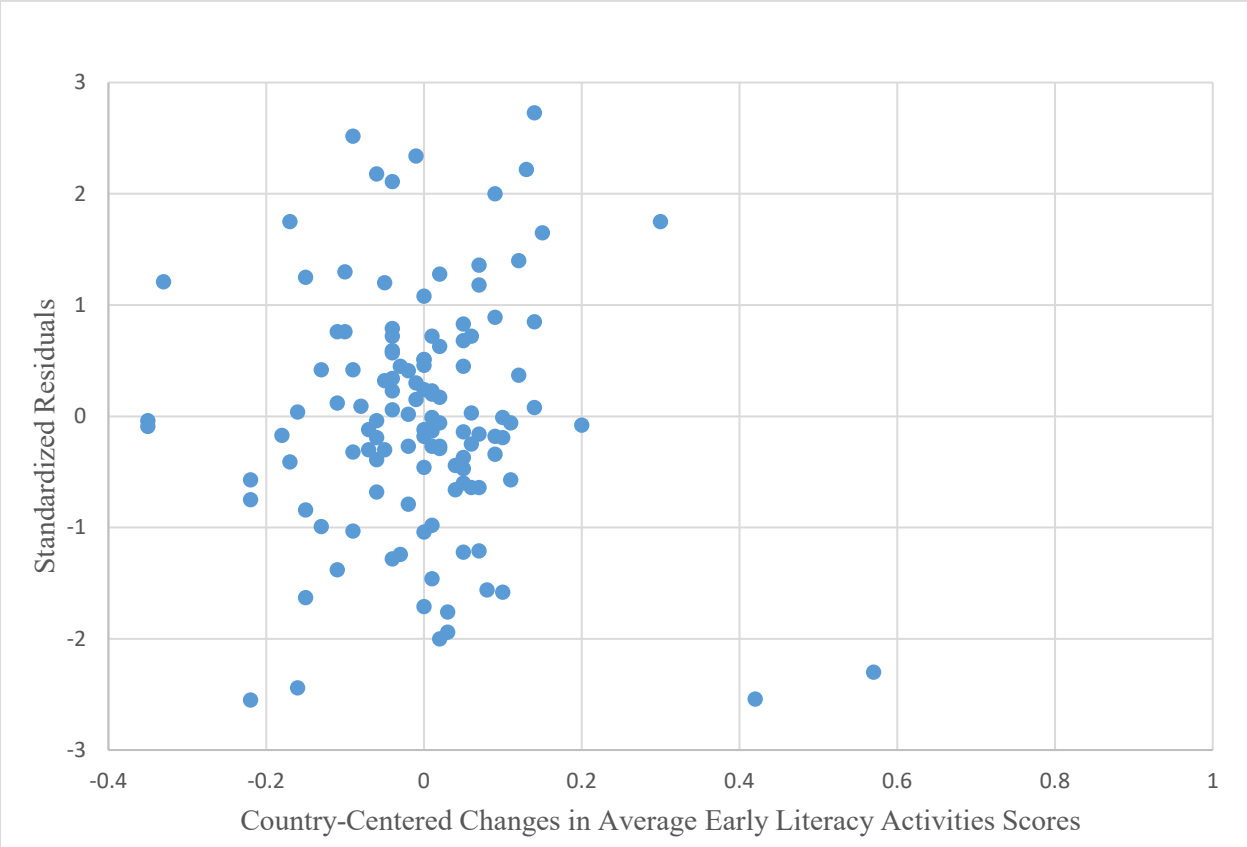
**Figure A.3: Scatterplots for Evaluating Country-Level Heteroscedasticity for Model 11**



## Phase 5: Analysis of Heteroscedasticity

To examine the heteroscedasticisty in the Phase 5 between-country analysis of subpopulation variance, the standardized residuals were plotted against the distribution of early literacy activites. As can be seen in the plot, there seems to be variabililty across the early literacy activities distribution, with a few outliers with scores between 0.4 and 0.6 in the distribution.

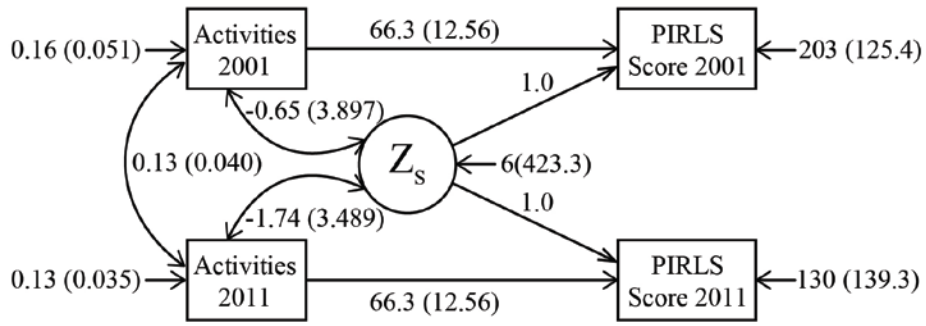
**Figure A.4: Scatterplots for Evaluating Country-Level Heteroscedasticity for Model 12**



## Phase 6: Comparisons of Fixed-Effects Coefficient Estimates across Groups

### Figure A.5: Detailed Results for Gender Subroups, Null Model

Girls Model




Boys Model

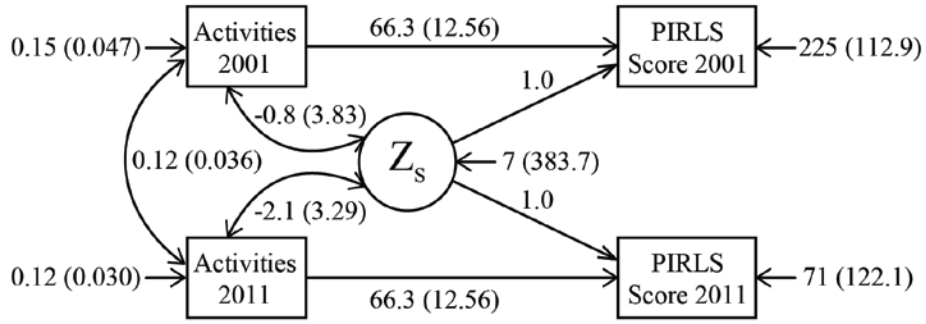
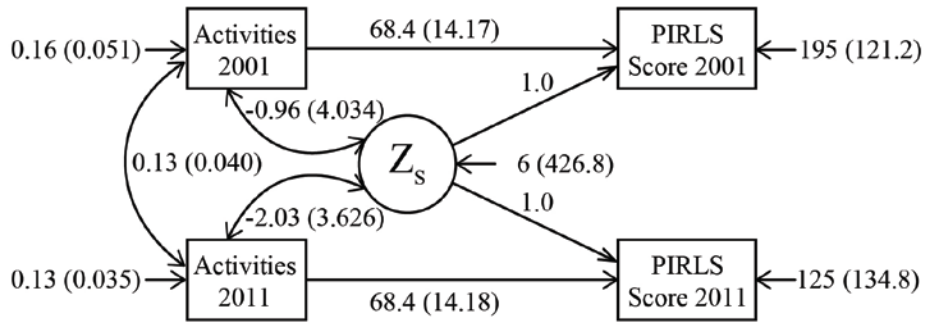


**Figure A.6: Detailed Results for Gender Subgroups, Alternative Model**

Girls Model




Boys Model

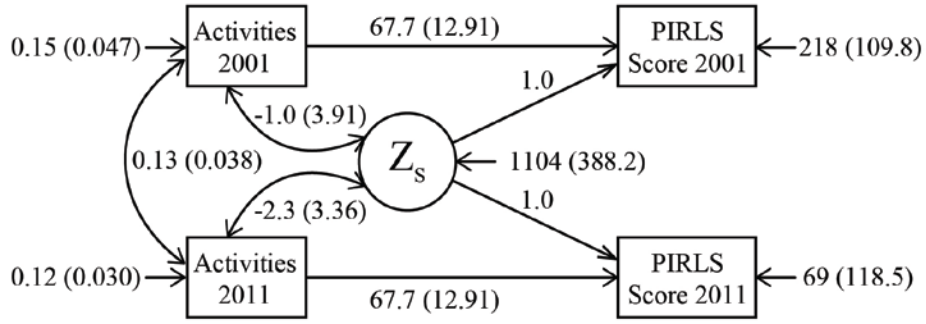


**Figure A.7: Detailed Results for Highest Parental Education Subgroups, Null Model**

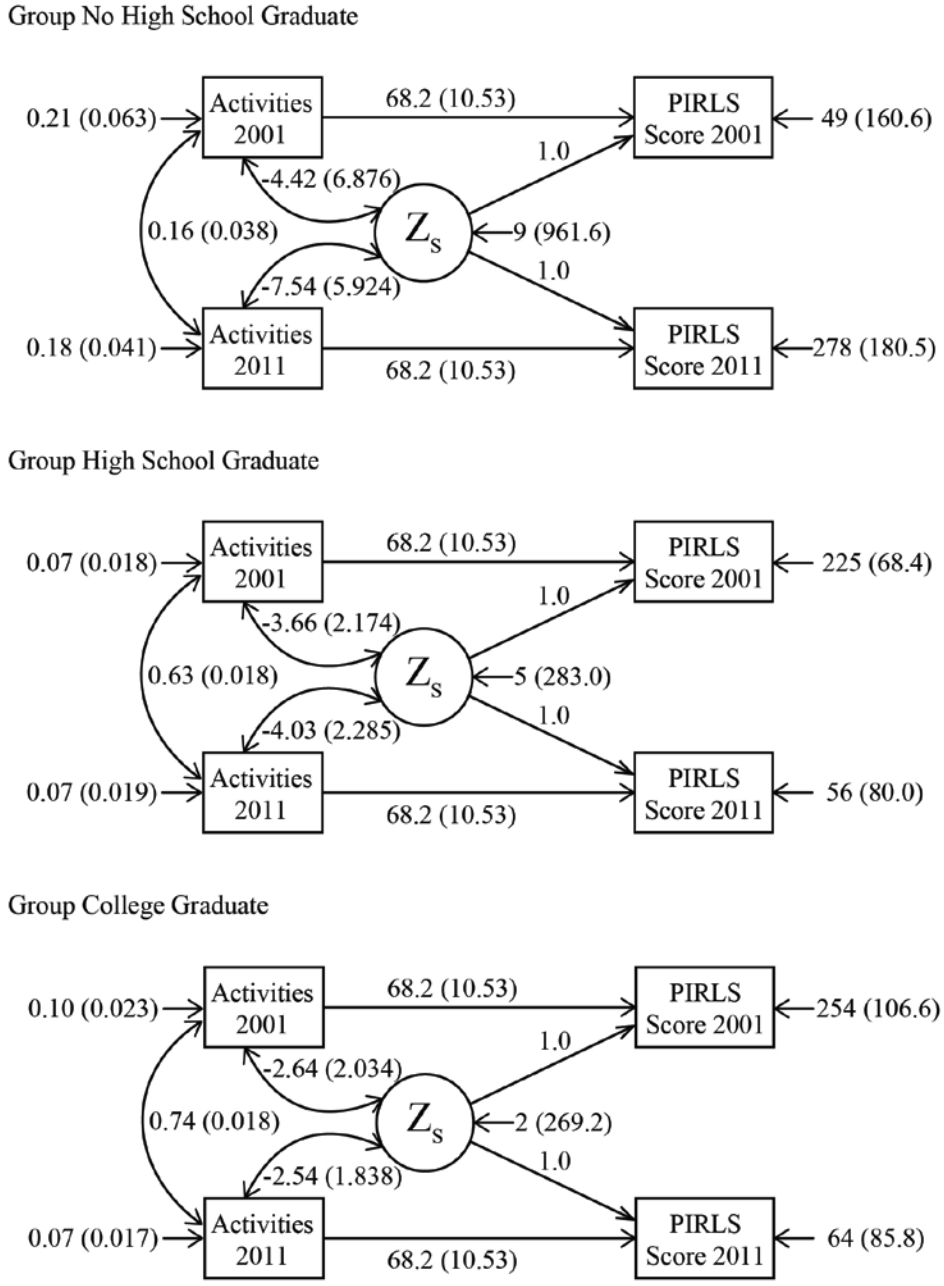


**Figure A.8: Detailed Results for Highest Parental Education Subgroups, Alternative Model**