TWO ESSAYS ON EDUCATIONAL RESEARCH: (1) USING MAXIMUM CLASS SIZE RULES TO EVALUATE THE CAUSAL EFFECTS OF CLASS SIZE ON MATHEMATICS ACHIEVEMENT: EVIDENCE FROM TIMSS 2011; (2) POWER CONSIDERATIONS FOR MODELS OF CHANGE

By

Wei Li

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Educational Policy - Doctor of Philosophy
Measurement and Quantitative Methods - Dual Major

2015

ABSTRACT

TWO ESSAYS ON EDUCATIONAL RESEARCH: (1) USING MAXIMUM CLASS SIZE RULES TO EVALUATE THE CAUSAL EFFECTS OF CLASS SIZE ON MATHEMATICS ACHIEVEMENT: EVIDENCE FROM TIMSS 2011; (2) POWER CONSIDERATIONS FOR MODELS OF CHANGE

By

Wei Li

This dissertation is a collection of two essays that address issues of class size effects on student achievement and power analysis methods for model of changes.

Class size reduction policies have been widely implemented around the world in the past decades. However, findings about the effects of class size on student achievement have been mixed. In addition, most of the studies about class size effects have focused on the effects on the average achievement for all students. Only a few studies have focused on the differential class size effects across the student achievement distribution, and their findings have been mixed. The first essay (Chapter 1 and Chapter 2) was designed to evaluate class size effects on student achievement. In particular, Chapter 1 employed instrumental variables (IV) methods to examine the causal effects of class size on fourth grade mathematics achievement using data from TIMSS (Trends in International Mathematics and Science Study). While I found some evidence of class size effects in Romania and the Slovak Republic, overall there were no systematic patterns of class size effects. The results indicate that in most European and Asian countries class size reduction may not improve mathematics achievement in fourth grade.

The first essay also evaluated the differential class size effects across mathematics achievement distribution. In particular, Chapter 2 employed quantile regression analysis, coupled with instrumental variables methods, to examine the causal effects of class size on

fourth grade mathematics achievement. While I found some evidence of quantile-specific class size effects in Romania and the Slovak Republic, overall there were no systematic patterns of class size effects. What is more, there was no evidence to show that high- or low-achievers benefited more from smaller classes. The results indicate that in most European and Asian countries class size reduction may not increase or reduce the achievement gap between low- and high-achieving students in fourth grade.

The second essay of this dissertation (Chapter 3) was designed to provide methods for three-level models in studies of polynomial change. Experiments that involve nested structures often assign entire groups to treatment conditions and follow them over time to assess group differences in the average of change, rate of acceleration, or higher degree polynomial effect. Chapter 3 provide methods for power analysis in three-level polynomial change models for cluster randomized designs (i.e., treatment at the third level) and block randomized designs (i.e., treatment at the second level). Both unconditional models and conditional models that include covariates at the second (e.g., student) and the third (e.g., school) levels are discussed. The power computations take into account nesting effects at the second and at the third level, the duration of study, sample size effects (e.g., the numbers of schools and students), and covariates effects. Chapter 3 also provided illustrative examples to show how powers are influenced by the study duration, sample sizes and covariates at the second and the third level.

ACKNOWLEDGEMENTS

This dissertation has been a great challenge to me. I have benefited a lot from many people I worked with at different stages of my life. I would like to acknowledge their support, guidance, and help.

First of all, I would like to thank my dissertation committee for their support and guidance. I sincerely thank Dr. Spyros Konstantopoulos for his long term academic guidance and unlimited support in my doctoral studies. He introduced me into class size effects analysis and power analysis. Dr. Konstantopoulos is a very knowledgeable and generous person, who has shared countless great ideas with me and helped me make this dissertation much better than I can imagine. He has been a great mentor during my journey of Ph.D. study. His insightful direction has greatly shaped my thinking and career goals as a quantitative educational researcher. I particularly appreciate his time, patience, kindness, and encouragement that helped me get though the hard times during my doctoral study. I was so fortunate to have Dr. Konstantopoulos as my advisor. I am indebted to him for all of his support, guidance and encouragement. I am looking forward to continuing to work with him and learn from him in the future.

I thank Dr. Barbara Schneider, who provided me great opportunities to work in several large-scale research projects. I worked under her as a graduate assistant, and gained valuable experiences of analyzing the real data and built up my quantitative skills. I also thank her thoughtful advice and feedback as I was working on this dissertation and searching for an academic job. I am thankful to Dr. Amita Chudgar and Dr. Kenneth Frank for agreeing to serve on my dissertation committee. I specially thank Dr. Chudgar for her

support and caring that helped me survive when I first came to MSU. I thank Dr. Frank for his critical feedback and valuable advice that I have learned a lot from.

I would like to acknowledge the generous financial support from Dr. Michael Sedlak and the Educational Policy program that supported my doctoral study and this dissertation. I also express my thanks to Dr. David Arsen and Dr. Kim Maier for their guidance and support.

I thank my teachers in the Graduate School of Education at Peking University. I particularly thank Dr. Ding Yanqing who encouraged me to pursue a doctoral degree in the U.S. and suggested me do a second major in MQM. He gave me countless advice and direction during the past years, which have changed my life. I am also thankful to my master advisor, Dr. Yue Changjun, for his consistent guidance and support when I was studying in GSE and at MSU. I thank Dr. Yan Fengqiao, Dr. Song Yingquan, Dr. Ding Xiaohao, Dr. Yang Po, and Dr. Ma Liping for their encouragement and help.

Moreover, I am thankful to my friends and fellow students from Educational Policy and MQM at MSU. Specifically, I thank my close friend, classmate, and coauthor- Dr. Yisu Zhou. I got countless support and help from him at every stage of my doctoral study. I am very thankful to Dr. Anne Traynor, Dr. Min Sun, Dr. Yongmei Ni, Guan Saw, Liyang Mao and Na Liu for their help and suggestion with respect to my research and my career. I would like to thank my close friend Keyin Wang who took care of my wife and me in the past years at MSU. I am thankful to Siwen Guo, Xin Luo, Ran Xu, and Yi Wei. I wish them all the best in their graduate study and subsequent life. I also would like to thank my close friend Wang Feng and his wife for taking care of my wife when I was in the U.S. while she was in Beijing.

I owe many thanks to my family. I thank my parents, Li Yueqing and Yin Shuwen, and my parents in law, Zhaoqiang and Di Youlan, for their unconditional love during my study in China and in the U.S. Finally, I would like to thank my wife, Meng Zhao, who has been there to laugh with me and to cry with me since we met eleven years ago.

This dissertation is dedicated to my grandparents, Yu Qingzhen and Li Guangxin, who gave me the best childhood.

TABLES OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

This dissertation is a collection of two essays that address issues of class size effects and power analysis method for model of changes in longitudinal randomized control trails.

The first essay (Chapter 1 and Chapter 2) focused on the effects of class size on fourth graders mathematics achievement. Many countries have recently enacted class size reduction policies. Mixed research findings leave policy makers, practitioners, and researchers wondering if class size reduction policy is an effective way of improving student achievement. Chapter 1 addresses the effects of class size on student average achievement. Specifically, Chapter 1 investigated the effects of class size on mathematics achievement for fourth graders using data from the Trends in International Mathematics and Science Study (TIMSS). Typical statistical method such as ordinary least square regression may produce biased estimates of class size effects because student and teacher allocation to classes is likely non-random. For example, students might be assigned to classes based on their prior achievement; however, there was no prior achievement provided in TIMSS. To account for the non-randomness of student assignment and to facilitate causal inference, I created a class size index that is independent of the unobserved process of student assignment, which is usually called Instrumental Variable (IV). In particular, I computed the grade and school specific average class size based on the maximum class size requirement in a country as the instrument. Generally, no systematic pattern of association between class size and mathematics achievement was found in my study. These results indicate that class size reduction may not improve fourth grade mathematics achievement.

Besides improving average student achievement, another critical objective of education interventions is to increase achievement for students at risk, and thus reduce the achievement gap between lower- and higher-achieving students. Class size reduction has been advocated as such an intervention by some researchers; however, no recent study has used current data to evaluate if CSR closes the achievement gap. Chapter 2 attempts to fill in that gap in the literature by exploring the differential class size effects for students with different levels of achievement. I employed quantile regression analysis, coupled with IV, to estimate the causal effects of class size on student achievement in the middle as well as the lower and upper tails of the achievement distribution. I also compared the differences of estimated effects of class size between low- and high-achieving students. Overall, there was were systematic differential class size effects across achievement distribution, and in most countries class size reduction may not reduce the achievement gap between low- and high-achieving fourth grade students.

Chapter 3 addresses power analysis methods in three-level polynomial change models. An important part of the design phase of an experiment involves power analysis. Statistical power is the probability of detecting the treatment effect of interest when it exists. A priori power analyses help educational researchers identify how big a student, classroom, or school sample they need to ensure a good enough chance (e.g., > 80 percent) of detecting a treatment effect assuming it is true. It is common in education to employ designs where students are assigned randomly to a treatment and a control condition, and then they are followed over time. The main objective in these studies is to determine whether treatment effects fade or have lasting benefits over time. Previous work has presented methods for power analysis of two-level (e.g., repeated measures nested in students) models.

Nonetheless, populations in education have frequently more complicated structures. For example, students are also nested within classes or schools and so forth. As a result, it is natural to extend methods for power analysis for tests of treatment effects from two to three-levels. My second essay (Chapter 3) was designed to provide methods for power analysis in three-level models. Both methods for cluster randomized designs (i.e., treatment at the third level) and block randomized designs (i.e., treatment at the second level) were discussed.

CHAPTER 1 CLASS SIZE EFFECTS ON FOURTH GRADE MATHEMATICS ACHIEVEMENT

Introduction

Identifying the best allocation of school resources to improve student achievement has been a fundamental objective in education for a long time. As a result, school resources such as teacher pay, per-pupil funding, and class size have received considerable attention in the past three decades. The underlying assumption is that these school resources can have positive effects on student achievement.

The effects of class size on student achievement have received particular attention in education research and policy. Results from experiments have indicated positive effects of small classes on student achievement (e.g., Finn & Achilles, 1990; Molnar et al., 1999). Specifically, evidence from Project STAR (Student-Teacher Achievement Ratio) in Tennessee has strongly indicated achievement improvements for students in small classes compared to students in regular size classes (e.g., Nye, Hedges, & Konstantopoulos, 2000; Krueger, 1999). However, results from quasi-experiments have indicated much smaller class size effects. For example, Angrist and Lavy (1999) found significant but smaller class size effects in Israel than those reported in Project STAR. Also, Hoxby (2000) analyzed data from a natural experiment in Connecticut and found that class size did not have a significant effect on student achievement.

Findings about class size effects have informed policies in different countries and, as a result, various countries have enacted class size reduction (CSR) polices. Such policies have been quite popular in the U.S. especially during the past decade. Twenty-one states in the U.S. had a CSR policy in place in 2007-2008 (*Education Week*, 2008). In Asia,

countries such as South Korea, Japan, Singapore, and districts such as Hong Kong and Chinese Taipei, have implemented CSR policies aiming to increase student achievement in recent years.

Similarly, in Europe, most countries have adopted CSR policies. In particular, two thirds of the European Union countries had introduced maximum class size requirements until 2011 in an attempt to ensure that class size does not exceed 30 students per class. Some European countries have lowered their upper class size limits in the last few years. For example, in Austria, since the 2007-2008 school year the number of students per classroom has been reduced at primary schools, general secondary schools, academic secondary schools and pre-vocational schools (EACEA Eurydice, 2011). Also, Scotland has reduced lately class size in first grade from a maximum of 30 students to 25 students (EACEA Eurydice, 2011). Other countries however, have stopped setting upper class size limits or have increased their upper class size limits in primary education. Norway for instance has stopped setting upper class size limits since 2003.2004. Also, Italy and Portugal have increased their upper class size limits from 25 and 24 in 2006- 2007 to 26 and 28 in 2010-2011 respectively.

Class size reduction policies require considerable investments in education. But economic budgets allocated to education at the federal and local levels are typically limited. Policy makers, practitioners, and researchers are still wondering whether CSR policies are an effective way of improving student achievement. Chapter 1 attempts to provide additional evidence about the effects of class size on student performance using data from a large-scale international assessment program. Specifically, the purpose of Chapter 1 is to examine the effects of class size on mathematics achievement using data from the 2011

fourth grade sample of TIMSS. My sample included hundreds of schools and thousands of students in 18 Asian and European countries and districts (see Table 1.1). I employed regression methods to analyze the data. To facilitate causal inferences of class size effects I used instrumental variables (IV) that take advantage of the maximum class size rule.

My study contributes to the existing literature in two ways. First, I used the most recent TIMSS data from 2011 that allows us to evaluate recent, concurrent CSR policies and compare class size effects across 18 Asian and European countries and districts. Second, I used maximum class size rules that allowed us to construct instruments to estimate the causal effects of class size on mathematics achievement in fourth grade across countries.

Literature Review

During the past three decades, researchers explored the effects of class size reduction on student achievement through meta-analyses, experimental and quasi-experimental designs, as well as other advanced statistical methods such as IV.

Meta-analytic reviews of early work on small class effects indicated positive relationship between small classes and student achievement, but the magnitude of the effects was small. For example, Glass and Smith (1979) synthesized 77 studies and found that the average effect-size when class sizes were reduced from 25 to 15 was 0.13 standard deviations (SD). Using a subset of the Glass and Smith (1979) sample that employed random assignment or initial controls for student quality, Slavin (1989) found extremely small effects of class size on achievement. He concluded that the class size effects are consistent, but small in kindergarten through third grades, slightly smaller in fourth through eighth grades, and non-existent in ninth through twelfth grades.

Project STAR is viewed as the most impressive and most powerful field experiment about class size effects in education (Mosteller, 1995). There have been numerous analyses of the Tennessee STAR data that have produced high internal validity estimates. Finn and Achilles (1990) were the first to analyze these data and found that students in small classes performed higher than those in regular classes in all subject areas, in every year of the experiment (kindergarten through third grade). Nye, Hedges, and Konstantopoulos (2000) analyzed the validity of Project STAR and suggested that the effects of class size might be under-estimated because of imperfect implementation. They also found that the estimated class size effects were consistent with those from Glass and Smith (1979). Other studies by Krueger (1999) and Konstantopoulos (2008) produced similar findings about the positive effects of small classes on student achievement in early grades.

Studies that have used observational data, especially data from large-scale surveys, have usually produced results with high external validity (i.e., generality). However, the internal validity (or unbiasedness) of estimates in observational or quasi-experimental studies is not so easy to achieve. That is, researchers have to use advanced statistical methods to warrant the internal validity of estimates. For instance, traditional ordinary least square (OLS) regression may produce biased estimates because of omitted variables bias (i.e., predictors may not be orthogonal to the error term).

Previous work has utilized different analytic methods to examine class size effects on student achievement. For example, Pong and Pallas (2001) used multilevel models to analyze TIMSS data from 1995 in nine different countries and found no class size effects on eighth grade achievement except in the U.S. Other researchers have used IV methods to analyze observational data in an attempt to explore the causal effects of class size

reduction. For example, Akerhielm (1995) used two instruments for class size-the average class size for a given subject in the student's school and the eighth grade enrollment in the school-to analyze class size effects on eighth graders' mathematics, science, English, and history scores using data from 1988 NELS. Her results indicated a significant and negative relationship between class size and student achievement. Hoxby (2000) used data from a natural experiment and used IV methods to estimate the effects of class size on student achievement in Connecticut. Her method exploited random variation in class size due to random variation in births from year to year in schools and district catchment areas. She found no class size effects in fourth and sixth grades. Cho, Glewwe, and Whitler (2012) applied Hoxby's (2000) method to compute class size effects in Minnesota and found positive effects of smaller classes on student achievement, but these estimates were smaller than estimates from Project STAR. Moreover, Wossmann and West (2006) examined class size effects in 11 countries that participated in TIMSS 1995. Their results indicated that there was no clear pattern of whether or when class size affects student achievement.

One of the best IV used to capture class size effects was introduced by Angrist and Lavy (1999). Specifically, their study used the Maimonides rule that sets the maximum class size to 40 students per classroom in order to evaluate the effect of class size on student achievement in Israel. The maximum class size rule of 40 was used to construct IV estimates of class size on test scores. The study reported a statistically significant effect of small classes on fifth grade reading and mathematics scores. In fourth grade the benefit of being in small classes was significant in reading, but not in mathematics. However, in third grade no significant effects of class size on achievement scores were detected.

Several other researchers have also used maximum class size rules as IV to evaluate class size effects. For instance, Bonesronning (2003) investigated class size effects using a maximum class size rule of 30 students per classroom in Norway. His analysis indicated small class effects. Wossmann (2005) explored class size effects in Europe using data from TIMSS 1995 for eighth grade students. He found two statistically significant relationships between class size and student achievement: a marginally significant effect in Norway and a highly significant effect in Iceland. He also found a statistically significant but positive relationship between class size and student achievement in Switzerland. For Denmark, France, Germany, Greece, Ireland, Spain, and Sweden, the estimates were not significant. A recent study about class size effects on fourth grade reading achievement in Greece also reported statistically insignificant estimates (Konstantopoulos & Traynor, 2014). Urquiola (2006) studied 10,018 third-grade students in Bolivia and found significant class size effects with effect sizes as large as 0.30 SDs, bigger than effects found in Project STAR in the U.S. and in Israel.

Methods

Data

I used data from TIMSS latest survey in 2011. TIMSS is the largest international database that measures trends in mathematics and science achievement at fourth and eighth grades. First conducted in 1995, TIMSS provides reliable and timely data about mathematics and science achievement every four years. It also collects extensive information about students, teachers, school principals, and curriculum experts via background questionnaires.

A stratified two-stage cluster-sampling design was used in TIMSS, where schools are sampled at the first stage and one or more intact classes are sampled at the second stage in each of the sampled schools (Martin & Mullis, 2012). TIMSS has produced high-quality assessment measures. Also, teachers reported class size information on all intact classrooms that were sampled. Other useful information about students, teachers, and schools has also been collected. It is noteworthy, that TIMSS was designed to yield a national probability sample of fourth (or eighth) graders. With the use of appropriate weights, one can make projections to the population of fourth (or eighth) graders in each country, which points to the high external validity of the estimates.

The stratified two-stage cluster-sampling design used in TIMSS makes the computation of the standard errors of estimates complicated because student data within the same school are correlated rather than independent. Following the suggestion from TIMSS technical report (Martin & Mullis, 2012), I used the jackknife repeated replication technique (JRR) to estimate the sampling variance because it is computationally straightforward and provides approximately unbiased estimates of the sampling errors (see Martin & Mullis, 2012). That is, JRR standard errors take into account clustering effects induced by the multi-stage sampling scheme.

Country selection

I used fourth grade data from the fifth and latest administration of the TIMSS assessment in 2011. I focused on fourth grade mathematics achievement because class size effects are typically expected in elementary grades (Nye et al., 2000). Twenty five European countries were surveyed in TIMSS 2011. I selected 14 countries of those 25 participating countries that had known clear rules about maximum class size limits for

fourth graders in 2011 (see Table 1.1). The highest upper class size limit in the fourth grade was in Malta and the Czech Republic with a maximum number of 30 students per classroom. The lowest upper class size limit of 24 students per classroom was in Lithuania. The most common upper class size limit was also 28 students per classroom (EACEA Eurydice, 2012). I also selected four Asian countries and districts that set clear maximum class size limit in the fourth grade in 2011. Compared the rules in Europe, the upper class size limits were quite larger in Asia, which ranged from 30 (Hong Kong) to 40 (Japan and Singapore). Table 1 provides detail about the selected countries as well as their upper class size limits.

Measures

The dependent variable was mathematics achievement represented by five plausible values. Because the item pool of TIMSS 2011 was too large for students to finish in two hours, TIMSS used an incomplete booklet design that had each student complete only a proportion of the item pool (Martin & Mullis, 2012). Then, multiple imputation methods were used to construct a distribution of scores that the students might have obtained had they completed the full test. The plausible values are a sample of scores from this distribution that incorporates the uncertainty about student scores (Martin & Mullis, 2012). It has been shown that five plausible values can produce reliable and consistent estimates of student achievement (Schafer, 1999).

The main independent variable was class size and was reported by teachers. Specifically, the class size measure was the number of students in a sampled classroom provided by the teachers. Student, teacher, classroom, and school variables of interest were also used as covariates. The student covariates included gender (e.g., a dummy for female),

Table 1.1: Maximum Class Size Rules: TIMSS 2011

Country	Maximum Class Size Rule	Country	Maximum Class Size Rule
Austria (AUT)	25	Lithuania (LTU)	24
Croatia (HRV)	28	Malta (MLT)	30
Czech Republic (CZE)	30	Portugal (PRT)	28
Denmark (DNK)	28	Romania (ROM)	25
Germany (DEU)	29	Slovak Republic (SVK)	25
Hungary (HUN)	26	Slovenia (SVN)	28
Italy (ITA)	26	Spain (ESP)	25
Hong Kong (HKG)	30	Singapore (SPG)	40
Japan (JPN)	40	Chinese Taipei (TWN)	32

The teacher covariates included education (e.g., dummy for completing post-secondary education), years of teaching experience, gender (e.g., a dummy for female), and teacher's instruction time per week. Classroom covariates included class level SES represented by aggregate measures of number of books in the home and average number of items in the home. The proportion of female students in the classroom and the average student positive affect to mathematics were also used as classroom covariates. School covariates included percent of economically disadvantaged students, percent of students having the tested language as their native language, income level of the school immediate area, and fourth grade enrollment and its square. Missing data flags (i.e., dummies) were included in the models to account for missing data effects. The Appendix A provides the full list of variables as well as detailed description about coding.

Multiple Regression

To examine the class size effects on student mathematics achievement, I employed first a multiple regression model that included class size and student, teacher/classroom, and school covariates namely

$$Score_i = \beta_0 + \beta_1 ClassSize_i + \mathbf{ST_i B_2} + \mathbf{CL_i B_3} + \mathbf{SC_i B_4} + \varepsilon_i \qquad (1.1)$$

where $Score_i$ represents mathematics scores, β_0 is the constant term, $ClassSize_i$ is the main independent variable, β_1 represents the class size effect and is the regression coefficient of interest, $\mathbf{ST_i}$ is a row vector of student background characteristics, $\mathbf{B}_2$ is a column vector of regression coefficients of student characteristics, $\mathbf{CL_i}$ a is row vector of

classroom or teacher characteristics, $\mathbf{B}_3$ is a column vector of regression coefficients of teacher and classroom characteristics, $\mathbf{SC_i}$ is a row vector of school characteristics, $\mathbf{B}_4$ is a column vector of regression coefficients of school characteristics, and $\mathcal{E}_i$ is the error term. Because TIMSS used a complicated cluster sampling design (i.e., sampled schools at the first stage and then classes within schools), the clustering effect needs to be incorporated in the estimation of the standard errors. To achieve this we used JRR techniques to obtain a cluster robust standard error as suggested by Martin and Mullis (2012).

Instrumental Variables

Typical regression could provide consistent estimates of class size under the assumption that class size is not correlated with unobserved processes that may take place in schools. These unobservables are represented by the error term in model (1.1). However, such an assumption is strong and rarely met in observational studies. The assignment of students and teachers to classrooms is not random typically, and thus class size could be correlated with unobserved factors related to student, parent, and teacher characteristics. For example, students may be assigned to classes based on their prior achievement or motivation. Parents may also influence assignment to classes. For instance, parents may want their children to be assigned in the classroom with the highest quality teacher or a specific peer composition (e.g., their children's friends). Teachers may also influence assignment by either selecting high achieving students in their classrooms by teaching the class with the higher proportion of high achieving students. If such processes were to take place, the estimated class size effect from equation (1.1) would be biased.

Because students and teachers are rarely randomly assigned to classrooms in a grade class size might be correlated with unobserved characteristics of students or teachers. For example, in order to help low achieving students, some schools might assign higher quality teachers to classes with higher proportions of low achievers. Variables that determine assignment of students and teachers to classes are not typically measured. For example, student motivation, family income, parental pressure, teacher quality, etc. are rarely available in observational datasets. In addition, cross-sectional data rarely provide indexes of prior ability or performance. Although we included as many covariates as we could in our multiple regression analysis, it is still possible that unobservable factors that are part of the error term in equation (1.1) are correlated with class size. If that were true, then the estimated class size effect in equation (1.1) would be biased.

One way to overcome this potential shortcoming and facilitate causal inferences, is to compute an index of class size that is independent of unobserved student, teacher or school variables. Specifically, we used the maximum class size rule in each country to compute school and grade specific average class size. This new variable was then used as an instrument to exclude unobserved variables from the teacher reported class size. In other words, this method creates a new class size variable that is "error free" and should not be related to unobserved variables. Our method is similar to that used by Angrist and Lavy (1999). The first step in this approach is to compute the average class size in fourth grade in each school. Specifically, the average class size in fourth grade in each school based on the maximum class size requirement is calculated as

$$f_i = E_i / [\text{int}((E_i - 1) / rule) + 1] \qquad (1.2)$$

where E_i denotes the enrollment in grade four in a school; f_i denotes the computed average class size based on the maximum class size rule; *rule* denotes the upper class size limit in a given country; for any positive number n, the function int(n) is the immediate smaller integer less than n. For example, if grade enrollment $E = 70$ and the maximum class size rule is 30 then int(n) = int(2.33) = 2. The upper class size limit generates discontinuities of the computed class size as the enrollment count increases to multiples of the upper class size limit. For example, if the maximum class size rule is 30 in a specific country, the above equation captures the fact that the maximum class size rule allows enrollment of cohorts of 1-30 to be grouped in a single class, while enrollment of cohorts 31-60 are split into two classes with average class sizes 15.5-30, and so on.

The second step was to regress the teacher reported class size on the instrument (i.e., the school and grade specific average class size we computed in equation 1.2), as well as other covariates (see variables section). This step is designed to eliminate the unobservables (i.e., the error) from teacher reported class size.

Specifically, the regression equation is

$$ClassSize_i = \pi_0 + \pi_1 f_i + \mathbf{ST_i \Pi_2} + \mathbf{CL_i \Pi_3} + \mathbf{SC_i \Pi_4} + u_i \qquad (1.3)$$

where f_i is the computed average class size (i.e., the instrument) in a school based on the maximum class size rule and u_i is the error term. All other terms have been defined previously. The π's are the regression estimates that need to be computed. The fitted

values of this regression are computed and will be used in the third step as the new class size variable that is "free" of error.

The third and final step of this procedure used a regression where the fitted values (denoted below of FV_i) from the regression equation (1.3) represent class size and are the main independent variable in the following achievement regression

$$Y_i = \delta_0 + \delta_1 FV_i + \mathbf{ST}_i\Delta_2 + \mathbf{CL}_i\Delta_3 + \mathbf{SC}_i\Delta_4 + \xi_i \qquad (1.4)$$

where Y indicates mathematics scores, ξ_i is an error term and all other terms have been defined previously. The coefficient δ_1 represents the relationship between mathematics achievement and class size, adjusted for student, teacher/classroom, and school characteristics. Appropriate student weights were used in both regressions (equations 1.3 and 1.4). The δs indicate regression estimates that need to be computed. The student, classroom/teacher, and school covariates included in equation (1.4) are the same as those included in equation (1.3).

The method (i.e., instrumental variables) described above has been used in previous work to estimate causal class size effects (e.g., Angrist & Lavy, 1999; Krueger, 1999). We used JRR techniques to estimate the standard errors of the regression coefficients. The TIMSS sampling design makes the JRR techniques particularly well suited for estimating the standard errors in complex sampling surveys such as TIMSS (Martin & Mullis, 2012). Our analysis was conducted for each plausible value separately, and then the five sets of estimates were combined to construct one set of final estimates of class size effects. To combine estimates we used formulae provided by Shafer (1999). The standard error of the

class size effects was a combination of the sampling variance obtained through JRR techniques and the variance between plausible values (see Martin & Mullis, 2012). The standard errors of the regression coefficients were also corrected for the two-stage estimation (i.e., equations 1.3 and 1.4) before they were combined across plausible values.

There were two key conditions that the computed average class size f_i must meet in order for the instrument to be valid: (1) schools should follow the maximum class size rule very well. In other words, f_i should be correlated significantly with reported class size; and (2) the instrument cannot be correlated with any of the unobserved student, teacher, or school characteristics (i.e., f_i should not be correlated with the error term in equation 1.1). The first condition can be checked through the first stage regression (equation 1.3). If the coefficient of the computed average class size (the instrument) is significantly different from zero, then the assumption that reported class size and the instrument are related holds. If the instrument is only marginally significant, our instrument may be weak. When instruments are weak, then the standard IV estimates, hypothesis tests, and confidence intervals may be unreliable (Stock, Wright, & Yogo, 2002). When multiple instruments are used the rule of thumb is that the F-statistic of all instruments in the first-stage regression should be larger than 10 (Staiger & Stock, 1997). In our study only one instrument is used (i.e., average class size per school) and thus we employ a t-test. The t-statistic of the regression coefficient of the instrument (π_1 in equation 1.3) should be greater than 3.20 and significant in the first stage regression. The t-statistic denotes the statistic for testing the hypothesis of a zero coefficient for the instrument (computed average class size using maximum class size rule) in the first stage regression.

The second condition which is called "exogenous assumption" or "exclusion restriction" indicates that computed average class size influences student mathematics achievement only through reported class size controlling for grade enrollment and other covariates. The question is essentially whether the instrument might be correlated with unmeasured factors that influence student assignment to classes. For example, private schools could manipulate the maximum class size requirement through adjusting their tuition or enrollment to avoid creating additional classrooms (see Urquiola & Verhoogen, 2009). Unfortunately, I cannot identify public or private schools based TIMSS data. Parents could manipulate the class size rule as well if school choice is an option in their country. In other words, some parents might take advantage of the rules and make their kids study in schools with smaller classes. There was some evidence that showed associations between smaller class size and higher student SES level in Spain and Malta based on some regression analysis, which indicates parents with higher SES might manipulates the rules and raises some concern of the validity of the IV in these two countries.

Results

Descriptive statistics

Table 1,2 presents descriptive statistics for selected student, teacher, and school variables of interest as well as samples sizes for students, classes, and schools. The national average mathematics scores for all countries participating in TIMSS have been set to a mean of 500 and a SD of 100. Fourteen countries in Table 1.2 had average scores greater than 500. Asian countries' score were much higher than European countries. Singapore had the highest average score (605.79), closely followed by Hong Kong, Chinese Taipei,

and Japan. Denmark had the highest scores among European countries, closely followed by Lithuania, Portugal, and Germany. With an average score of 482.28, Romania had the lowest average score. Spain, Croatia, and Malta also had average scores lower than 500. About half of the students were females for all countries. At least 70 percent of students almost always spoke the tested language at home for all countries except Chinese Taipei, Malta, Singapore, and Spain.

The average class sizes in grade four for European countries were much smaller than those in Asian countries. In Europe, the smallest average class size with 19 students per class was in Austria, closely followed by Slovenia, the Slovak Republic, Lithuania and Romania. With nearly 23 students per classroom on average, Spain had the largest classes. The largest average grade four enrollment (70.13) was in Italy; while the smallest average grade four enrollment (25.6) was in the Slovak Republic. In Asia, the largest average class size were found in Singapore (37). Teacher experience varied across countries. The highest average teacher experience was in Lithuania (24 years), while the lowest was in Singapore (nearly ten years). Almost all teachers completed post-secondary education in all countries in our sample except Italy and Romania. More than 75 percent of teachers were females in all European countries in our sample except Denmark, where only about half of the teachers were females; while among Asian countries, it ranged from 56 percent to 82 percent. In addition, school size was much larger in Asia than in Europe.

The numbers of students and schools per country sample are also presented in Table 1.2. The number of schools ranged from 96 in Malta to 216 in Denmark; the number of classes ranged from 197 in Malta to 351 in Singapore; the smallest sample of students was in Malta (3607), while the largest sample of students was in Singapore (6368).

Table 1.2: Descriptive Statistics for Some Variables of Interest of TIMSS 2011 Samples: Means and Standard Deviations

	AUT	CZE	DEU	DNK	ESP	HRV	HUN	ITA	LTU	MLT	PRT	ROM	SVK	SVN	HKG	JPN	SGP	TWN
Student Variables																		
Mathematics Achievement	508.31	510.85	527.74	536.96	482.43	490.17	515.40	507.82	533.69	495.77	532.26	482.28	506.77	513.03	601.61	585.37	605.79	591.21
	(62.70)	(70.39)	(62.14)	(70.77)	(70.31)	(67.07)	(89.79)	(72.17)	(74.02)	(77.71)	(68.68)	(105.36)	(79.63)	(68.52)	(66.42)	(72.31)	(78.18)	(73.22)
Female	0.49	0.48	0.49	0.51	0.49	0.50	0.49	0.50	0.48	0.49	0.49	0.48	0.49	0.48	0.46	0.49	0.49	0.47
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Age in Years	10.33	10.42	10.37	11.02	9.97	10.67	10.77	9.81	10.85	9.92	10.01	10.77	10.38	9.92	10.07	10.62	10.90	10.24
	(0.44)	(0.44)	(0.49)	(0.38)	(0.44)	(0.32)	(0.51)	(0.36)	(0.37)	(0.42)	(0.49)	(0.65)	(0.64)	(0.34)	(0.51)	(0.28)	(0.46)	(0.31)
Almost Always Speaking Tested Language at Home	0.76	0.86	0.73	0.78	0.67	0.85	0.97	0.78	0.82	0.16	0.89	0.88	0.79	NA	0.65	0.85	0.32	0.50
	(0.43)	(0.35)	(0.45)	(0.41)	(0.47)	(0.36)	(0.18)	(0.42)	(0.39)	(0.37)	(0.31)	(0.33)	(0.41)		(0.48)	(0.36)	(0.47)	(0.50)
SES: Numbers of Books in the Home	2.94	3.17	3.17	2.96	2.95	2.55	3.01	2.74	2.57	2.90	2.73	2.29	2.89	2.98	2.82	2.75	3.08	2.90
	(1.13)	(1.09)	(1.10)	(1.08)	(1.16)	(1.07)	(1.25)	(1.15)	(1.04)	(1.05)	(1.07)	(1.15)	(1.13)	(1.04)	(1.15)	(1.07)	(1.08)	(1.26)
SES: Numbers of Items in the Home	6.30	8.47	8.09	8.89	5.31	7.48	8.39	6.47	8.61	8.94	6.92	6.28	7.34	8.04	6.28	7.95	7.76	5.85
	(1.21)	(1.69)	(1.65)	(1.29)	(0.92)	(1.35)	(1.94)	(1.82)	(1.79)	(1.76)	(1.55)	(2.55)	(1.80)	(1.50)	(2.03)	(1.79)	(1.93)	(1.71)
Classroom Variables																		
Class Size	19.33	21.13	21.61	21.25	22.63	20.61	22.09	20.10	20.00	21.40	20.91	20.00	19.67	19.67	32.13	28.90	37.00	28.03
	(4.12)	(5.30)	(3.89)	(3.98)	(4.17)	(5.73)	(5.45)	(4.59)	(5.11)	(4.85)	(4.78)	(5.86)	(4.84)	(4.13)	(5.38)	(8.54)	(5.57)	(4.60)
Classroom SES: Average Numbers of Books in the Home	2.94	3.17	3.17	2.96	2.95	2.55	3.01	2.74	2.57	2.90	2.73	2.29	2.89	2.98	2.82	2.75	3.08	2.90
	(0.48)	(0.45)	(0.46)	(0.42)	(0.57)	(0.48)	(0.65)	(0.45)	(0.48)	(0.37)	(0.53)	(0.69)	(0.59)	(0.39)	(0.50)	(0.37)	(0.47)	(0.45)
Classroom SES: Average Numbers of Items in the Home	6.30	8.47	8.09	8.89	5.31	7.48	8.39	6.47	8.61	8.94	6.92	6.28	7.34	8.04	6.28	7.95	7.76	5.85
	(0.40)	(0.57)	(0.66)	(0.48)	(0.31)	(0.53)	(0.98)	(0.64)	(0.85)	(0.62)	(0.74)	(1.72)	(1.07)	(0.52)	(1.02)	(0.52)	(0.77)	(0.55)
Percent of Female Students	0.49	0.48	0.49	0.51	0.49	0.50	0.49	0.50	0.48	0.49	0.49	0.48	0.49	0.48	0.46	0.49	0.49	0.47
	(0.14)	(0.13)	(0.11)	(0.11)	(0.10)	(0.12)	(0.13)	(0.11)	(0.13)	(0.26)	(0.14)	(0.12)	(0.12)	(0.11)	(0.17)	(0.07)	(0.21)	(0.07)
Teacher Variables																		
Experience in Years	21.54	18.76	19.29	15.73	21.02	20.75	23.96	23.98	24.01	12.70	17.29	23.22	19.94	20.69	14.49	17.33	9.81	14.56
	(11.58)	(10.28)	(12.27)	(10.74)	(11.00)	(9.79)	(9.92)	(10.02)	(8.49)	(8.29)	(8.62)	(11.11)	(10.02)	(9.67)	(8.24)	(11.69)	(9.11)	(7.15)
Complete Post-Secondary Education	0.99	0.92	0.87	0.86	0.94	0.98	0.97	0.21	0.97	0.86	0.97	0.57	0.99	0.99	0.96	0.92	0.87	0.98
	(0.11)	(0.26)	(0.34)	(0.35)	(0.23)	(0.13)	(0.17)	(0.41)	(0.16)	(0.34)	(0.17)	(0.50)	(0.10)	(0.11)	(0.19)	(0.27)	(0.34)	(0.14)
Female	0.91	0.95	0.78	0.53	0.77	0.96	0.94	0.91	0.99	0.81	0.86	0.87	0.92	0.96	0.56	0.59	0.71	0.82
	(0.28)	(0.23)	(0.41)	(0.50)	(0.42)	(0.19)	(0.24)	(0.29)	(0.10)	(0.39)	(0.34)	(0.34)	(0.27)	(0.19)	(0.50)	(0.49)	(0.45)	(0.38)
Instruction Time in Hours	3.98	4.13	4.08	3.11	4.57	3.74	4.04	5.82	4.06	5.43	7.17	4.02	3.73	4.44	4.16	3.72	5.49	3.14
	(0.90)	(0.90)	(0.97)	(0.41)	(0.69)	(0.83)	(1.21)	(1.37)	(0.95)	(1.32)	(1.08)	(1.03)	(0.12)	(0.78)	(0.76)	(0.27)	(1.68)	(0.91)

Table 1.2 (cont'd)

	AUT	CZE	DEU	DNK	ESP	HRV	HUN	ITA	LTU	MLT	PRT	ROM	SVK	SVN	HKG	JPN	SGP	TWN
School Variables																		
Grade 4 Enrollment	27.15	26.50	42.49	35.86	36.22	52.82	33.43	70.13	35.18	41.48	26.45	29.82	25.60	30.54	105.78	58.61	273.05	108.96
	(22.37)	(21.49)	(25.37)	(21.51)	(23.13)	(33.47)	(25.20)	(46.61)	(55.25)	(26.80)	(25.58)	(32.26)	(26.82)	(22.55)	(48.65)	(41.50)	(88.83)	(120.63)
Income Level of the School's Immediate Area: Low	0.25	0.48	0.21	0.22	0.29	0.34	0.59	0.18	0.68	0.14	0.39	0.65	0.56	0.42	0.06	0.05	0.07	0.06
	(0.43)	(0.50)	(0.41)	(0.41)	(0.46)	(0.47)	(0.49)	(0.39)	(0.47)	(0.35)	(0.49)	(0.48)	(0.50)	(0.49)	(0.23)	(0.21)	(0.25)	(0.24)
Income Level of the School's Immediate Area: Medium	0.71	0.50	0.71	0.64	0.66	0.64	0.40	0.71	0.32	0.85	0.60	0.33	0.43	0.56	0.40	0.80	0.74	0.60
	(0.45)	(0.50)	(0.45)	(0.48)	(0.47)	(0.48)	(0.49)	(0.46)	(0.47)	(0.36)	(0.49)	(0.47)	(0.50)	(0.50)	(0.49)	(0.40)	(0.44)	(0.49)
Income Level of the School's Immediate Area: High	0.04	0.02	0.07	0.14	0.04	0.02	0.01	0.11	0.00	0.01	0.01	0.02	0.01	0.02	0.54	0.15	0.19	0.34
	(0.19)	(0.14)	(0.26)	(0.35)	(0.20)	(0.14)	(0.08)	(0.31)	(0.00)	(0.10)	(0.11)	(0.14)	(0.10)	(0.14)	(0.50)	(0.36)	(0.39)	(0.48)
City Size: 0-3,000	0.56	0.58	0.27	0.38	0.27	0.36	0.45	0.11	0.58	0.29	0.43	0.46	0.58	0.53	0.37	0.21	1.00	0.05
	(0.50)	(0.50)	(0.44)	(0.49)	(0.44)	(0.48)	(0.50)	(0.32)	(0.50)	(0.46)	(0.50)	(0.50)	(0.49)	(0.50)	(0.49)	(0.41)	(0.00)	(0.22)
City Size: 3,001-15,000	0.26	0.17	0.27	0.22	0.15	0.37	0.21	0.40	0.13	0.60	0.27	0.39	0.15	0.22	0.50	0.30	0.00	0.24
	(0.44)	(0.38)	(0.44)	(0.41)	(0.35)	(0.48)	(0.41)	(0.49)	(0.33)	(0.49)	(0.44)	(0.49)	(0.36)	(0.42)	(0.50)	(0.46)	(0.00)	(0.43)
City Size: 15,001-50,000	0.03	0.10	0.22	0.19	0.16	0.08	0.13	0.18	0.09	0.10	0.12	0.06	0.13	0.09	0.07	0.13	0.00	0.26
	(0.18)	(0.31)	(0.42)	(0.39)	(0.37)	(0.28)	(0.33)	(0.39)	(0.28)	(0.30)	(0.32)	(0.24)	(0.34)	(0.29)	(0.25)	(0.34)	(0.00)	(0.44)
City Size: 50,001-100,000	0.02	0.06	0.06	0.09	0.13	0.06	0.05	0.12	0.02	0.00	0.03	0.03	0.07	0.03	0.05	0.30	0.00	0.27
	(0.13)	(0.24)	(0.23)	(0.29)	(0.33)	(0.23)	(0.23)	(0.33)	(0.15)	(0.00)	(0.18)	(0.17)	(0.25)	(0.18)	(0.21)	(0.46)	(0.00)	(0.44)
City Size: 100,001-500,000	0.05	0.06	0.10	0.05	0.16	0.05	0.09	0.07	0.13	0.00	0.07	0.05	0.04	0.08	0.01	0.06	0.00	0.15
	(0.21)	(0.23)	(0.30)	(0.21)	(0.37)	(0.22)	(0.28)	(0.26)	(0.33)	(0.00)	(0.26)	(0.22)	(0.19)	(0.27)	(0.12)	(0.24)	(0.00)	(0.36)
City Size: >500,000	0.08	0.03	0.09	0.08	0.14	0.08	0.08	0.11	0.06	0.01	0.08	0.02	0.03	0.04	0.00	0.00	0.00	0.03
	(0.27)	(0.18)	(0.28)	(0.28)	(0.34)	(0.27)	(0.27)	(0.31)	(0.24)	(0.10)	(0.28)	(0.14)	(0.18)	(0.20)	(0.00)	(0.00)	(0.00)	(0.17)
Schools	158	177	197	216	151	152	149	202	154	96	147	148	197	195	136	149	176	150
Classes	276	235	205	216	200	295	249	239	277	197	240	246	314	243	137	149	351	155
Students	4668	4578	3995	3987	4183	4584	5204	4200	4688	3607	4042	4673	5616	4492	3957	4411	6368	4284

Note: Weighted means are reported. Standard deviations are in parentheses. Variable "SES: Numbers of Books in the Home" takes values from one to five indicating 0-10 books, 11-25 books, 26-100 books, 101-200 books and more than 200 books; it was used as a continuous variable in our analysis for simplicity.

Country abbreviations: AUT = Austria, CZE = Czech Republic, DEU = Germany, DNK = Denmark, ESP = Spain, HRV = Croatia, HUN = Hungary, ITA = Italy, LTU = Lithuania, MLT = Malta, PRT = Portugal, ROM = Romania, SVK = Slovak Republic, SVN = Slovenia, JPN = Japan, TWN = Chinese Taipei, HKG = Hong Kong, SGP = Singapore.

Regression Results

The class size estimates of the OLS regression analysis are summarized in Table 1.3. Negative coefficients of class size indicate that student achievement increases as class size decreases; while positive coefficients indicate that student achievement increases as class size increases. The regression coefficients of class size were negative in eight of the 18 countries, but none of them was significant at the 0.05 level after controlling for student, teacher/classroom, and school characteristics.

Significant and positive class size coefficients were found in Croatia, Hong Kong and Malta. As we discuss in the method section, OLS results might be biased because of omitted variables. I analyzed the impact of the omitted variables (unobservable confounding variables) using approach in Frank (2000). The method is based on the idea that for a confounding variable to change the significance of the variable of interest (e.g., class size) it should be correlated with both the variable of interest and the dependent variable. Frank (2000) developed formulas to calculate the minimum correlations necessary to invalidate the inference. He defined the Impact Threshold for a Confounding Variable (ITCV) as the lowest product of the partial correlation between the dependent variable and the confounding variable and the partial correlation between the variable of interest and the confounding variable that makes the coefficient insignificant. The higher the absolute value of the ITCV is, the more robust the OLS estimate is. Table 1.4 presents the ITCVs, and their corresponding minimum correlations between students score and confounding variables and correlation between class size and the confounding variable, which would invalidate the reference of OLS results for countries and districts with the significant estimates. It should be noted that the correlation coefficients shown in Table

Table 1.3: OLS Regression Estimates and Standard Errors of Class Size

	AUT	CZE	DEU	DNK	ESP	HRV	HUN	ITA	LTU	MLT	PRT	ROM	SVK	SVN	HKG	JPN	SGP	TWN
Class size	-0.78	0.55	0.80	0.68	-0.39	0.60*	0.00	-0.41	-0.51	0.98*	-0.62	0.19	-0.87	0.17	2.71*	-0.08	0.60	-0.84
	(0.59)	(0.54)	(0.51)	(0.72)	(0.67)	(0.30)	(0.53)	(0.78)	(0.42)	(0.42)	(0.84)	(1.08)	(0.60)	(0.57)	(0.45)	(0.22)	(0.45)	(0.75)
Number of Schools	158	173	176	156	139	147	144	179	151	89	143	138	192	188	119	148	176	145
Number of Students	4637	4517	3565	2965	3883	4427	4858	3690	4556	3346	3791	4359	5413	4366	3452	4389	6240	4155
R-sq	0.258	0.254	0.311	0.222	0.277	0.203	0.439	0.193	0.283	0.213	0.304	0.352	0.328	0.236	0.381	0.219	0.453	0.240

* $p < 0.05$

Country abbreviations: AUT = Austria, CZE = Czech Republic, DEU = Germany, DNK = Denmark, ESP = Spain, HRV = Croatia, HUN = Hungary, ITA = Italy, LTU = Lithuania, MLT = Malta, PRT = Portugal, ROM = Romania, SVK = Slovak Republic, SVN = Slovenia, JPN = Japan, TWN = Chinese Taipei, HKG = Hong Kong, SGP = Singapore.

Table 1.4: Analysis of the impact of unobservable confounding variables

	Croatia	Malta	Hong Kong
ITCV|z	0.058	0.072	0.132
Rxcv|z	0.241	0.269	0.363
Rycv|z	0.241	0.269	0.363

1.4 are partial correlations that condition on the covariates included in equation (1.1). Therefore, for instance, in Hong Kong, the result indicates that to sustain an inference an omitted variable would have to be correlated at 0.363 with class size and at 0.363 with mathematics achievement, conditional on all the covariates in equation (1.1). The partial correlation coefficients shown in Table 1.4 ranged from 0.241 to 0.363, which were somewhat large since they conditioned on a group of student, teacher/classroom and school covariates. However, it is still difficult to tell if the significant coefficients in Croatia, Malta and Hong Kong were robust to omitted variables because TIMSS does not provide information such as prior achievement and family income, which are usually highly correlation with student achievement and class size.

The positive class size coefficients are somewhat puzzling. One possible explanation is that parents chose high quality schools for their kids, which increased school enrollment and thus increased the average class size in high quality schools.

IV Results

The first stage regression results are summarized in Table 1.5, and the IV estimates of class size effects in Table 1.6. In 12 countries the first stage regression coefficients of computed average class size were significant and positive, and the t-statistic of the instrument was bigger than 3.20. This indicates that the correlation between reported class size and the instrument is strong enough in these countries (Staiger & Stock, 1997). It should be noted that, for Hong Kong, although the absolute value of the t-statistic was larger than 3.20, it was negative, which indicates the computed class size and the teacher reported class size were negative correlated. That is because, in Hong Kong, the maximum class size rules were only applicable to part of schools but not all of the primary schools.

However, there was no information from TIMSS data to identify which schools should follow the rules. Therefore, our IV methods were not appropriate to Hong Kong. In Denmark, the first stage regression coefficients of class size were also significant and positive, but the t-statistic of the instrument was smaller than 3.20. In Croatia, Italy Malta, and Singapore, the first stage regression coefficients of class size were insignificant, which indicate the IVs were quite weak in these countries.

To sum up, the IVs might not be valid in Hong Kong, Malta and Spain; also, in Croatia, Denmark, Italy, Malta and Singapore, the IVs were weak, which made the IV estimates and inference unreliable. Therefore, I will focus on IV estimates from countries with valid and strong IVs.

The IV estimates of class size are summarized in Table 1.6. The coefficients for Austria, Lithuania, Portugal, Slovenia, Japan and Chinese Taipei were negative but insignificant. The coefficients for the Czech Republic, Germany, and Hungary were positive but insignificant. The estimated class size effects were negative and significant at the 0.05 level in only two countries: Romania and the Slovak Republic. The magnitude of class size coefficients for Romania and the Slovak Republic were about 4.5, which is equivalent to 0.045 SD among all fourth graders who participated in TIMSS 2011. Such results indicate that a one student reduction would increase about 4.5 points (or 0.045 SD) of student mathematics achievement on average in the TIMSS scale.

To facilitate interpretation, we transformed our estimates to effect sizes (standard deviations units) assuming a reduction in class size of 10 students. The effect size was 0.48 SD and 0.44 SD respectively for Romania and the Slovak Republic. Such effect sizes are quite substantial in magnitude and larger than estimates reported in prior studies (e.g.,

Table 1.5: First Stage Regression Estimates and Standard Errors of the Computed Average Class Size

	Countries with Strong IV													Countries with Weak IV				
	AUT	CZE	DEU	ESP	HUN	LTU	PRT	ROM	SVK	SVN	JPN	TWN	HKG	DNK	HRV	ITA	MLT	SGP
IV: Computed Average Class Size	0.50*	0.62*	0.47*	0.70*	0.50*	0.59*	0.38*	0.59*	0.49*	0.45*	0.82*	0.73*	-0.50*	0.35*	0.10	0.26	0.28	0.28
	(0.10)	(0.08)	(0.09)	(0.08)	(0.09)	(0.11)	(0.11)	(0.08)	(0.07)	(0.08)	(0.10)	(0.11)	(0.15)	(0.14)	(0.13)	(0.16)	(0.16)	(0.19)
T-Statistic for IV	5.23	7.94	5.06	8.96	5.28	5.49	3.38	7.38	6.62	5.52	8.12	7.45	-3.28	2.54	0.72	1.7	1.71	1.51
Number of Schools	158	173	176	139	144	151	143	138	192	188	148	145	119	156	147	179	89	176
Number of Students	4637	4517	3565	3883	4858	4556	3791	4359	5413	4366	4389	4155	3452	2965	4427	3690	3346	6240

* $p < 0.05$

Note: Standard errors are in parentheses.

Country abbreviations: AUT = Austria, CZE = Czech Republic, DEU = Germany, DNK = Denmark, ESP = Spain, HRV = Croatia, HUN = Hungary, ITA = Italy, LTU = Lithuania, MLT = Malta, PRT = Portugal, ROM = Romania, SVK = Slovak Republic, SVN = Slovenia, JPN = Japan, TWN = Chinese Taipei, HKG = Hong Kong, SGP = Singapore.

Table 1.6: Second Stage Regression Estimates and Standard Errors of Class Size

	AUT	CZE	DEU	HUN	LTU	PRT	ROM	SVK	SVN	JPN	TWN
Class Size	-1.82	0.24	1.33	0.45	-1.25	-3.80	-4.84*	-4.40*	-1.87	-0.81	-0.83
	(1.27)	(1.16)	(1.26)	(1.49)	(1.28)	(2.67)	(2.28)	(1.58)	(1.32)	(0.46)	(1.23)
Number of Schools	158	173	176	144	151	143	138	192	188	148	145
Number of Students	4637	4517	3565	4858	4556	3791	4359	5413	4366	4389	4155

* $p < 0.05$

Note: Standard errors are in parentheses.

Country abbreviations: AUT = Austria, CZE = Czech Republic, DEU = Germany, HUN = Hungary, LTU = Lithuania, PRT = Portugal, ROM = Romania, SVK = Slovak Republic, SVN = Slovenia, JPN = Japan, TWN = Chinese Taipei.

Angrist & Lavy, 1999). A reduction of eight students, which was by and large the average reduction in number of students between regular size and small size classes in Project STAR, would indicate an increase in mathematics achievement nearly one-third of a SD. This is a considerable effect knowing that the average benefit for students in small classes in Project STAR was nearly 0.20 SD.

Comparison of Regression and IV Estimates

Finally, I examined whether IV estimates were indeed different than regression estimates that could be biased. To compare OLS and IV estimates, we used the Durbin-Wu-Hausman test (Durbin, 1954; Hausman, 1978; Wooldridge, 2010; Wu, 1973). Specifically, we ran the regression

$$Score_i = \delta_0 + \delta_1 Residual_i + \delta_2 ClassSize_i + \mathbf{ST}_i\Delta_2 + \mathbf{CL}_i\Delta_3 + \mathbf{SC}_i\Delta_4 + \xi_i \qquad (1.6)$$

where $Residual_i$ is the residual term from regression equation (1.3). The idea is that once we control for reported class size (and other covariates) the coefficient of the residuals should not be significant unless there is omitted variable bias. The significance of δ_1 indicates that the regression and IV estimates are different. The significance of δ_1 indicates the reported class size is endogenous, that is, reported class size is correlated with omitted variables that are part of the error term of equation (1.3). Table 1.7 summarizes the results of the Durbin-Wu-Hausman test for the full samples. Significant estimates at 0.05 level were found in Romania and the Slovak Republic; while significant estimates at 0.10 level were found in Portugal, Slovenia, and Japan. The results suggest the regression and

Table 1.7: Results from Durbin-Wu-Hausman Test

	AUT	CZE	DEU	HUN	LTU	PRT	ROM	SVK	SVN	JPN	TWN
First Stage Residual	1.35	0.43	-0.69	-0.53	0.87	3.46+	6.67*	4.38*	2.55+	0.95+	-0.02
	(1.34)	(1.14)	(1.38)	(1.52)	(1.26)	(2.00)	(2.22)	(1.68)	(1.33)	(0.50)	(1.43)
Number of Schools	158	173	176	144	151	143	138	192	188	148	145
Number of Students	4637	4517	3565	4858	4556	3791	4359	5413	4366	4389	4155

* $p < 0.05$, + $p < 0.10$

Note: Standard errors are in parentheses.

Country abbreviations: AUT = Austria, CZE = Czech Republic, DEU = Germany, HUN = Hungary, LTU = Lithuania, PRT = Portugal, ROM = Romania, SVK = Slovak Republic, SVN = Slovenia, JPN = Japan, TWN = Chinese Taipei.

IV estimates are different in these countries. They also indicate that reported class size was endogenous and thus correlated with omitted variables in these countries. These results support the notion that that IV analysis was necessary and that the IV estimates should capture the causal effects of class size on student achievement in these two countries. For other countries with strong and valid instruments -Austria, the Czech Republic, Germany, Hungary, and Lithuania, and Chinese Taipei- the results indicate that estimates from regression and IV analyses were overall similar. These findings may suggest that there is little bias from omitted variables in the regression analysis in these countries.

Discussion

I investigated the effects of class size on mathematics achievement for fourth graders in 18 countries and districts in 2011 using rich data from TIMSS. These European and Asian countries and districts had maximum class size limits, which allowed me to use an IV approach to explore the causal effects of class size on student achievement. Both regression analyses and IV analyses were conducted. By and large, I did not observe significant class size effects in most countries. Significant class size coefficients at the 0.05 level were found in Romania and the Slovak Republic. These coefficients indicated that class size reductions increased mathematics achievement significantly and meaningfully. The estimates produced from the IV analysis were somewhat different than those from the OLS analysis in some countries. The Durbin-Wu-Hausman test provided some evidence that reported class size was correlated with omitted variables in some countries and that the IV analysis was necessary and provided valid estimates of class size effects in Romania

and the Slovak Republic. In other countries however, the regression estimates were similar to the IV estimates, which suggests that regression estimates were as good as IV estimates.

Generally, the results indicated no systematic pattern of association between class size and achievement. For nine of the eleven countries and districts with strong and valid IV no class size effects were found. The exceptions were Romania and the Slovak Republic. These significant class size effects were quite substantial in magnitude compared to prior studies (e.g., Angrist & Lavy, 1999). Nonetheless, my findings are in congruence with findings of previous work that used prior cycles of TIMSS assessments and have indicated generally no significant relationships between class size and achievement (Pong & Palls, 2001; Wossmann, 2005; Wossmann & West, 2006). Romania and the Slovak Republic are not as wealthy or developed countries compared to the other European countries in our sample, which might indicate that school resources such as class size reduction may play a more important role in less wealthy countries.

Unfortunately TIMSS does not provide data about classroom dynamics, instruction, and practices and therefore it is difficult to know exactly why we failed to detect class size effects in most countries. Prior studies have suggested that class size have positive effects when teachers spend more time on individualized instruction or when pupils become more involved in learning activities (e.g., Finn & Achilles, 1990). Perhaps in most of my samples teachers did not utilize individualized instruction when class size was reduced. Also, perhaps students were not as actively involved in learning activities when class size was reduced.

One possible limitation of our estimates is related to the enrollment information we used in our models. Specifically, enrollment information from the beginning of the school

year can predict average class size more accurately (see Angrist & Lavy, 1999). However, the enrollment information available in TIMSS is at the time of testing, which is near the end of the school year. Thus, we could not control for any enrollment changes during the school year. If potential changes of enrollment are not random, our results might be biased, and that's a potential limitation of our study (Wossmann, 2005).

Another potential limitation is that our IV method may not be valid. Although we tested if covariates were locally balanced across schools around cut-offs, it is unclear whether enrollment influences student achievement only though class size once enrollment and other important covariates are controlled for. However, if class size is related to unobserved variables that we could not control for (e.g., parental education level or family income) then our IV estimates may be biased.

CHAPTER 2 DOES CLASS SIZE REDUCTION CLOSE THE ACHIEVEMENT GAP

Introduction

The effects of class size on student achievement have been discussed repeatedly in education research and policy in the past decades. Meta-analytic reviews of early work on small class effects (e.g., Glass & Smith, 1979) and studies using data from a high-quality large-scale experiment (e.g., Finn & Achilles, 1990) indicated a positive relationship between small classes and student achievement. In particular, evidence from Project STAR (Student-Teacher Achievement Ratio) in Tennessee has strongly indicated achievement improvements for students in small classes compared to regular size classes (e.g., Krueger, 1999; Nye, Hedges, & Konstantopoulos, 2000). These findings suggest that reducing class size is a promising policy option to increase academic achievement, on average, for all students.

Besides improving average student achievement, another critical objective of education interventions is to increase achievement for students at risk, and thus reduce the achievement gap between lower- and higher-achieving students. Class size reduction has been advocated as such an intervention by some researchers (e.g., Finn & Achilles, 1990). One way to evaluate whether CSR can close the achievement gap is to examine the interaction effect between class size and student background such as gender, socioeconomic status (SES), minority status, etc. Prior studies have focused typically on the average effects of class size on student achievement for all students. Only a few studies have examined the differential class size effects for subgroups of students, most of which have used data from Project STAR. The findings of these studies were mixed. For example,

Finn and Achilles (1990) found some evidence that the positive effects of small classes were larger for minority students, especially in kindergarten and first grade, while Nye, Hedges, and Konstantopoulos (2002) found weak or no evidence for differential effects of small classes on minority and low-SES students. Another way to evaluate whether CSR can close the achievement gap is to estimate the differential class size effects across student achievement distribution using quantile regression. Konstantopoulos (2008) used quantile regression to evaluate the small size effects for student in the middle and tails of the achievement distribution using data from Project STAR and found that reductions in class size did not reduce the achievement gap between low- and high-achievers in the early grades. Later studies have found similar findings using the same data (Ding & Lehrer, 2011; Jackson & Page, 2013). Nevertheless, there is some evidence that the cumulative effects of being in a small class from kindergarten through third grade may reduce the achievement gap in reading and science in some of the later grades four through eight (Konstantopoulos & Chung, 2009). However, no recent study has used current data to evaluate if CSR closes the achievement gap.

Chapter 2 was designed to fill in that gap in the literature and explore the differential class size effects for students with different levels of achievement. In particular, Chapter 2 examined the effects of class size across the student achievement distribution (i.e., middle and upper or lower tails), in an attempt to address the question of whether CSR closes the achievement gap between high- and low-achievers, using the latest cycle of a large-scale international assessment program.

Specifically, I used the data from the 2011 fourth grade sample of the Trends in International Mathematics and Science Study (TIMSS). I utilized maximum class size rules

available in some countries to gauge class size effects on mathematics achievement. I employed quantile regression to estimate class size effects on student achievement in the middle as well as in the lower and upper tails of the achievement distribution. To deal with the potential endogeneity of class size, I computed the average class size in a school based on the maximum class size rule in each country, which was used as an instrumental variable (IV) for class size. I used the control function approach (see Lee, 2007) to estimate the differential causal effects of class size effect on fourth graders' mathematics achievement.

Chapter 2 contributes to the existing literature in two ways. First, I used the most recent TIMSS data from 2011 that allowed us to evaluate recent, concurrent CSR policies, and to compare class size effects across Asian and European countries and districts. Second, I used quantile regression coupled with IV to evaluate causal class size effects across the achievement distribution. To my knowledge, the TIMSS data have not been used to examine differential class size effects, although some researchers have used previous cycles of TIMSS assessment to evaluate average class size effects (e.g., Pong & Palls, 2001; Wossmann, 2005; Wossmann & West, 2006).

Literature Review

During the past three decades, researchers explored the effects of class size reduction on student achievement through meta-analyses, experimental and quasi-experimental designs (e.g., RD), as well as other advanced statistical methods such as IV. Most researchers have focused exclusively on estimating mean differences in student achievement between small and regular-size classes (Konstantopoulos, 2008). For example, meta-analytic reviews of early work on small class effects have indicated a positive

relationship between small classes and student achievement, but the magnitude of the effect was small (e.g., Glass and Smith, 1979; Slavin, 1989).

Project STAR is viewed as the most impressive and most powerful field experiment about class size effects in education (Mosteller, 1995). There have been numerous analyses of the Tennessee STAR data that have produced high internal validity estimates. Finn and Achilles (1990) were the first to analyze these data, and they found that students in small classes performed higher than those in regular classes in all subject areas, and in every year of the experiment (kindergarten through third grade). Nye, Hedges, and Konstantopoulos (2000) examined the validity of Project STAR, and they suggested that the effects of class size might be under-estimated because of imperfect implementation. They also found that the estimated class size effects were consistent with those from Glass and Smith (1979).

Researchers also attempted to evaluate average class size effects using observational data. The main difficulty of analyzing observational data is that the internal validity (or unbiasedness) of estimates in observational or quasi-experimental studies is not so easy to achieve. That is, researchers have to use advanced statistical methods to warrant the high internal validity of estimates for observational data. Previous work has utilized different analytic methods to examine class size effects on student achievement. For example, Pong and Pallas (2001) used multilevel models to analyze TIMSS 1995 data in nine different countries and found no class size effects on eighth grade achievement except in the U.S. Other researchers have used IV methods to analyze observational data in an attempt to explore the causal effects of class size reduction (e.g., Akerhielm, 1995; Hoxby, 2000; Cho, Glewwe, & Whitler, 2012; Wossmann & West, 2006).

One of the best instruments used to capture class size effects was introduced by Angrist

and Lavy (1999). They used the Maimonides rule that sets the maximum class size to 40 students per classroom in order to evaluate the effect of class size on student achievement in Israel. The authors used this maximum class size rule of 40 to construct IV estimates of class size on test scores. They found a statistically significant effect of small classes on fifth grade reading and mathematics scores. However, they found no significant effects of class size on third grade scores.

Several other researchers have also used maximum class size rules as IV to evaluate class size effects. For instance, Bonesronning (2003) investigated class size effects using a maximum class size rule of 30 students per classroom in Norway. His analysis indicated small class effects. Wossmann (2005) explored class size effects in Europe using data from TIMSS 1995 for eighth grade students. He found two statistically significant and negative relationships between class size and student achievement in Norway and Iceland. He also found a statistically significant but positive relationship between class size and student achievement in Switzerland. For Denmark, France, Germany, Greece, Ireland, Spain, and Sweden, the estimates were not significant. A recent study about class size effects on fourth grade reading achievement in Greece also reported statistically insignificant estimates (Konstantopoulos & Traynor, 2014). Urquiola (2006) studied third-grade students in Bolivia and found significant class size effects, with effect sizes as large as 0.30 standard deviations, bigger than the effects found in Project STAR in the U.S. and in Israel.

Class size reduction can potentially affect average student achievement as well as the achievement gap among subgroups of students. In other words, interactions between class size effects and student background, such as student SES or achievement level, are possible (Konstantopoulos & Chung, 2009). If economically disadvantaged students or low-

achieving students benefit more from being in smaller classes, CSR would decrease the achievement gap. However, most prior studies have focused on the average class size effects, while only few studies have explored the interaction effects between class size and student backgrounds or achievement levels.

The differential effects of class size have traditionally been determined through statistical interactions between class size and student variables such as gender, SES, and race. Project STAR data have been used to examine such interaction effects. For example, early analyses have reported that class size reduction had larger positive effects for minority students (see Finn & Achilles, 1990). These average differences were significant for reading achievement for the first two years of the experiment (kindergarten and first grade). However, more recent studies could not fully replicate these findings. For example, Nye, Hedges, and Konstantopoulos (2000) found weak evidence that class size reduction had larger benefits for minority students. Also, Nye, Hedge and Konstantopoulos (2002) examined the differential effects of small classes for students who were low-achievers in previous grades, and they found no evidence of additional small class benefits for these students.

Several non-experimental studies have also evaluated class size effects for subgroups of students, and almost all of them did not find differential class size effects. For example, Hoxby (2000) analyzed data from a natural experiment in Connecticut and found no evidence of class-size effects at schools that served high percentages of economically disadvantaged or minority students. In a similar study, Cho, Glewwe, and Whitler (2012) found the estimated class size effects did not differ by race/ethnicity, gender, or free lunch eligibility. One exception was the study by Jepsen and Rivkin (2009), which found

differential class size effects among subgroups. They analyzed the CSR policy in California and found that this policy initially helped economically advantaged (both in family background and performance) students more than their less affluent peers.

One appropriate method of examining differential class size effects at different levels of achievement is quantile regression, which examines class size effects across the entire student achievement distribution. Konstantopoulos (2008) employed this approach to estimate class size effects at the tenth, twenty-fifth, fiftieth, seventy-fifty, and ninetieth quantiles, using data from Project STAR. He also constructed t-tests to examine whether the estimates were statistically different across quantiles and found some evidence that higher-achieving students benefited more from being in small classes in certain early grades than other students. Later studies confirmed such findings (e.g., Ding & Lehrer, 2011; Jackson & Page, 2013). Nevertheless, Konstantopoulos and Chung (2009) examined the long-term effects of class size across the student achievement distribution. They found that for certain grades (fourth and sixth grade) in reading and science, low- achievers benefited more from being in small classes consistently in the early grades, while for other grades, no differential class size effects were found.

Very few previous studies examined quantile-specific class size effects using non-experimental data. To our knowledge, there were only two studies. One is by Levin (2001), who used quantile regression as well as IV methods through two-stage least absolute deviations (2SLAD) (Amemiya, 1987) to estimate the causal effects of class size on scholastic achievement across various points in the conditional distributions of mathematics and languages achievement of Dutch primary school students. He did not find any significant class size effects at any quantile. Levin (2001) did not examine differences

between estimates across quantiles, and thus it is not clear whether CSR reduced the achievement gap. Ma and Koenker (2006) reanalyzed Levin's data and found that, for mathematics scores, lower-achieving students benefited more from smaller classes while average and high-achieving students did not get benefit from smaller classes.

To sum up, it is not very clear if class size reduction would decrease achievement gap or not; also, there were quite limited studies that evaluated class size effects across achievement distribution. It is necessary to provide more evidence of class size effects across achievement distribution using concurrent data.

Method

In this chapter, I also used the data from TIMSS 2011, and focused on fourth grade mathematics achievement. I analyzed the same countries as I did in Chapter 1. Table 1.1 provides detail about the selected countries as well as their upper class size limits.

Quantile Regression

The objective of my study was to examine class size effects across the distribution of fourth graders' mathematics achievement, especially the effects in the upper and lower tails of the distribution. Ordinary least squares (OLS) regression fails to describe the full distributional impact of class size on student achievement, unless the lower-achievers and higher-achievers benefit the same from smaller classes as students in the middle of the achievement distribution. Quantile regression (Koenker and Bassett, 1978) is a tool that allows researchers to estimate quantile-specific class size effects, not only in the middle but also in the tails of the conditional student mathematics achievement distribution. Thus, we used quantile regression, and compared quantile-specific class size effects across

different quantiles of the achievement distribution to evaluate whether CSR closes or enlarges the achievement gap.

I evaluated class size effects at the tenth, twenty-fifth, fiftieth, seventy-fifth, and ninetieth quantiles through the following equation

$$Score_i = \beta_0 + \beta_1 ClassSize_i + \mathbf{ST}_i\mathbf{B}_2 + \mathbf{CL}_i\mathbf{B}_3 + \mathbf{SC}_i\mathbf{B}_4 + \varepsilon_i \quad (2.1)$$

where $Score_i$ represents mathematics scores, β_0 is the constant term, *ClassSize* is the main independent variable, β_1 represents the class size effect and is the coefficient of interest, $\mathbf{ST_i}$ is a row vector of student background characteristics, $\mathbf{B}_2$ is a column vector of regression coefficients of student characteristics, $\mathbf{CL_i}$ is a row vector of classroom or teacher characteristics, $\mathbf{B}_3$ is a column vector of regression coefficients of teacher and classroom characteristics, $\mathbf{SC_i}$ is a row vector of school characteristics, $\mathbf{B}_4$ is a column vector of regression coefficients of school characteristics, and ε_i is the error term.

Instrumental Variable and Control Function

An important issue to consider in estimating quantile-specific class size effects is that class size may be endogenous because of omitted variable bias. The relative position of students in the conditional achievement distribution could be related to systematic differences in unobservables, such as motivation, family background, school or teacher quality, etc. In that case, the estimated class size effect from equation (2.1) cannot reflect the true quantile-specific class size effect.

Because students and teachers are rarely randomly assigned to classrooms in a grade

class size might be correlated with unobserved characteristics of students or teachers. For example, in order to help low achieving students, some schools might assign higher quality teachers to classes with higher proportions of low achievers. Variables that determine assignment of students and teachers to classes are not typically measured. For example, student motivation, family income, parental pressure, teacher quality, etc. are rarely available in observational datasets. In addition, cross-sectional data rarely provide indexes of prior ability or performance. Although we included as many covariates as we could in our multiple regression analysis, it is still possible that unobservable factors that are part of the error term in equation (2.1) are correlated with class size. If that were true, then the estimated class size effect in equation (2.1) would be biased.

In general, there are two sources of omitted variable bias that are related to student mathematics achievement, and to class size as well. First, students do not choose schools randomly but typically attend schools in their neighborhoods. Therefore, students within the same school might share common characteristics, such as parents' education, parents' occupations, and family income. That is, class size may be correlated with SES manifested via parents' occupations or family income. Such variables were not measured or reported in the TIMSS 2011 fourth grade student survey. Second, students and teachers are rarely randomly assigned to classrooms, and thus class size might be correlated with unobserved student or teacher characteristics. For example, students may be assigned to classes based on their ability or motivation. TIMSS 2011, being a cross-sectional survey, did not include information about prior achievement (a proxy for ability). In the same vein, in order to help low-achieving students, some schools might assign higher quality teachers to classes with higher proportions of low-achievers. There were only very few teacher characteristics

reported in the TIMSS 2011 teacher survey, such as their gender, experience, and education level, which may capture only partially teacher "quality." Although we included as many covariates as we could in our analysis, there are likely unobservable factors that could be correlated with class size that are part of the error term of equation (2.1).

Just as with OLS, endogeneity of class size renders quantile-specific estimates biased. To overcome this potential shortcoming and to facilitate causal inferences, we used IV methods. Specifically, we created a grade and school specific average class size variable using the maximum class size rule, and we used it as an instrument for class size. Our method is similar to the one used by Angrist and Lavy (1999) and is the same as we did in Chapter 1. The average class size in fourth grade, based on the maximum class size requirement, could be calculated through the following equation

$$f_i = E_i / [\text{int}((E_i - 1) / rule) + 1] \qquad (2.2)$$

where E_i denotes the enrollment in grade four in a school; f_i denotes the computed school and grade specific average class size based on the maximum class size rule; *rule* denotes the upper class size limit in a given country; and for any positive number n, the function int(n) is the largest integer less than or equal to n.

I adopted the control function approach proposed by Lee (2007) to get quantile-specific IV estimates. Lee's approach fits our study for two reasons: first, his estimation approach is computationally convenient and simple to implement through the "qreg" command in STATA; second, the required assumptions by Lee's control approach hold in general settings (see Lee, 2007). The control function approach is also a two-stage

estimation method that is similar to two-stage-least square (2SLS). The basic idea is to add a control variable to equation (2.1) such that, once we condition on this variable, the teacher reported class size will be independent of omitted variables (see Wooldridge, 2010). This so-called control variable usually needs to be estimated through a first stage regression, because it cannot be observed or measured directly. In our study, the first stage regression equation is

$$ClassSize_i = \pi_0 + \pi_1 f_i + \mathbf{ST}_i\Pi_2 + \mathbf{CL}_i\Pi_3 + \mathbf{SC}_i\Pi_4 + u_i \quad (2.3)$$

where f_i is the computed average class size in a school based on the maximum class size rule, and u_i is the error term. All other terms have been defined previously. The π's are the regression estimates that need to be computed.

Researchers typically use the estimated residuals from equation (2.3) as the control variable. Residuals can be estimated from a quantile regression, or even an OLS regression (Lee, 2007). It should be noted that Lee's control function method is only applicable to continuous endogenous variables. Although class size is conceptually continuous, it has only a finite number of distinct values. In this case, Lee (2007) suggested using OLS regression in the first stage. I calculated residual $\hat{u}_i$ through the following equation

$$\hat{u}_i = ClassSize_i - \widehat{ClassSize_i}$$

where $\widehat{ClassSize_i}$ is the fitted value of $ClassSize_i$ from equation (2.3), the OLS regression. Contrary to the conventional control function approach that inserts $\hat{u}_i$ into equation (2.1)

as the second stage regression, Lee (2007) proposed inserting a power series or kernel of $\hat{u}_i$. He showed that with proper conditions, the estimator from his control function approach is consistent (See Appendix B for a proof). In this study, I added a fifth order polynomial of $\hat{u}_i$, denoted as $\lambda(\hat{u}_i)$, into equation (1). Specifically, the second stage regression in each quantile (i.e., tenth, twenty-fifth, fiftieth, seventy-fifth, and ninetieth) is

$$Y_i = \delta_0 + \delta_1 ClassSize_i + \lambda(\hat{u}_i) + \mathbf{ST}_i\Delta_2 + \mathbf{CL}_i\Delta_3 + \mathbf{SC}_i\Delta_4 + \xi_i \qquad (2.4)$$

The coefficient δ_1 represents the relationship between mathematics achievement and class size, adjusted for student, teacher/classroom, and school characteristics; $\lambda(\hat{u}_i)$ represents a fifth order polynomial of $\hat{u}_i$. The δ 's indicate regression estimates that need to be computed. The student, classroom/teacher, and school covariates included in equation (2.4) are the same as those included in equation (2.3) (see Appendix A). Appropriate student weights were used in both regressions (equations 2.3 and 2.4).

It should be noted that due to the two-step feature of the model, the standard errors of estimates in equation (2.4) were adjusted by nonparametric bootstrap techniques using 1000 replications. I used the bootstrap method introduced by Kelnikov (2010), which is suitable for complex survey data and corrects the potential clustering effects (i.e., students nested within schools). Also, my analysis was conducted for each plausible value separately, and then the averages of the five sets of estimates were calculated and reported as the final estimates of class size effects for each quantile (see Schafer & Olsen, 1998). The standard error of the class size effects was a combination of the sampling variance obtained through

bootstrap techniques and the variance between plausible values (see Martin & Mullis, 2012).

Similar to the case in 2SLS context, there were two key assumptions that the computed average class size f_i must meet in order for the variable to be a valid IV: (1) f_i should be correlated with actual class size, and (2) f_i should not be correlated with the error term in equation (2.1).

The first assumption indicates that schools followed the maximum class size requirement when they assigned students to classrooms. In a 2SLS context, such an assumption can easily be tested through the first stage regression. If the instrument is only marginally significant, our instrument could be weak. When instruments are weak, then the standard IV estimates, hypothesis tests, and confidence intervals may be unreliable (Stock, Wright, & Yogo, 2002). The rule of thumb is that the t-statistic of the instrument in the first-stage regression should be larger than 3.2 (Stock, Wright, & Yogo, 2002). Results from Table 1.6 in Chapter 1 had shown that there were five countries or districts - Denmark, Croatia, Italy, Malta and Hong Kong- whose IVs were weak. In addition, the significant but negative coefficient in Hong Kong indicated that the IV was valid in Hong Kong.

Results

I only evaluated the class size effects for countries and districts with strong and valid IVs. The quantile-specific IV estimates of class size are summarized in Table 2.1. To compare the results between OLS regression and median regression (quantile regression at the fiftieth quantile), estimates from 2SLS are also presented. Negative coefficients of class

size indicate that student achievement increases as class size decreases, which is what researchers and policy makers expect. In Romania, the Slovak Republic, Slovenia, Japan

Table 2.1: 2SLS and Quantile Regression Estimates and Standard Errors of Class Size

	2SLS	Quantile				
		10th	25th	50th	75th	90th
AUT	-1.82	0.35	-2.07	-2.26	-2.51	-1.71
	(1.27)	(3.37)	(3.04)	(2.04)	(2.15)	(2.79)
CZE	0.24	0.83	0.67	0.75	-0.10	-0.95
	(1.16)	(1.74)	(1.48)	(1.28)	(1.56)	(1.48)
DEU	1.33	2.25	2.28	2.26	1.58	-0.30
	(1.26)	(2.28)	(1.97)	(1.68)	(2.16)	(2.42)
HUN	0.45	1.31	-0.50	0.70	1.19	1.17
	(1.49)	(2.65)	(2.24)	(1.78)	(1.92)	(2.44)
LTU	-1.25	-0.06	-0.15	-0.79	-2.15	-2.81
	(1.28)	(3.27)	(1.65)	(1.75)	(1.71)	(2.41)
PRT	-3.80	-2.84	-2.89	-2.89	-3.05	-4.68
	(2.67)	(4.63)	(3.10)	(2.78)	(3.10)	(3.69)
ROM	-4.84*	-5.46	-5.52	-5.72*	-4.86+	-4.23
	(2.28)	(3.63)	(3.71)	(2.64)	(2.92)	(3.46)
SVK	-4.40*	-4.42*	-3.59	-4.10+	-4.68*	-4.33
	(1.58)	(2.24)	(2.57)	(2.19)	(2.16)	(2.78)
SVN	-1.87	-1.13	-1.69	-2.03	-2.65	-2.70
	(1.32)	(2.69)	(2.09)	(1.86)	(2.31)	(2.52)
JPN	-0.81	-1.45	-1.03	-0.71	-0.51	-0.18
	(0.46)	(0.92)	(0.81)	(0.57)	(0.63)	(0.79)
TWN	-0.83	-1.83	-0.27	-0.63	-0.15	0.21
	(1.23)	(2.53)	(2.27)	(1.79)	(1.86)	(1.93)

*$p \leq .05$ + $p \leq 0.1$

Note: Bootstrap standard errors are in parentheses.

and Chinese Taipei, the magnitude of the coefficients in the median regression were similar to those from 2SLS. In Germany, Hungary, Lithuania, and the Czech Republic, the magnitude of the coefficients in the median regression were quite different from those from 2SLS. In terms of significance, the estimates from 2SLS and the estimates from median regression were quite similar and, by and large, insignificant. In addition, the standard errors from the median regression were larger than those from 2SLS.

The coefficients of class size were negative but insignificant across all five quantiles

in Lithuania, Japan, Portugal, and Slovenia. The coefficients for Austria, the Czech Republic, Germany, Hungary, and Chinese Taipei were mixed: for some quantiles, they were positive, while for the other quantiles, they were negative. However, none of the quantile estimates were significant. Negative and significant quantile-specific class size estimates were only found in Romania and the Slovak Republic. In Romania, the class size coefficient at the fiftieth quantile was negative and significant at the 0.05 level. Also, the class size coefficient at the seventy-fifth quantile was negative and significant at the 0.10 level. Such results indicate that students in the middle and upper tail of the achievement distribution benefitted from being in smaller classes. For instance, a one student reduction corresponds to an increase of about 5.7 points of mathematics achievement in the TIMSS scale for students in the middle of the achievement distribution. This is equivalent to about 0.057 standard deviations (SD) among all fourth graders who participated in TIMSS 2011. For the other three quantiles, the estimates were negative but insignificant. The magnitude of the class size coefficients were similar across quantiles and ranged between 4.23 at the ninetieth quantile to 5.72 at the fiftieth quantile.

In the Slovak Republic, the estimates at the tenth quantile and seventy-fifth quantile were significant and negative at the 0.05 level. The estimate at the fiftieth quantile was negative and significant at the 0.10 level. Such results indicate that students in the lower tail, median or upper tail of the achievement distribution benefitted from smaller classes. For the other two quantiles (twenty-fifth and ninetieth quantiles), the estimates were negative but insignificant. The magnitude of the class size coefficients were similar across quantiles and ranged between 3.59 at the twenty-fifth quantile to 4.68 at the seventy-fifth quantile.

To facilitate interpretation, I transformed the estimates to effect sizes expressed in SD units, assuming a reduction in class size of eight students, which was the average class size reduction in Project STAR. For Romania, the effect sizes were about 0.46 SD at the fiftieth quantile, and about 0.39 SD at the seventy-fifth quantile. For the Slovak Republic, the effect sizes were about 0.36 SD at the tenth quantile and the seventy-fifth quantile, and about 0.33 SD at the fiftieth quantile. Such effect sizes are quite substantial in magnitude and larger than the conditional mean estimates reported in prior studies (e.g., Angrist and Lavy, 1999; Nye, Hedges & Konstantopoulos, 2004). For example, the average effect size for Project STAR was about 0.20 SD.

In Japan the magnitude of the coefficients indicated that the class size effects were consistently larger for low-achievers than for other students. For example, the magnitude of the coefficient estimated at the tenth quantile was more than eight times larger than that at the ninetieth quantile. In countries such as the Czech Republic, Lithuania, Portugal, and Slovenia, the magnitude of the coefficients indicated that the class size effects were consistently larger for higher-achievers than for other students. For example, the magnitude of the coefficient estimated at the ninetieth quantile was about 47 times as large as that at the tenth quantile in Lithuania. Overall these results seem mixed. In some countries, the results seem to support the notion that high-achieving students may benefit more from being in small classes than other students. In contrast, in other countries low-achievers seem to benefit more from smaller classes than other students. Still, one needs to examine whether the estimates across these different quantiles were statistically significant.

A bootstrap procedure was employed to compute the standard errors of the differences between two quantile-specific estimates (Kelnikov, 2010). Table 2.2 summarizes the

differences between estimated class size coefficients and their bootstrap standard errors. I calculated the difference between two specific-quantile estimates by subtracting the estimated class size coefficient of lower achievers from the estimated class size coefficients

Table 2.2: Differences in Quantile Regression Estimates

	90th vs. 10th Quantile	90th vs. 25th Quantile	90th vs. 50th Quantile	75th vs. 10th Quantile	75th vs. 25th Quantile	75th vs. 50th Quantile	50th vs. 25th Quantile	50th vs. 10th Quantile
AUT	-2.07	0.36	0.54	-2.86	-0.44	-0.25	-2.61	-0.19
	(3.01)	(2.93)	(1.83)	(2.28)	(2.16)	(0.94)	(1.73)	(1.60)
CZE	-1.78	-1.62	-1.71	-0.93	-0.77	-1.05	0.26	0.23
	(1.17)	(1.26)	(1.17)	(1.35)	(1.24)	(1.11)	(1.13)	(1.09)
DEU	-2.42	-2.59	-2.56	-0.96	-0.71	-0.68	-0.14	-0.02
	(2.02)	(1.52)	(1.37)	(1.77)	(1.36)	(0.98)	(1.57)	(0.83)
HUN	-0.14	1.32	0.13	-0.12	1.57	0.38	-0.39	1.19
	(2.05)	(1.90)	(1.40)	(1.83)	(1.55)	(1.11)	(1.59)	(1.26)
LTU	-2.75	-2.66	-2.02	-2.08	-1.99	-1.36	-0.73	-0.64
	(2.79)	(1.55)	(1.32)	(2.59)	(1.25)	(1.00)	(2.02)	(0.99)
PRT	-2.45	-1.79	-2.21	-0.72	-0.48	0.03	-0.18	0.41
	(4.76)	(3.27)	(2.55)	(4.22)	(2.55)	(1.67)	(3.70)	(2.02)
ROM	1.23	1.30	1.50	0.60	0.66	0.86	-0.27	-0.20
	(3.08)	(3.26)	(2.18)	(2.18)	(2.42)	(1.22)	(1.45)	(1.86)
SVK	0.20	-0.74	-0.23	-0.09	-0.79	-0.51	0.18	-0.51
	(2.30)	(2.01)	(1.42)	(2.02)	(1.69)	(1.02)	(1.55)	(1.42)
SVN	-1.57	-1.01	-0.67	-1.53	-0.97	-0.62	-0.90	-0.34
	(1.92)	(1.56)	(1.42)	(1.70)	(1.29)	(0.90)	(1.46)	(1.10)
JPN	1.27*	0.85	0.50	0.94	0.52	0.25	0.65	0.26
	(0.60)	(0.66)	(0.41)	(0.61)	(0.62)	(0.29)	(0.50)	(0.45)
TWN	2.04	0.48	0.85	1.68	0.12	0.49	1.19	-0.37
	(1.81)	(1.37)	(1.20)	(1.61)	(1.15)	(0.92)	(1.51)	(0.81)

*$p \leq .05$ + $p \leq 0.1$

Note: Bootstrap standard errors are in parentheses.

of higher achievers. Thus, a negative difference indicated that high-achievers benefitted more from small classes than low-achievers. For example, in Japan, the difference of the class size coefficients between the ninetieth and the tenth quantile was 1.27, which indicates that a one student reduction in class size would increase achievement by 1.27 points in the mathematics achievement scale between these two quantiles (favoring the tenth quantile). In other words, negative difference indicates an increase in the achievement gap between high-achievers and low-achievers as class size decreases. In contrast, a

positive difference indicates a decrease in the achievement gap between high-achievers and low-achievers as class size decreases.

The results in Table 2.2 show that almost all differences between any two specific-quantile estimates were insignificant with only one exception, which indicates that in general CSR did not reduce the achievement gap between high- and low-achievers. By and large, CSR is likely to have no impact on the achievement gap across countries, which is inconsistent with prior studies, especially the studies using data from Project STAR (e.g., Konstantopoulos, 2008; Ding & Lehrer, 2011; Jackson & Page, 2013) that consistently found high-achieving students got more benefit from small classes and thus achievement gap increased.

Discussion

I investigated the differential effects of class size at different levels of mathematics achievement for fourth graders, using rich data from TIMSS 2011. The European and Asian countries and districts I selected had maximum class size rules, which allowed me to use an IV approach to explore the causal effects of class size on student achievement across the achievement distribution. Specifically, I used a control function approach, coupled with quantile regression, to examine differential class size effects for students in the middle, lower, and upper tails of the achievement distribution.

Generally, the findings from the quantile regression indicated no systematic patterns of association between class size and achievement. In nine of the eleven European and Asian countries and districts that had strong IV and valid RE design, we found insignificant class size effects. The only two exceptions were Romania and the Slovak Republic, where

significant class size effects were detected in some quantiles. These significant class size effects were quite substantial in magnitude compared to prior studies (e.g., Angrist & Lavy, 1999). Nonetheless, my findings are in congruence with the findings of previous work that used prior cycles of TIMSS and have indicated generally insignificant relationships between class size and achievement (Pong & Palls, 2001; Wossmann, 2005; Wossmann & West, 2006). I also compared class size coefficients at the lower and upper tails of the achievement distribution. These results suggest no differential class size effects across the achievement distribution. In sum, our findings suggest that CSR has no impact on achievement gap between low- and high-achieving students. In other words, lower-achieving students did not get hurt from CSR policies. Such findings are not in congruence with findings of previous works that used high-quality experimental data or (e.g., Konstantopoulos, 2008; Nye, Hedges & Konstantopoulos, 2002). In addition, our findings indicates that, for some specific countries such as Romania and the Slovak Republic, CSR is a promising policy that would increase student achievement but not increase achievement gap between low- and higher-achieving students.

CHAPTER 3 POWER CONSIDERATION FOR MODEL OF CHANGE

Introduction

In recent years, there has been an increased interest in assessing the effects of educational interventions via experimental designs where students, classrooms, or schools are randomly assigned to a treatment and a control condition. An important part of the design phase of an experiment involves power analysis. Statistical power is the probability of detecting the treatment effect of interest when it exists (Boruch & Gomez, 1977; Cohen, 1988). A priori power computations are critical in designing experiments because they inform empirical researchers about the sampling scheme needed to detect a treatment effect. Specifically, a priori power analyses help educational researchers identify how big a sample is needed at the student, classroom, or school level to ensure a high probability (e.g., > 80 percent) of detecting a treatment effect if it were true (Lipsey 1990; Konstantopoulos, 2008a).

The recent resurgence of experiments in education has been an attempt to establish rigorous research in the field. That is, currently much of the empirical research in education employs randomized experiments that are typically large in scale. These field experiments allow education researchers to examine the effects of school, or student interventions on student performance. In addition, education experiments incorporate often times a longitudinal component where students are followed over time. The main objectives in these studies include assessing whether the treatment effects are cumulative or have lasting benefits or whether they fade over time. For example, the effect of a novel mathematics curriculum is evaluated through an experiment (i.e., novel versus traditional mathematics

curriculum) where measurements of student outcomes (e.g., mathematics achievement) are collected repeatedly over time (e.g., every spring for a few years).

In repeated measures experiments each student has their own trajectory which is a function of time and indicates the rate of change over time (Raudenbush & Bryk, 2002). The central goal in such studies is not only to estimate the treatment effect in the first year of the study (e.g., immediate effects), but also gauge longer term effects over time. For example, a researcher may be interested in the change or growth of mathematics achievement for students who use a novel mathematics curriculum vis-a-vis students who use a traditional mathematics curriculum. In this case, it is important for the researcher to compare trajectories of students who received the treatment (i.e., novel curriculum) versus those who did not receive the treatment (i.e., traditional curriculum).

The change in measurements over time does not always follow a linear trend. Instead, trajectories sometimes point to nonlinearities such as curvilinear trends. For example, Huttenlocher et al. (1991) studied how children's vocabulary is accelerated in early years. One way of defining trajectories of change is via polynomial functions (Raudenbush & Liu, 2001). The first degree polynomial indicates linear rate of change, the second degree polynomial indicates a quadratic rate of change, the third degree polynomial indicates a cubic rate of change and so forth. That is, treatment effects are estimated for linear rates or non-linear rates of change.

Studies about polynomial change may be viewed as having a nested structure. For example, measurements are nested within individuals and this nesting needs to be taken into account in the design phase of the study as well as in the statistical analysis phase. Prior work has utilized two-level models (e.g., measurements within students) for repeated

measurements designs (see Raudenbush & Bryk, 2002). In particular, the authors presented methods for power analysis of treatment effects in studies of polynomial change with one level of nesting. Power is a function of the magnitude of the treatment effect, the sample size of individuals, the duration of the study, and the frequency of measurements over time. Researchers should take into account all of these parameters in the design phase of the experiment to ensure that treatment effects will be detected.

Nonetheless, populations in education have frequently more complicated structures. For example, students are also nested within classes or schools and so forth. In addition, education interventions typically assign either schools or students randomly to treatment or control groups. For instance, students are assigned to small or regular classes within schools. Or schools are randomly assigned to an assessment program or not. It seems natural to extend methods for power analysis for tests of treatment effects in studies of polynomial change from two to three-levels. Consider for example, a nested structure where measurements are nested within students and students in turn are nested within schools. That is, the first level is repeated measurements, the second level is students, and the third level is schools. Spybrook et al. (2011) reported in the optimal deign manual formulae to calculate power for three-level polynomial change models without covariates.

This study extends previous methods by Raudenbush and Liu (2001) and Spybrook et al. (2011), and provide methods for power analysis of tests of treatment effects in studies of polynomial change with two levels of nesting (e.g., students and schools) where the treatment is either at the third level (e.g., school intervention) or at the second level (e.g., student intervention). In particular, I present first methods for power analysis for cluster randomized designs (CRD) where for instance schools are randomly assigned in a

treatment and a control group, students are nested within schools, and repeated measurements are nested within students. This design assumes that schools are sampled randomly from a larger population at the first stage and then students within schools are randomly sampled. That is, both schools and students are random effects. Within CRD I briefly present the unconditional model (i.e., no covariates at any level), and then I expand the model to include covariates in the second and third levels. Second we provide methods for power analysis for block randomized designs (BRD) where the treatment is at the second level (e.g., student intervention) and the third level units (e.g., schools) serve as blocks. For example, students are assigned to treatment and control conditions within schools. In this design both schools and students are also treated as random effects. In addition, we will discuss how study duration, sample size (number of third and second level units), and covariates influence power through two illustrative samples.

The Polynomial Change Model

A polynomial is an algebraic expression that contains more than one term and is described as a sum of terms of the same variable (e.g., time) in different powers (Kirk, 2012). For example, student achievement growth could be modeled through a polynomial equation of the third degree as

$$Y = \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3 a^3 + \varepsilon \qquad (3.1)$$

where Y is student achievement, a is a measure of time such as age at each time of measurement, β_0 is a constant, $\beta_1 a$ is a linear component, $\beta_2 a^2$ is a quadratic component,

$\beta_3 a^3$ is a cubic component, and ε is an error term. One disadvantage of equation (3.1) is that the trend components are highly correlated, which leads to multicollinearity. To resolve the dependency problem, one can utilize orthogonal polynomial contrast coefficients, which have been frequently used to fit trends of repeated measures. Equation (3.1) can then be constructed as

$$Y = \alpha_0 c_0 + \alpha_1 c_1 + \alpha_2 c_2 + \alpha_3 c_3 + u \tag{3.2}$$

where c_1, c_2, and c_3 are orthogonal polynomial coefficients that are independent with each other and thus enable researchers to independently test a null hypothesis for each of the three components (Kirk, 2012). Orthogonal polynomial coefficients have been used to fit trends since the early 20th century (e.g., Fisher, 1928). Jennrich and Sampson (1971) provided an algorithm to generate the orthogonal polynomial contrast coefficients. They are provided in tables of many experimental design texts (e.g., Kirk, 2012).

Previous work has discussed sample size and statistical power considerations for group comparisons using repeated measures, most of which however are focused on single-level models (e.g., Bloch, 1986; Hedeker, Gibbons, & Waternaux, 1999). Raudenbush and Liu (2001) extended this work and provided power analysis and sample determination methods for repeated measures in two-level models. They focused on studies in which two groups were followed over time to assess group differences in the average rate of change, rate of acceleration, or a higher degree polynomial effect. Through a two-level model combined with orthogonal polynomial contrasts at the first level, the authors examined how the duration of the study, frequency of observation, and number of participants affected

statistical power. They found that power increases as the study duration or the number of students increases. Meorbeek (2008) discussed how the costs of including more persons or taking more measurement influence powers in a two-level polynomial growth models and provided methods of comparing alternative design on the basis of their costs and sample size. She also took drop-out into consideration, and found that power decrease as the dropout increase, and thus increasing the study duration might have a negative effect on the power.

Power analysis methods for growth models with two levels of nesting have rarely been discussed in prior literature. One exception was by Jong, Moerbeek, and Van der Leeden (2010), who discussed power estimation methods for three-level growth models with linear rate of change only. They have demonstrated that power is influenced by intraclass correlation coefficients, level of randomization, sample size, covariates and drop-out rates. However, their methods could not be applied to models with higher order of change rates (e.g., quadratic rate of change). The optimal design manual by Spybrook et al. (2011) has provided power calculation formulae for three-level models in studies of polynomial change where the treatment is at the third level (e.g., schools), but has not incorporated the effects of covariates.

Both Randenbush and Liu (2001) and Spybrook et. al. (2011) discussed unconditional models that did not include any covariates at any level. However, prior studies have shown that covariates (e.g., students and school characteristics) could increase power significantly. Hedges and Hedberg (2007) documented that prior test scores and demographic covariates such as SES account for nearly one-third of the variance at the student level. Bloom, Richburg-Hayes, and Black (2007) found that controlling for baseline covariates could

improve the precision of CRD studies that examine the impact of school interventions. Konstantopoulos (2012) showed that covariates at different levels of the hierarchy potentially explain a considerable proportion of the variance at the corresponding levels, and centering of lower level covariates plays an important role in this (see also Snijders & Bosker, 1999).

Statistical Models

Design I: Treatment Assigned at Third Level (Cluster Design)

Unconditional Model

Consider a simple three-level growth design where level-3 units (e.g., clusters such as schools) are randomly assigned to treatment or control conditions (i.e., clusters are nested within treatment). The first level for change over time of level-2 unit *i* in cluster *j* can de expressed as a polynomial function, namely

$$Y_{gij} = \alpha_{0ij}c_{0g} + \alpha_{1ij}c_{1g} + \alpha_{2ij}c_{2g} + ... + \alpha_{(P-1)ij}c_{(P-1)g} + u_{gij} \quad (3.3)$$

where c_{pg} represent orthogonal polynomial contrasts of degree p (p = 0, 1, …, P-1) at measurement g (g = 1, …, G), α_{pij}'s represent the mean and the rates of change (linear, quadratic, cubic, etc.), and u_{gij} is the within level-2 unit random term with variance σ_e^2. When p = 0, $c_{0g} = 1$ and α_{0ij} represents the average outcome for level-2 unit *i* in level-3 unit *j*. When p = 1, c_{1g} is a linear contrast and α_{1ij} is the linear rate of change for level-2

unit *i* in level-3 unit *j*, and so forth. We work with orthogonal polynomial contrasts because they facilitate the computations of estimators and their standard errors, and simplify power analysis (see Raudenbush & Liu, 2001). The results apply to studies of any length and for polynomials of any degree (Kirk, 2012).

Orthogonal polynomial contrast coefficients should satisfy two conditions: The *p*th polynomial contrasts trend sum to zero, and the sum of the product of the *p*th and *p'*th polynomial contrasts is equal to zero (see Kirk, 2012)

$$\sum_{g=1}^{G} c_{pg} = 0$$
$$\sum_{g=1}^{G} c_{pg} c_{p'g} = 0.$$

(3.4)

Orthogonal polynomial coefficients that meet conditions shown in equation (3.4) are not unique because, any group of orthogonal polynomial coefficients denoted as $C_{pg} = k_p c_{pg}$, also meet these two conditions, where k_p could be any constant (see Appendix C for a detailed proof). With equally spaced time points, the following formulae could be used to calculate orthogonal polynomial coefficients

$$\begin{aligned} c_{pg} &= k_p C_{pg} \\ C_{0g} &= 1 \\ C_{1g} &= g - \sum_{g=1}^{G} \frac{g}{G} \\ C_{p+1,g} &= C_{1g} C_{pg} - \frac{p^2 (G^2 - p^2)}{4 \cdot (4p^2 - 1)} C_{p-1,g} \end{aligned} \qquad (3.5)$$

(see Jennrich & Sampson, 1971), where c_{pg} is one possible orthogonal polynomial coefficient of degree *p* at measurement *g* as defined before, and k_p could be any constant. That is, researchers could choose any k_p to get their own orthogonal polynomial contrast coefficients.

For example, when $k_0 = 1, \quad k_1 = 1, \quad k_2 = \frac{1}{2}, \text{ and } k_3 = \frac{1}{6}$, one can compute the first four orthogonal coefficients as

$$
\begin{aligned}
c_{0g} &= 1 \\
c_{1g} &= g - \sum_{g=1}^{G} \frac{g}{G} \\
c_{2g} &= \frac{1}{2}\left[\left(g - \sum_{g=1}^{G} \frac{g}{G} \right)^2 - \frac{G^2 - 1}{12} \right] \\
c_{3g} &= \frac{1}{6}\left[\left(g - \sum_{g=1}^{G} \frac{g}{G} \right)^3 - \frac{3G^2 - 7}{20} \cdot \left(g - \sum_{g=1}^{G} \frac{g}{G} \right) \right]
\end{aligned}
\tag{3.6}
$$

(see Appendix C for a detailed proof). When G = 4, then the values of the orthogonal coefficients are

$$
\begin{aligned}
c_0 &= (1,\ 1,\ 1,\ 1) \\
c_1 &= (-1.5,\ -0.5,\ 0.5,\ 1.5) \\
c_2 &= (0.5,\ -0.5,\ -0.5,\ 0.5) \\
c_3 &= (-0.05,\ 0.15,\ -0.15,\ 0.05).
\end{aligned}
\tag{3.7}
$$

Least squares estimates of each level-2 unit's change parameter as well as their variance can be computed as

$$\hat{\alpha}_{pij} = \frac{\sum_{g=1}^{G} c_{pg} Y_{gij}}{\sum_{g=1}^{G} c^2_{pg}}$$
$$Var(\hat{\alpha}_{pij}) = \frac{\sigma_e^2}{\sum_{g=1}^{G} c^2_{pg}} \qquad (3.8)$$

(see Seber & Lee, 2003), where

$$\sum_{g=1}^{G} c^2_{pg} = k_p^2 \cdot \frac{(p!)^4}{(2p)! \cdot (2p+1)!} \cdot \frac{(G+p)!}{(G-p-1)!} \qquad (3.9)$$

(see Appendix D for proof).

In the second level model each of the parameters α_{pij} (e.g., the average polynomial change for each individual) from the first level equation varies between level-2 units (e.g., individuals) within level-3 units (e.g., schools), namely

$$\alpha_{pij} = \beta_{p0j} + \xi_{pij}, \qquad (3.10)$$

where β_{p0j}'s represent the average polynomial effects within level-3 units such as schools and the ξ_{pij}'s are individual specific random effects within level-3 units for each

polynomial change parameter. The random effects follow a multivariate normal distribution with zero means, variances τ_{pp}^2, and covariance $\tau_{pp'}$ between the random effects ξ_{pij} and $\xi_{p'ij}$.

At the third level each of the parameters, β_{p0j}'s (average polynomial change for each level-3 unit) vary across third level units such as schools, namely

$$\beta_{p0j} = \gamma_{p00} + \gamma_{p01} T_j + \eta_{p0j}, \tag{3.11}$$

where γ_{p00}'s represent the average polynomial effects across level-3 units, γ_{p01}'s represent the average difference between the treatment and the control group for each polynomial change parameter, and the η_{p0j}'s are level-3 unit specific random effects for each polynomial change parameter. These random effects follow a multivariate normal distribution with zero means, variances ω_{pp}^2, and covariance $\omega_{pp'}$ between the random effects η_{p0j} and $\eta_{p'0j}$.

Suppose there are *N* level-2 units within each level-3 unit and *m* level-3 units within each treatment condition, which means that the total number of level-3 units is $M = 2m$ and thus the total number of level-2 units is *MN*. Then, the estimate of the variance of the treatment effect for polynomial *p* is

$$Var(\gamma_{p01}) = \frac{2}{mN}(N\omega_{pp}^2 + \tau_{pp}^2 + \sigma_p^2),\ \sigma_p^2 = \frac{\sigma_e^2}{\sum_{g=1}^{G} c_{pg}^2} \tag{3.12}$$

and $\sum_{g=1}^{G} c_{pg}^2$ is defined in equation (3.9) (see Konstantopoulos, 2008a; Raudenbush & Liu, 2001; Spybrook et al., 2011).

Suppose that a researcher wants to test the hypothesis that γ_{p01} is different from zero and carries out the usual *t*-test. The test statistic is defined as

$$t = \hat{\gamma}_{p01} / \sqrt{Var(\hat{\gamma}_{p01})} \,. \tag{3.13}$$

When the null hypothesis is true, the test statistic *t* has a Student's *t*-distribution with *2m-2* degrees of freedom. When the null hypothesis is false, the test statistic *t* has the non-central *t*-distribution with *2m-2* degrees of freedom and non-centrality parameter λ. The non-centrality parameter is defined as the expected value of the estimate of the treatment effect divided by the square root of the variance of the estimate of the treatment effect, namely

$$\lambda = \gamma_{p01} \sqrt{\frac{mN}{2}} \sqrt{\frac{1}{(N\omega_{pp}^2 + \tau_{pp}^2 + \sigma_p^2)}} \,. \tag{3.14}$$

To calculate power, we need to define a standardized effect size first. Prior literature provided three definitions of standard effect size for three level models (e.g., Hedge, 2010; Konstantopoulos, 2008a, 2008b). The first option of defining the standardized effect size for a polynomial degree *p* in three-level models is the group differences divided by the

square root of the total variance (Hedges, 2010; Jong, Moerbeek, & Van der Leeden, 2010; Konstantopoulos, 2008a)

$$ES_1 = \frac{\gamma_{p01}}{\sqrt{\omega_{pp}^2 + \tau_{pp}^2 + \sigma_p^2}}. \tag{3.15}$$

However, the denomination of ES_1 depends on σ_p^2, which is a function of the study duration as shown in equation (3.8). In other words, ES_1 changes as the study duration varies. Because this study evaluates various designs with alternative study duration but with fixed effect size, ES_1 is not appropriate.

Another two ways of defining the standardized effects size are

$$ES = \frac{\gamma_{p01}}{\sqrt{\omega_{pp}^2 + \tau_{pp}^2}} \text{ or } ES_2 = \frac{\gamma_{p01}}{\omega_{pp}}, \tag{3.16}$$

where *ES* is the group differences divided by the square root the sum of level-2 variance and level-3 variance (Jong, Moerbeek, & Van der Leeden, 2010; Spybrook et. al., 2011); while ES_2 is the group differences divided by the square root level-3 variance. Both *ES* and ES_2 could be used as the standardized effect size in three-level models (see Hedges, 2011). It should be noted that ES_2 is larger than *ES* for the same model if $\tau_{pp}^2 > 0$, especially when the level-2 variance account for a large proportion of the total variance. For example, in our illustrative example using data from Project STAR in a later section, the effect size was larger than one if ES_2 is used; however the effect size from Project STAR was about 0.2

using Cohen's *d*. Cohen (1988) suggested that 0.2 is considered as a small effect size, 0.5 is considered as a medium effect size, and 0.8 is considered as a large effect size. Therefore, small or medium effect size might be interpreted as large effect size without cautiousness if ES_2 is used. In order to avoid assuming a large standardized effect and keep consistent with Cohen's definition of small, medium and large effect size, I use *ES* as the definition of standardized effect size in this study. Note that researchers still need to be cautious to interpret *ES*, which trends to be larger than ES_1 since it does not take the variance at the first level into consideration.

Then, the non-centrality parameter λ of the *t*-test in equation (3.14) simplifies to

$$\lambda = \sqrt{\frac{mN}{2}} ES \sqrt{\frac{\omega_{pp}^2 + \tau_{pp}^2}{N\omega_{pp}^2 + \tau_{pp}^2 + \sigma_p^2}}. \qquad (3.17)$$

The power of a two-tailed *t*-test for a specified significance level α is defined as

$$p_1 = 1 - \mathrm{H}\,[c(\alpha/2,\, 2m\text{-}2),\, (2m\text{-}2),\, \lambda] + \mathrm{H}\,[\text{-}c(\alpha/2,\, 2m\text{-}2),\, (2m\text{-}2),\, \lambda] \qquad (3.18)$$

where $c(\alpha, v)$ is the level *a* one-tailed critical value of the t-distribution with *v* degrees of freedom (e.g., $c(0.05,20)=1.72$), and $\mathrm{H}(x, v, \lambda)$ is the cumulative distribution function of the non-central t-distribution with *v* degrees of freedom and non-centrality parameter λ. Alternatively, one can use an *F*-test with 1, $2m - 2$ degrees of freedom and a non-centrality parameter λ^2.

Covariates at Second and Third Levels

When covariates are included at the second level equation (3.10) becomes

$$\alpha_{pij} = \beta_{p0j} + \mathbf{X_{ij}B_{p2j}} + \xi_{Apij}, \tag{3.19}$$

where $\mathbf{X_{ij}}$ is a row vector of *k* level-2 unit characteristics, and $\mathbf{B_{p2j}}$ is a row vector of *k* coefficients of level-2 unit characteristics. The ξ_{Apij}'s are level-2 specific random effects within level-3 units for each polynomial change parameter, and subscript *A* indicates adjustment in the error term because of covariates. The random effects follow a multivariate normal distribution with zero means, variances τ^2_{Rpp}, covariance $\tau_{Rpp'}$ between random effects ξ_{pij} and $\xi_{p'ij}$, and subscript *R* indicates residual variance because of covariates. All other terms have been defined previously.

Similarly, the third level model of equation (3.11) becomes

$$\beta_{p0j} = \gamma_{p00} + \gamma_{Ap01}T_j + \mathbf{Z_{P02}\Gamma_j} + \eta_{Ap0j} \tag{3.20}$$

where $\mathbf{Z_{P02}}$ is a row vector of *q* level-3 unit characteristics, and $\mathbf{\Gamma_j}$ is a column vector of coefficients of level-3 unit characteristics. The η_{Ap0j}'s are level-3 specific random effects for each polynomial change parameter, where subscript *A* indicates adjustment because of covariates (see Konstantopoulos, 2008a). These random effects follow a multivariate normal distribution with zero means, variances ω^2_{Rpp}, covariance $\omega_{Rpp'}$ between random

effects η_{p0j} and $\eta_{p'0j}$, and subscript *R* indicates residual variance because of covariates. All other terms have been defined previously.

As a result, the non-centrality parameter of the *t*-test for the three-level model with covariates at second and third levels is defined as

$$\lambda_A = \gamma_{Ap01}\sqrt{\frac{mN}{2}}\sqrt{\frac{1}{Nw_3\omega_{pp}^2 + w_2\tau_{pp}^2 + \sigma_p^2}}, \tag{3.21}$$

where

$$w_3 = \omega_{Rpp}^2 / \omega_{pp}^2, w_2 = \tau_{Rpp}^2 / \tau_{pp}^2, \tag{3.22}$$

that is, w_2 indicates the proportion of the variance at the second level that is still unexplained; while w_3 indicates the proportion of the variance at the third level that is still unexplained. For example, when w_3 = 0.8, it indicates that the variance at the third level decreased by 20% because of inclusion of covariates at the third level (assuming a centering approach where covariates can explain variance in the outcome only at their corresponding levels). In other words, the covariates at the third level explain 20% of the variance at the third level.

We assume that the coefficient of the treatment does not change after adding covariates at the second and third level ($\gamma_{Ap01} = \gamma_{p01}$), which is reasonable since in experimental

designs the treatment (T_j) should be independent of any covariates (observed or unobserved). Then the non-centrality parameter λ_A in equation (3.21) simplifies to

$$\lambda_A = \sqrt{\frac{mN}{2}} ES \sqrt{\frac{\omega_{pp}^2 + \tau_{pp}^2}{N w_3 \omega_{pp}^2 + w_2 \tau_{pp}^2 + \sigma_p^2}} . \quad (3.23)$$

The power of a two-tailed *t*-test for a specified significance level α is defined as

$$p_2 = 1 - \mathrm{H}\,[c(\alpha/2,\, 2m\text{-}q\text{-}2),\, (2m\text{-}q\text{-}2),\, \lambda_A] + \mathrm{H}\,[-c(\alpha/2,\, 2m\text{-}q\text{-}2),\, (2m\text{-}q\text{-}2),\, \lambda_A] \quad (3.24)$$

where *q* is the number of covariates at the third level. As mentioned previously, an F-test could be used instead.

Design II: Treatment Assigned at Second Level (Block Randomized Design)

Unconditional Model

The first level model is identical to equation (3.3). The second level model incorporates the treatment (T_{ij}), namely

$$\alpha_{pij} = \beta_{p0j} + \beta_{p1j} T_{ij} + \xi_{pij} , \quad (3.25)$$

where β_{p0j}'s represent the average polynomial effects within level-3 units, T_{ij} is a dummy variable coded as one if second level unit *i* in third level unit *j* is assigned to treatment or control conditions and zero otherwise, β_{p1j} is the treatment effect within level-3 units, and

the ξ_{pij} 's are level-2 random effects within level-3 units for each polynomial change parameter. The random effects follow a multivariate normal distribution with zero means, variances τ_{pp}^2, and covariance $\tau_{pp'}$ between random effects ξ_{pij} and $\xi_{p'ij}$.

The third level equations for the intercept (β_{p0j}) and the treatment effect (β_{p1j}) are

$$\begin{aligned} \beta_{p0j} &= \gamma_{p00} + \eta_{p0j} \\ \beta_{p1j} &= \gamma_{p10} + \eta_{p1j} \end{aligned}, \tag{3.26}$$

where γ_{p00} 's represent the average polynomial effects across level-3 units, the η_{p0j} 's are level-3 unit specific random effects for each polynomial change parameter, γ_{p10} 's represent the average difference between the treatment and the control groups for each polynomial change parameter across level-3 units, and the η_{p1j} 's are treatment by level-3 unit random effects (interaction effects) for each polynomial change parameter. The η_{p0j} 's follow a multivariate normal distribution with zero means and variances ω_{pp}^2, whilst the treatment by level-3 unit random effects also follow a normal distribution with a mean of zero and a variance ω_{Tpp}^2, where subscript *T* indicates treatment at the second level whose effect varies at the third level.

Suppose there are *M* level-3 units and *n* level-2 units within each treatment condition within each level-3 unit, which means that the total number of level-2 units in each level-3 unit is $N = 2n$ and thus the total number of level-2 units is *MN*. Then, the estimate of the variance of the treatment effect for polynomial *p* is

$$Var(\gamma_{p10}) = \frac{2}{Mn}(n\omega_{Tpp}^2 + \tau_{pp}^2 + \sigma_p^2),\ \sigma_p^2 = \frac{\sigma_e^2}{\sum_{g=1}^{G} c_{pg}^2} \tag{3.27}$$

where subscript *T* indicates treatment at the second level whose effect varies at the third level.

and $\sum_{g=1}^{G} c_{pg}^2$ is defined in equation (3.9).

Suppose that a researcher wants to test the hypothesis that γ_{p10} is different from zero and carries out a *t*-test. The test statistic is defined as

$$t = \hat{\gamma}_{p10} / \sqrt{Var(\hat{\gamma}_{p10})}. \tag{3.28}$$

When the null hypothesis is true, the test statistic *t* has a Student's *t*-distribution with *M-1* degrees of freedom (Konstantopoulos, 2008b). When the null hypothesis is false, the test statistic *t* has the non-central *t*-distribution with *M-1* degrees of freedom and non-centrality parameter λ. The non-centrality parameter is defined as the expected value of the estimate of the treatment effect divided by the square root of the variance of the estimate of the treatment effect, namely

$$\lambda = \gamma_{p10}\sqrt{\frac{Mn}{2}}\sqrt{\frac{1}{(n\omega_{Tpp}^2 + \tau_{pp}^2 + \sigma_p^2)}}. \tag{3.29}$$

We define the standardized effect size for a polynomial degree *p* as

$$ES = \frac{\gamma_{p10}}{\sqrt{\omega_{Tpp}^2 + \omega_{pp}^2}}. \tag{3.30}$$

Then, the non-centrality parameter λ of the t-test simplifies to

$$\lambda = \sqrt{\frac{Mn}{2}} ES \sqrt{\frac{\omega_{Tpp}^2 + \tau_{pp}^2}{n\omega_{Tpp}^2 + \tau_{pp}^2 + \sigma_p^2}}. \tag{3.31}$$

The power of a two-tailed *t*-test for a specified significance level α is defined as

$$p_3 = 1 - \mathrm{H}\,[c(\alpha/2, M\text{-}1), (M\text{-}1), \lambda] + \mathrm{H}\,[-c(\alpha/2, M\text{-}1), (M\text{-}1), \lambda] \tag{3.32}$$

where $c(\alpha,v)$ is the level *a* one-tailed critical value of the t-distribution with *v* degrees of freedom, and $\mathrm{H}(x, v, \lambda)$ is the cumulative distribution function of the non-central *t*-distribution with *v* degrees of freedom and non-centrality parameter λ. As noted previously one could use an F-test instead.

Covariates at Second and Third Levels

When covariates are included at the second level equation (3.25) becomes

$$\alpha_{pij} = \beta_{p0j} + \beta_{Ap1j} T_{ij} + \mathbf{X_{ij}} \mathbf{B_{p2j}} + \xi_{Apij}, \tag{3.33}$$

where $\mathbf{X_{ij}}$ is a row vector of *k* level-2 unit background characteristics, and $\mathbf{B_{p2j}}$ is a row vector of *k* coefficients of level-2 unit characteristics. The ξ_{Apij} 's are level-2 specific random effects within level-3 units for each polynomial change parameter, where subscript *A* indicates adjustment because of covariates. The random effects follow a multivariate normal distribution with zero means, variances τ^2_{Rpp}, and covariance $\tau_{Rpp'}$ between random effects ξ_{pij} and $\xi_{p'ij}$. The subscript *R* indicates residual variance because of covariates. All other terms have been defined previously.

When covariates are included at the third level equation (3.26) becomes

$$\begin{aligned}\beta_{p0j} &= \gamma_{p00} + \mathbf{Z_{P1}\Gamma_{p0j}} + \eta_{Ap0j} \\ \beta_{Ap1j} &= \gamma_{Ap10} + \mathbf{Z_{P1}\Gamma_{p1j}} + \eta_{Ap1j},\end{aligned} \tag{3.34}$$

where $\mathbf{Z_{P1}}$ is a row vector of *q* level-3 unit characteristics and the Γ 's include regression coefficients. The η_{Ap0j} 's are level-3 unit specific random effects for each polynomial change parameter, and the η_{Ap1j} 's are treatment by level-3 unit random effects (interaction effects) for each polynomial change parameter. The η_{Ap0j}'s follow a multivariate normal distribution with zero means, variances ω^2_{Rpp}, and the treatment by level-3 unit random effects also follows a normal distribution with a mean of zero and a variance ω^2_{RTpp}, subscript *R* indicates residual variance because of covariates. The non-centrality parameter of the *t*-test when covariates are added at the second and third levels is defined as

$$\lambda_A = \gamma_{p10}\sqrt{\frac{Mn}{2}}\sqrt{\frac{1}{(nw_3\omega^2_{Tpp} + w_2\tau^2_{pp} + \sigma^2_p)}} \quad (3.35)$$

where subscript *A* indicates adjustment because of covariates (see Konstantopoulos, 2008b) and

$$w_3 = \omega^2_{RTpp} / \omega^2_{Tpp}, w_2 = \tau^2_{Rpp} / \tau^2_{pp}, \quad (3.36)$$

that is, w_2 indicates the proportion of the variance at the second level that is still unexplained, and w_3 indicates the proportion of the treatment by level-3 unit variance at the third level that is still unexplained. We assume the coefficient of the treatment does not change after adding covariates at the second and third level ($\gamma_{Ap10} = \gamma_{p10}$), which is reasonable since in experimental designs the treatment (T_{ij}) should be independent of any covariates (observed or unobserved). Then the non-centrality parameter λ_A of the t-test in equation (3.35) simplifies to

$$\lambda_A = \sqrt{\frac{Mn}{2}}ES\sqrt{\frac{\omega^2_{Tpp} + \tau^2_{pp}}{nw_3\omega^2_{Tpp} + w_2\tau^2_{pp} + \sigma^2_p}}. \quad (3.37)$$

The power of a two-tailed *t*-test for a specified significance level α is defined as

$$p_4 = 1 - \mathrm{H}\,[c(\alpha/2, M\text{-}q\text{-}1), (M\text{-}q\text{-}1), \lambda_A] + \mathrm{H}\,[\text{-}c(\alpha/2, M\text{-}q\text{-}1), (M\text{-}q\text{-}1), \lambda_A], \quad (3.38)$$

where q is the number of covariates at the third level. As mentioned previously one could use an F-test instead.

Illustrative Examples

Cluster Randomized Design: A Linear Growth Model

To illustrate the applicability of the methods to assess consequences of study duration, sample sizes (students and schools), and covariates on power, we firstly utilized the data from a large scale experiment that was conducted in Indiana. This experiment employed a CRD, where students were nested within schools, and schools were nested within treatment and control groups. Random assignment took place at the school level, that is, schools were randomly assigned to treatment and control conditions. Schools in the treatment group adopted specific diagnostic assessment tools to measure student learning a few times during the 2009-2010 school year and to provide diagnostic information to teachers to improve ongoing instruction. The study incorporated a longitudinal component and thus student mathematics and reading achievement were measured three times in the spring of 2010, 2011, and 2012 (see Konstantopoulos, Miller, & van der Ploeg, 2013 for a more detailed introduction on this experiment). The total number of participating schools was 50 with 32 schools in the treatment group. Overall, nearly 20,000 students participated in the study during the 2009-2010 school year.

The outcome is standardized student mathematics achievement. Because the study duration was only 3 years, we used a linear rate of change model at level-1 (repeated measures), namely

$$Math_{gij} = \alpha_{0ij} c_{0g} + \alpha_{1ij} c_{1g} + u_{gij},\ u_{gij} \sim N(0,\ \sigma_e^2),$$

where $Math_{gij}$ is student mathematics achievement in year g, $c_{0g} = (1,\ 1,\ 1)$ and $c_{1g} = (-1,\ 0,\ 1)$ at g = 1, 2, 3 in accord with the orthogonal polynomials in equation (3.6). This model defines α_{0ij} as the mean mathematics achievement for student *i* in cluster *j*, and α_{1ij} is the average rate of linear change of mathematics achievement for student *i* in school *j*.

The second level model (student level) is

$$\alpha_{0ij} = \beta_{00j} + \xi_{0ij},\ \xi_{0ij} \sim N(0,\ \tau_{00}^2)$$

$$\alpha_{1ij} = \beta_{10j} + \xi_{1ij},\ \xi_{1ij} \sim N(0,\ \tau_{11}^2),$$

where β_{00j} is the mean mathematics achievement in school *j*, and β_{10j} is the average growth rate in school *j*.

The third level model (school level) is

$$\beta_{00j} = \gamma_{000} + \gamma_{001} T_j + \eta_{00j},\ \eta_{00j} \sim N(0,\ \omega_{00}^2)$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101} T_j + \eta_{10j},\ \eta_{10j} \sim N(0,\ \omega_{11}^2),$$

where γ_{000} is the grand mean, γ_{001} is the main effect of treatment for the mean, T_j is a binary indicator coded as one for treatment schools and zero for control schools, γ_{100} is the average rate of change, and γ_{101} is the main effect of treatment for the rate of change, which is my primary interest. We estimates of the relevant variances are

$$\sigma_e^2 = 0.00092,\ \tau_{11}^2 = 0.00091,\ \omega_{11}^2 = 0.00012\,.$$

To calculate power, we assumed a standardized effect size of 0.40 and a significance level of 0.05. We also assumed the sample size as m = 10 and N = 20, which indicates 10 schools in the treatment group (20 schools in total) and 20 students in each treatment or control school.

According to equation (3.8) and equation (3.9) with G = 3, p =1 and k_1 = 1, first I calculate

$$\sigma_1^2 = \frac{12 \cdot 0.00092}{4 \cdot 3 \cdot 2} = 0.00046\,.$$

Then, I calculate the non-centrality parameter of the t-test based on equation (3.17), namely

$$\lambda = \sqrt{\frac{mN}{2}} ES \sqrt{\frac{\omega_{pp}^2 + \tau_{pp}^2}{N\omega_{pp}^2 + \tau_{pp}^2 + \sigma_p^2}} = \sqrt{\frac{10 \cdot 20}{2}} \cdot 0.4 \cdot \sqrt{\frac{0.00012 + 0.00091}{300 \cdot 0.00012 + 0.00091 + 0.00046}} \approx 2.090\,.$$

Then, I compute the critical value of the test using the *t*-distribution with (2×10) - 2 = 18 degrees of freedom as $c(0.25, 48) \approx 2.101$. To compute power I use equation (3.18) as

$$p = 1 - \mathrm{H}\,[2.101, 18, 2.090] + \mathrm{H}\,[-2.101, 18, 2.090] \approx 0.508.$$

Tables 3.1 to 3.3 and Figure 3.1 to 3.3 show how variations of study duration and sample sizes affect power to detect the treatment effect for the linear rate of change in cluster designs, assuming two-tailed *t*-tests at the 0.05 significance level and effect size as 0.40. Table 3.1 and Figure 3.1 provide power estimates for designs that vary the study duration (*D*) and the number of schools (*M*), holding the number of students (*N*) in each school constant at 20. The estimate of power from above was 0.508 (see Table 3.1, row 2, column 2). As the study duration or number of schools increase, power increases. When the study duration is three and the number of schools is 40, power reaches to 0.80 (i.e., 0.822). Note that, power increases significantly as study duration increases from two to three, but then power only changes marginally. This suggests that for a fixed number of students, increasing the study duration beyond a certain point has only a small effect on powers. In addition, the number of schools has bigger effects on powers compared to the study duration. For example, when study duration is tripled from two to six, powers are less than doubled; while number of schools tripled from 10 to 30, powers are more than doubled.

Table 3.2 and Figure 3.2 provide power estimates for designs that vary the duration of study (*D*) and the number of students (*N*) in each school, holding the number of schools (*M*) constant at 20. As the study duration or the number of students grows, power becomes larger. In particular, power changes significantly when the study duration increases from two to three, and then powers does not change much as the study duration becomes longer. Similarly, increasing the number of students increases power to a specific number of students per school and beyond that number power does not change much. It is noteworthy that increasing the number of students is not an effective way of boosting power. For

Table 3.1: Effect of Study Duration (*D*) and Number of Schools (*M*) on Power Holding Number of Students (*N*) in Each School Constant at 20: CRD, Linear Rate of Change

D	*M*									
	10	20	30	40	50	60	70	80	90	100
2	0.201	0.395	0.562	0.693	0.791	0.861	0.910	0.942	0.964	0.977
3	0.257	0.508	0.696	0.822	0.900	0.945	0.971	0.985	0.992	0.996
4	0.273	0.538	0.728	0.849	0.920	0.959	0.980	0.990	0.995	0.998
5	0.279	0.549	0.740	0.858	0.926	0.963	0.982	0.992	0.996	0.998
6	0.282	0.554	0.745	0.862	0.929	0.965	0.983	0.992	0.996	0.998
7	0.283	0.556	0.747	0.864	0.931	0.966	0.984	0.992	0.997	0.998
8	0.284	0.558	0.748	0.865	0.931	0.966	0.984	0.993	0.997	0.998

Note. Effect size is 0.4 with a significance level of 0.05.

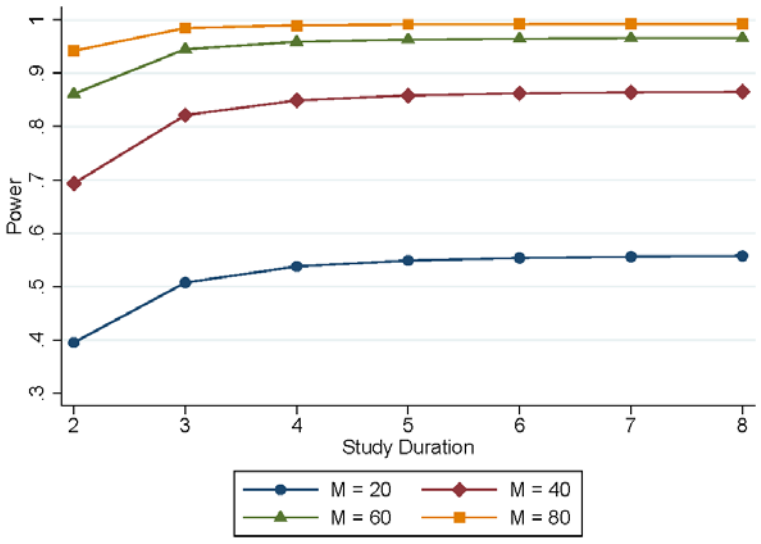

Figure 3.1: Effect of Study Duration (*D*) and Number of Schools (*M*) on Power, Holding Number of Students (*N*) in Each School Constant at 20: CRD, Linear Rate of Change

Note. Effect size is 0.4 with a significance level of 0.05

Table 3.2: Effect of Study Duration (*D*) and Number of Students (*N*) on Power Holding Number of Schools (*M*) Constant at 20: CRD, Linear Rate of Change

D	*N*									
	10	20	30	40	50	60	70	80	90	100
2	0.277	0.395	0.463	0.507	0.537	0.559	0.576	0.589	0.600	0.609
3	0.396	0.508	0.560	0.590	0.609	0.623	0.633	0.640	0.646	0.651
4	0.434	0.538	0.584	0.609	0.626	0.637	0.645	0.651	0.656	0.660
5	0.449	0.549	0.592	0.616	0.631	0.642	0.649	0.655	0.660	0.663
6	0.455	0.554	0.596	0.619	0.634	0.644	0.651	0.657	0.661	0.665
7	0.459	0.556	0.598	0.620	0.635	0.645	0.652	0.658	0.662	0.665
8	0.461	0.558	0.599	0.621	0.636	0.645	0.653	0.658	0.662	0.666

Note. Effect size is 0.4 with a significance level of 0.05.

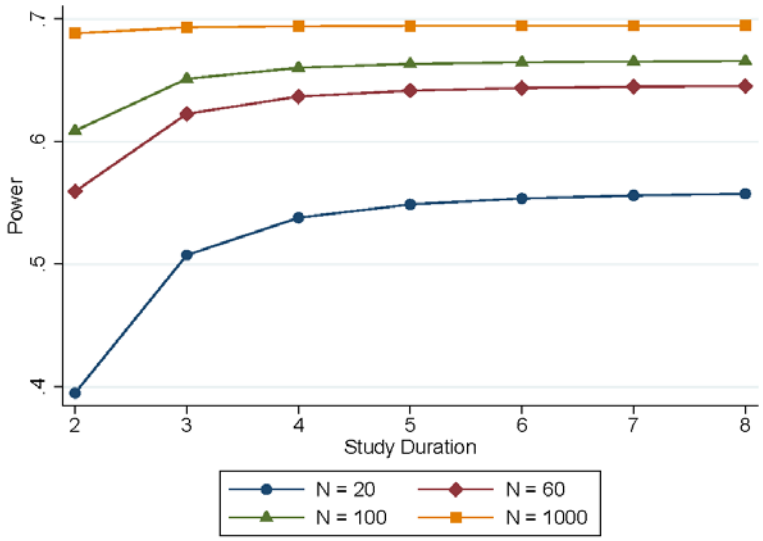

Figure 3.2: Effect of Study Duration (*D*) and Number of Students (*N*) on Power Holding Number of Schools (*M*) Constant at 20: CRD, Linear Rate of Change

Note. Effect size is 0.4 with a significance level of 0.05.

Table 3.3: Effects of Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 3: CRD, Linear Rate of Change

M	*N*									
	10	20	30	40	50	60	70	80	90	100
10	0.201	0.257	0.285	0.302	0.314	0.322	0.328	0.333	0.337	0.340
20	0.396	0.508	0.560	0.590	0.609	0.623	0.633	0.640	0.646	0.651
30	0.563	0.696	0.751	0.780	0.798	0.810	0.819	0.826	0.831	0.835
40	0.694	0.822	0.867	0.890	0.903	0.911	0.917	0.922	0.925	0.928
50	0.792	0.900	0.933	0.947	0.956	0.961	0.964	0.967	0.969	0.970
60	0.862	0.945	0.967	0.976	0.981	0.983	0.985	0.987	0.988	0.988
70	0.910	0.971	0.984	0.989	0.992	0.993	0.994	0.995	0.995	0.996
80	0.943	0.985	0.993	0.995	0.997	0.997	0.998	0.998	0.998	0.998
90	0.964	0.992	0.997	0.998	0.999	0.999	0.999	0.999	0.999	0.999
100	0.977	0.996	0.999	0.999	0.999	1.000	1.000	1.000	1.000	1.000

Note. Effect size is 0.4 with a significance level of 0.05.

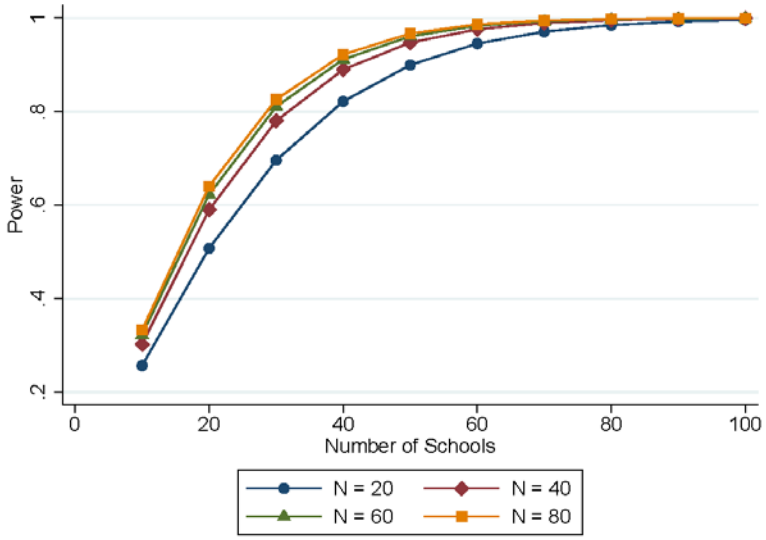

Figure 3.3: Effects of Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 3: CRD, Linear Rate of Change

Note. Effect size is 0.4 with a significance level of 0.05.

example, as shown in Figure 3.2, power is still less than 0.70 even if the number of students per school reaches to 1000.

Table 3.3 and Figure 3.3 provides power estimates for designs that vary the number of students (N) in each school and the number of schools (M), holding study duration constant at three. As the number of students per school or the number of schools increases, power increases initially and then does not change much. Power reaches to 0.80 with various combinations of the number of schools and the number of students per schools (e.g., M = 30 and N = 60, M = 40 and N = 20, and M = 60 and N = 10). It also should be noted that the number of schools affects power more significantly than the number of students in each school, holding the study duration fixed. For example, power is at least about tripled when the number of schools increases from ten to 100; while power less than doubled when the number of students increases from ten to 100.

Covariates also influence powers assuming they explain a certain proportion of variances at the second or the third level. Table 3.4 and Figure 3.4 shows how power varies as the proportion of the unexplained variances at the second and third levels vary for a design with M = 20 (or m = 10), N = 20, D = 3, and ES = 0.40. The degrees of freedom decrease when I add covariates at the third level. Assuming that five covariates are added at the third (q = 5), the degrees of freedom reduce to (2×10) - 5 - 2 = 13. As the unexplained variance decreases because of covariates, power increases. For example, when w_3 = 0.9 and w_2 = 0.9, which indicates the proportion of the unexplained variance at the second and the third level are 90% (or the covariates explain 10% of the variances at the second and the third level), the power is 0.526, which is larger than the power without covariates (0.508). In addition, covariates at the third level affect power significantly more than covariates at

Table 3.4: Effect of Covariates on Power: CRD, Linear Rate of Change

W3	*W2*								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.988	0.979	0.967	0.953	0.937	0.919	0.900	0.881	0.861
0.2	0.958	0.943	0.925	0.907	0.888	0.868	0.848	0.828	0.808
0.3	0.914	0.895	0.875	0.855	0.835	0.815	0.795	0.775	0.756
0.4	0.862	0.842	0.822	0.802	0.782	0.763	0.744	0.726	0.708
0.5	0.809	0.790	0.770	0.751	0.733	0.715	0.697	0.681	0.665
0.6	0.758	0.739	0.721	0.704	0.687	0.670	0.655	0.639	0.625
0.7	0.710	0.693	0.676	0.660	0.645	0.630	0.616	0.602	0.589
0.8	0.666	0.650	0.635	0.621	0.607	0.593	0.580	0.568	0.556
0.9	0.626	0.612	0.598	0.585	0.572	0.560	0.549	0.537	0.526

Note. The study duration is 3 with 20 schools and 20 students in each school; significance level is 0.05.

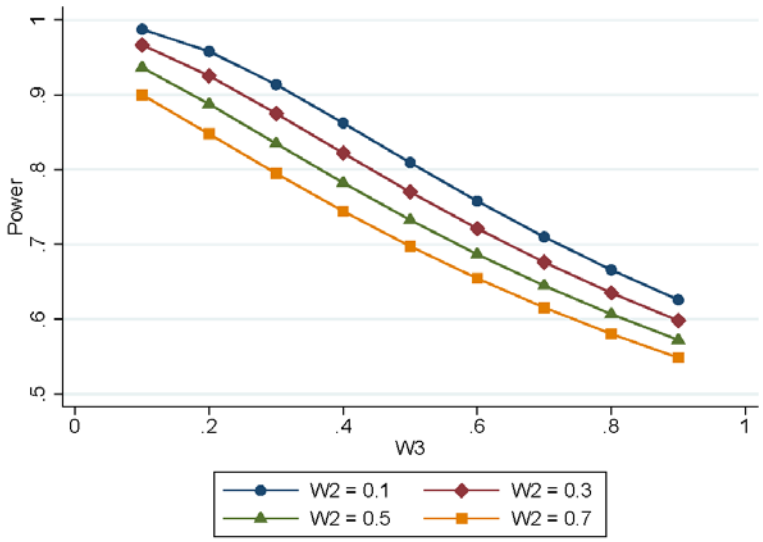

Figure 3.4: Effect of Covariates on Power: CRD, Linear Rate of Change

Note. The study duration is 3 with 20 schools and 20 students in each school; significance level is 0.05.

the second level, which is mainly because the ratio between variance of level-2 random effect and variance of level-3 random effect ($\tau_{11}^2 / \omega_{11}^2$) is small. For example, as the proportion of the unexplained variances at the second level (w_2) decreases from 0.9 to 0.1 with $w_3 = 0.9$, power increases slightly from 0.526 to 0.626. However, as the proportion of the unexplained variance at the third level (w_3) decreases from 0.9 to 0.1 with $w_2 = 0.9$, power increases significantly from 0.526 to 0.861.

To compare the powers between design with and without covariates, I also compute power estimates for designs that vary the number of students (N) in each school and the number of schools (M), assuming 40% of variances explained at the second and the third level ($w_2 = w_3 = 0.6$), holding study duration constant at three, which are presented by Table 3.5 and Figure 3.5. In general, power increases when covariates explain a certain proportion of variance at the second or the third level, comparing the power estimates in Table 3.3. There are only three exceptions (i.e., $M = 10$ and $N = 10$, $M = 10$ and $N = 20$, and $M = 10$ and $N = 30$), where power decreases when covariates were added. That is because degrees of freedom decreases as I assume five covariates added at the third level.

Block Randomized Design: A Linear Growth Model

The second example utilized data from Project STAR (Student-Teacher Achievement Ratio) in Tennessee (e.g., Finn & Achilles, 1990; Krueger, 1999; Nye, Hedges, & Konstantopoulos, 2000). This experiment employed a block randomized design, where within each school (the block) and grade, students and their teachers were randomly assigned to one of three treatment conditions: small classes (13.17 students), regular-size classes (22-25 students), and regular classes with a full-time teacher aide (22-25 students). Project STAR was a longitudinal study that started in the 1985-1986 school year. The

cohort of students who entered kindergarten in the 1985-1986 school year remained in the experiment until their third grade. Students' mathematics and reading achievement were measured four times in the end of kindergarten, first grade, second grade, and third grade. Overall, more than 11,000 students in 79 schools participated in the experiment over the four-year period.

The sample included students in small classes or regular classes only to ensure a balanced design. Students in regular classes with a full-time teacher aide were excluded from the analysis. The outcome is standardized student mathematics achievement. A linear rate of change was used at level-1 (repeated measures), namely

$$Math_{gij} = \alpha_{0ij} c_{0g} + \alpha_{1ij} c_{1g} + u_{gij},\ u_{gij} \sim N(0,\ \sigma_e^{\ 2}),$$

where $Math_{gij}$ is student mathematics achievement in year g, $c_{0g} = (1,\ 1,\ 1,\ 1)$ and $c_{1g} = (\text{-}1.5,\ -0.5,\ 0.5,\ 1.5)$ at g = 1, 2, 3, 4 following equation (3.6). This model defines α_{0ij} as the mean mathematics achievement for student *i* in school *j*, and α_{1ij} is the average linear rate of change of mathematics achievement for student *i* in school *j*.

The second level model (student level) is

$$\alpha_{0ij} = \beta_{00\,j} + \beta_{01\,j} \cdot T_{ij} + \xi_{0ij},\ \xi_{0ij} \sim N(0,\ \tau_{00}^2)$$
$$\alpha_{1ij} = \beta_{10\,j} + \beta_{11\,j} \cdot T_{ij} + \xi_{1ij},\ \xi_{1ij} \sim N(0,\ \tau_{11}^2)$$

where $\beta_{00\,j}$ is the mean mathematics achievement in school *j*, $\beta_{01\,j}$ is the average difference of mathematics achievement between students in small classes and students in

Table 3.5: Effects of Covariates, Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 3, $w_2 = 0.6$ and $w_3 = 0.6$: CRD, Linear Rate of Change

M	*N*									
	10	20	30	40	50	60	70	80	90	100
10	0.195	0.253	0.283	0.302	0.315	0.324	0.331	0.337	0.341	0.345
20	0.525	0.670	0.734	0.768	0.789	0.804	0.814	0.822	0.829	0.833
30	0.727	0.861	0.906	0.927	0.939	0.946	0.952	0.955	0.958	0.960
40	0.851	0.945	0.970	0.979	0.984	0.987	0.989	0.990	0.991	0.992
50	0.922	0.980	0.991	0.994	0.996	0.997	0.998	0.998	0.998	0.998
60	0.960	0.993	0.997	0.999	0.999	0.999	1.000	1.000	1.000	1.000
70	0.981	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	0.991	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
90	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Effect size is 0.4 with a significance level of 0.05.

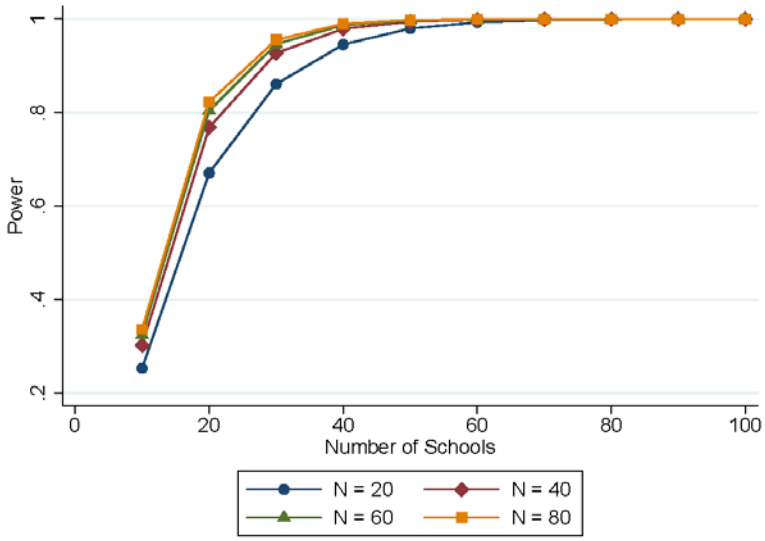

Figure 3.5: Effects of Covariates, Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 3, $w_2 = 0.6$ and $w_3 = 0.6$: CRD, Linear Rate of Change
Note. Effect size is 0.4 with a significance level of 0.05.

regular classes in school *j*, β_{10j} is the average linear growth rate in school *j*, and β_{11j} is the average difference of linear growth rate between students in small classes and students in regular classes in school *j*.

The third level model (school level) is

$$\beta_{00j} = \gamma_{000} + \eta_{00j},\ \eta_{00j} \sim N(0,\ \omega_{00}^2)$$
$$\beta_{01j} = \gamma_{010} + \eta_{01j},\ \eta_{01j} \sim N(0,\ \omega_{T00}^2)$$
$$\beta_{10j} = \gamma_{100} + \eta_{10j},\ \eta_{10j} \sim N(0,\ \omega_{10}^2)$$
$$\beta_{11j} = \gamma_{110} + \eta_{11j},\ \eta_{11j} \sim N(0,\ \omega_{T11}^2)$$

where γ_{000} is the grand mean, γ_{010} is the average treatment effect for all schools, γ_{100} is the average linear rate of change, and γ_{110} is the main effect of treatment for the linear change rate, which is my primary interest. The variance estimates are

$$\sigma_e^2 = 0.30369,\ \tau_{11}^2 = 0.00753,\ \omega_{T11}^2 = 0.02097$$

To calculate power, I assumed a standardized effect size of 0.40 and a significance level of 0.05. I also assumed sample sizes $M = 40$ and $N = 40$, which indicates there were 20 students in the treatment or control condition (40 students in total) in each school and there were 40 schools.

According to equation (3.8) and equation (3.9) with $G = 4$, $p = 1$ and $k_1 = 1$, first I compute

$$\sigma_1^2 = \frac{12 \cdot 0.30369}{5 \cdot 4 \cdot 3} = 0.060738.$$

Then, I calculate the non-centrality parameter of the t-test based on equation (3.31)

$$\lambda = \sqrt{\frac{Mn}{2}} ES \sqrt{\frac{\omega_{Tpp}^2 + \tau_{pp}^2}{nw_3\omega_{Tpp}^2 + w_2\tau_{pp}^2 + \sigma_p^2}} = \sqrt{\frac{40 \cdot 20}{2}} \cdot 0.4 \cdot \sqrt{\frac{0.02097 + 0.00753}{150 \cdot 0.02097 + 0.00753 + 0.060738}} \approx 1.933$$

.

The critical value of the test using the *t*-distribution with 40 - 1 = 39 degrees of freedom is $c(0.25, 39) \approx 2.022$.Finally, I computed power as

$$P = 1 - H\,[2.022, 39, 1.933] + H\,[-2.22, 39, 1.933] \approx 0.471.$$

Table 3.6 to 3.8 and Figure 3.6 to 3.8 show how variations of study duration and sample sizes affect the power to detect the treatment effect for the linear rate of change in block designs, assuming two-tailed *t*-tests at the 0.05 significance level and effect size as 0.40. Table 3.6 and Figure 3.6 show how power changes as study duration (*D*) and the number of schools (*M*) changes, holding the number of students (*N*) in each school constant at 40. As the duration of study increases, the power of detecting a linear rate of change increases slightly when the study duration increases from two to six, and remains virtually unchanged as the study duration increases from six to eight. However, as the number of schools increases, power increases significantly more. For example, when the number of schools increases from 20 to 60, the power is more than doubled. In particular, when *M* = 80 and *D* = 6, or *M* = 90 and *D* = 4, power reaches to 0.80.

Table 3.7 and Figure 3.7 provides power estimates for designs that vary the duration of study (*D*) and the number of students (*N*) in each school, holding the number of schools (*M*) constant at 40. These results re-confirm that the power of detecting a linear rate of

Table 3.6: Effect of Study Duration (*D*) and Number of Schools (*M*) on Power Holding Number of Students (*N*) in Each School Constant at 40: BRD, Linear Rate of Change

D	*M*									
	10	20	30	40	50	60	70	80	90	100
2	0.092	0.145	0.199	0.254	0.307	0.360	0.410	0.458	0.504	0.548
3	0.125	0.222	0.318	0.410	0.494	0.571	0.639	0.698	0.750	0.794
4	0.140	0.255	0.367	0.471	0.563	0.644	0.713	0.771	0.818	0.857
5	0.146	0.269	0.387	0.495	0.591	0.672	0.741	0.797	0.842	0.878
6	0.149	0.275	0.396	0.507	0.603	0.685	0.753	0.808	0.852	0.887
7	0.150	0.278	0.401	0.512	0.609	0.691	0.759	0.814	0.857	0.892
8	0.151	0.280	0.404	0.516	0.613	0.695	0.762	0.817	0.860	0.894

Note. Effect size is 0.4 with a significance level of 0.05

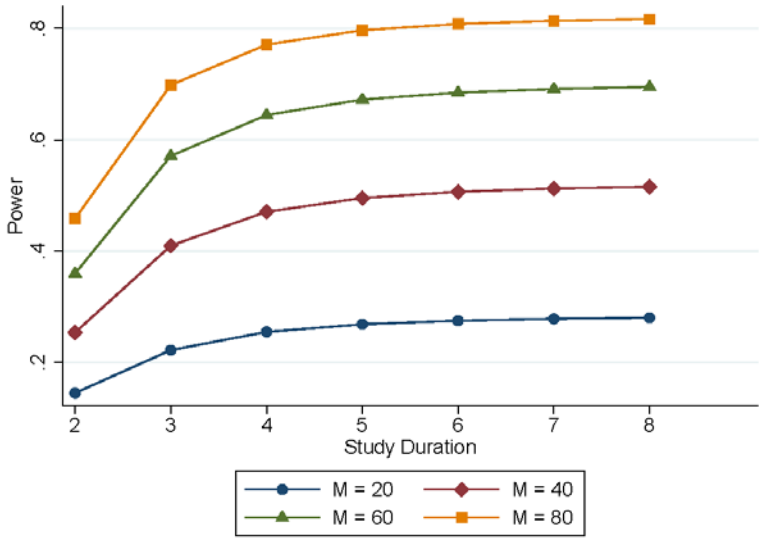

Figure 3.6: Effect of Study Duration (*D*) and Number of Schools (*M*) on Power Holding Number of Students (*N*) in Each School Constant at 40: BRD, Linear Rate of Change

Note: Effect size is 0.4 with a significance level of 0.05.

Table 3.7: Effect of Study Duration (*D*) and Number of Students (*N*) on Power Holding Number of Schools (*M*) Constant at 40: BRD, Linear Rate of Change

D	*N*									
	20	40	60	80	100	120	140	160	180	200
2	0.177	0.254	0.304	0.339	0.365	0.385	0.401	0.413	0.423	0.432
3	0.335	0.410	0.443	0.462	0.474	0.483	0.489	0.494	0.497	0.500
4	0.423	0.471	0.489	0.498	0.504	0.508	0.511	0.514	0.515	0.517
5	0.465	0.495	0.506	0.512	0.515	0.518	0.519	0.521	0.522	0.522
6	0.485	0.507	0.514	0.518	0.520	0.522	0.523	0.524	0.524	0.525
7	0.496	0.512	0.518	0.521	0.523	0.524	0.525	0.525	0.526	0.526
8	0.502	0.516	0.520	0.522	0.524	0.525	0.525	0.526	0.526	0.527

Note. Effect size is 0.4 with a significance level of 0.05

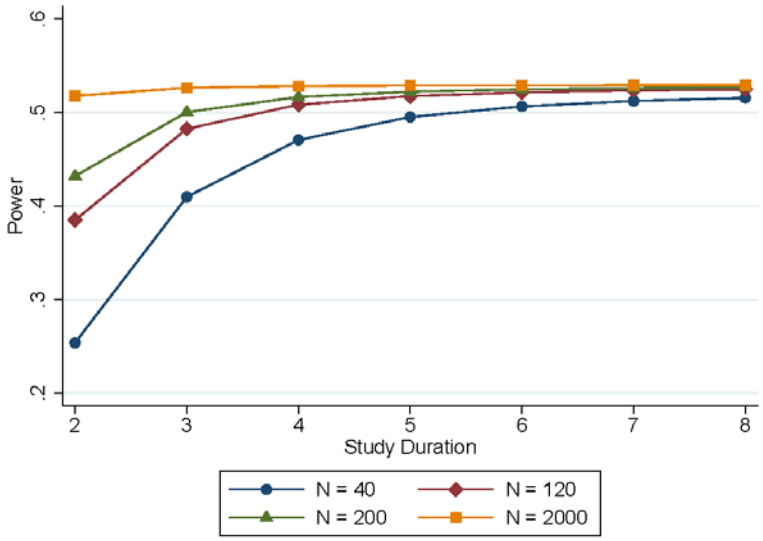

Figure 3.7. Effect of Study Duration (*D*) and Number of Students (*N*) on Power Holding Number of Schools (*M*) Constant at 40: BRD, Linear Rate of Change

Note: Effect size is 0.4 with a significance level of 0.05.

Table 3.8: Effects of Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 4: BRD, Linear Rate of Change

M	*N*									
	20	40	60	80	100	120	140	160	180	200
10	0.128	0.140	0.144	0.146	0.148	0.149	0.150	0.150	0.151	0.151
20	0.229	0.255	0.265	0.270	0.274	0.276	0.278	0.279	0.280	0.281
30	0.329	0.367	0.382	0.390	0.395	0.398	0.400	0.402	0.404	0.405
40	0.423	0.471	0.489	0.498	0.504	0.508	0.511	0.514	0.515	0.517
50	0.510	0.563	0.584	0.594	0.601	0.605	0.608	0.611	0.612	0.614
60	0.588	0.644	0.665	0.676	0.682	0.687	0.690	0.692	0.694	0.696
70	0.656	0.713	0.734	0.744	0.750	0.755	0.758	0.760	0.762	0.763
80	0.716	0.771	0.790	0.800	0.806	0.810	0.813	0.815	0.816	0.818
90	0.767	0.818	0.836	0.845	0.850	0.854	0.857	0.858	0.860	0.861
100	0.810	0.857	0.873	0.881	0.886	0.889	0.891	0.893	0.894	0.895

Note. Effect size is 0.4 with a significance level of 0.05

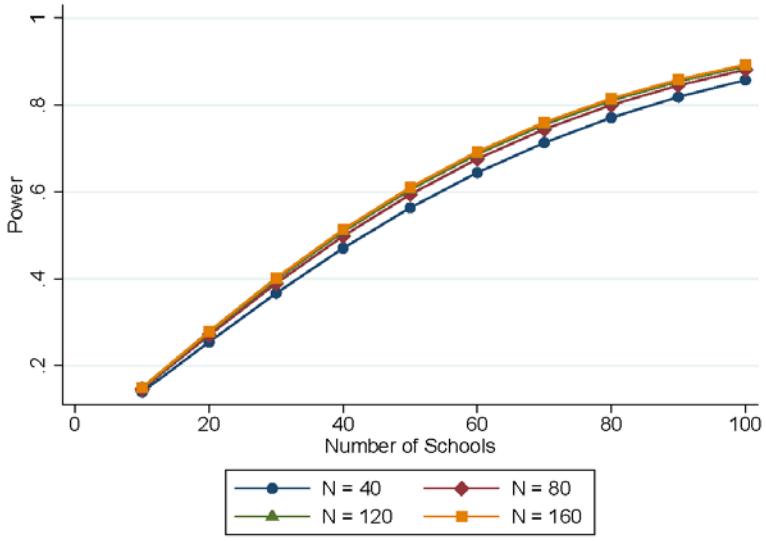

Figure 3.8: Effects of Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 4: BRD, Linear Rate of Change
Note: Effect size is 0.4 with a significance level of 0.05.

change does not increase consistently as study duration increases. In addition, when the number of students in each school is small (e.g., 20), power is impacted more as study duration increases from two to four, compared to the power estimates when the number of students is large (e.g., 200). Similarly, power does not increase consistently as the number of students increases, especially after a certain number of students. For example, the power does not change much as the number of students increases from 160 to 200. What is more, it is hardly to boost power through increasing the number of students per schools. For example, as shown in Figure 3.7, even if there are 2000 students per school, powers are still around 0.5.

Table 3.8 and Figure 3.8 provides power estimates for designs that vary the number of students (N) in each school and the number of schools (M), holding study duration constant at four. As the number of schools increases, power increases consistently. For example, power increases approximately 0.1 as the number of schools changes from ten to 50, and then powers increases around 0.06 for every ten school increase until they reach to 0.80. When M = 80 schools and N = 80 students, power becomes 0.80. In addition, power increases as the number of students increase from 20 to 80, but does not change much as the number of students increases from 100 to 200. Such results indicate that to boost power it is recommended to sample more schools rather than to sample more students per school.

Table 3.9 and Figure 3.9 shows how the power of detecting a linear rate of change is influenced by the proportion of unexplained variance at the second and third levels when M = 40, N = 40, D = 4, and ES = 0.40. I assume that five covariates are added at the third level (q = 5) and thus the degrees of freedom reduce to 40 - 5 - 1 = 34. The results show that power increases when covariates are added in the model, as expected. For example,

Table 3.9: Effect of Covariates on Power: BRD, Linear Rate of Change

W3	*W2*								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.983	0.982	0.982	0.981	0.980	0.980	0.979	0.978	0.977
0.2	0.931	0.929	0.928	0.927	0.926	0.925	0.923	0.922	0.921
0.3	0.858	0.857	0.855	0.854	0.853	0.851	0.850	0.848	0.847
0.4	0.782	0.781	0.780	0.778	0.777	0.776	0.774	0.773	0.772
0.5	0.712	0.711	0.710	0.709	0.707	0.706	0.705	0.704	0.703
0.6	0.650	0.649	0.648	0.647	0.646	0.645	0.644	0.643	0.642
0.7	0.596	0.595	0.594	0.593	0.592	0.591	0.590	0.589	0.589
0.8	0.549	0.548	0.547	0.547	0.546	0.545	0.544	0.543	0.543
0.9	0.508	0.508	0.507	0.506	0.506	0.505	0.504	0.504	0.503

Note. The study duration is 4 with 40 schools and 40 students in each school; significance level is 0.05

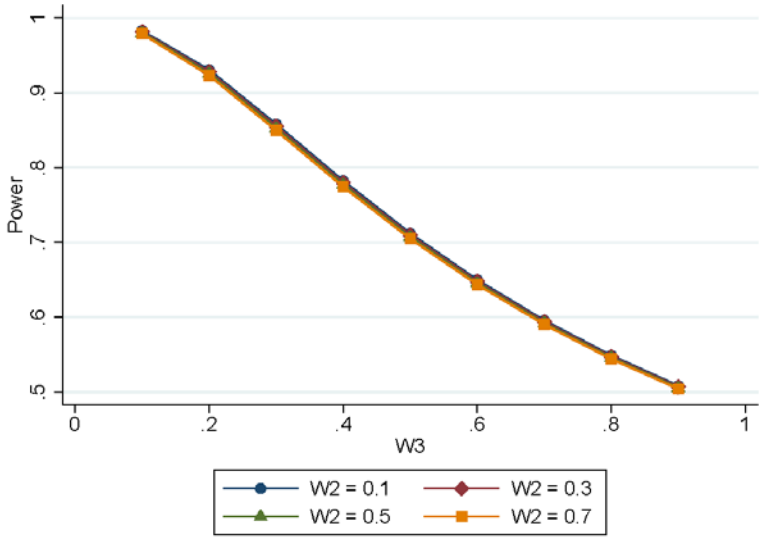

Figure 3.9: Effect of Covariates on Power: BRD, Linear Rate of Change
Note: The study duration is 4 with 40 schools and 40 students in each school; significance level is 0.05.

Table 3.10: Effects of Covariates, Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 4, w_2 = 0.6 and w_3 = 0.6: BRD, Linear Rate of Change

M	*N*									
	20	40	60	80	100	120	140	160	180	200
10	0.136	0.154	0.162	0.166	0.169	0.171	0.172	0.174	0.174	0.175
20	0.302	0.352	0.374	0.386	0.393	0.398	0.402	0.405	0.407	0.409
30	0.443	0.514	0.542	0.558	0.568	0.574	0.579	0.583	0.586	0.588
40	0.565	0.645	0.676	0.692	0.702	0.709	0.714	0.717	0.720	0.723
50	0.666	0.746	0.776	0.791	0.800	0.806	0.811	0.814	0.817	0.819
60	0.748	0.823	0.849	0.861	0.869	0.874	0.878	0.881	0.883	0.884
70	0.812	0.878	0.900	0.910	0.916	0.920	0.923	0.925	0.927	0.928
80	0.862	0.918	0.934	0.942	0.947	0.950	0.952	0.954	0.955	0.956
90	0.900	0.945	0.958	0.964	0.967	0.969	0.971	0.972	0.973	0.973
100	0.928	0.964	0.973	0.977	0.980	0.981	0.982	0.983	0.984	0.984

Note. Effect size is 0.4 with a significance level of 0.05

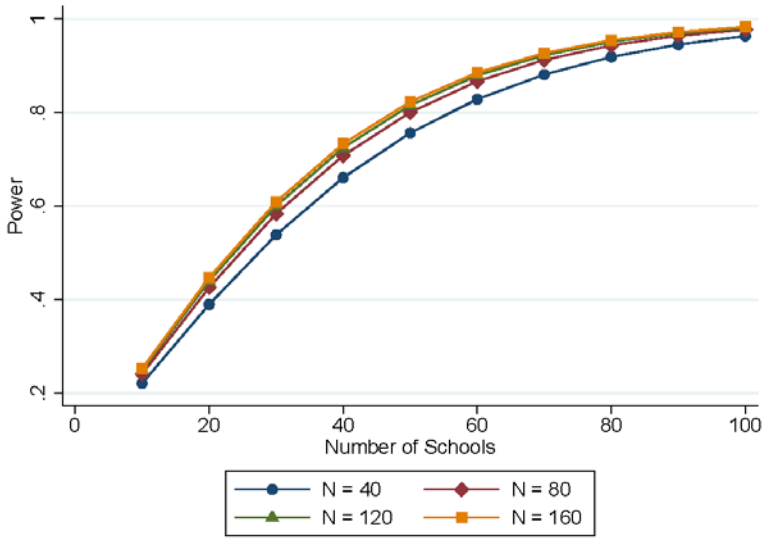

Figure 3.10: Effects of Covariates, Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 4, w_2 = 0.6 and w_3 = 0.6: BRD, Linear Rate of Change

Note. Effect size is 0.4 with a significance level of 0.05.

when covariates explain 40% of the variance at both the second and third level ($w_3 = w_2 = 0.6$), the power increases from 0.471 to 0.641. The power increase is much larger than that when adding 20 schools or adding 160 students in each school. In addition, covariates at the second level have little influence on power, while the covariates at the third level affect power significantly more. Powers does not change much as the proportion of variances explained at the second level increase from 10% to 90% regardless how much of the variances at the third level are explained, which is mainly because the variance of the treatment by school random effects (i.e., τ_{11}^2) only account for a small proportion of the total variance.

Table 3.10 and Figure 3.10 provide power estimates for designs that vary the number of students (N) in each school and the number of schools (M), assuming 40% of variances explained at the second and the third level and holding study duration constant at four. In general, power increases when covariates explain a certain proportion of variance at the second or the third level, comparing the power estimates in Table 3.8. In particular, it requires fewer schools or fewer students per school for power to reach to 0.80. For instance, with $N = 40$, only 60 schools are need to boost power to 0.80, which is 30 schools fewer comparing to the design without covariates included.

Block Randomized Design: A Quadratic Growth Model

I also used data from Project STAR to fit a model with quadratic rate of change at level - 1 (repeated measures), namely

$$Math_{gij} = \alpha_{0ij} c_{0g} + \alpha_{1ij} c_{1g} + \alpha_{2ij} c_{2g} + u_{gij},\ u_{gij} \sim N(0,\ \sigma_e^2),$$

where $Math_{gij}$ is student mathematics achievement in year g, $c_{0g} = (1, 1, 1, 1)$, $c_{1g} = (-1.5, -0.5, 0.5, 1.5)$ and $c_{2g} = (0.5, -0.5, -0.5, 0.5)$ at g = 1, 2, 3, 4 following equation (3.6). This model defines α_{2ij} as the average quadratic rate of change of mathematics achievement for student *i* in school *j*. All the other terms has been defined previously.

The second level model (student level) is

$$\alpha_{0ij} = \beta_{00j} + \beta_{01j} \cdot T_{ij} + \xi_{0ij},\ \xi_{0ij} \sim N(0,\ \tau_{00}^2)$$
$$\alpha_{1ij} = \beta_{10j} + \beta_{11j} \cdot T_{ij} + \xi_{1ij},\ \xi_{1ij} \sim N(0,\ \tau_{11}^2)$$
$$\alpha_{2ij} = \beta_{20j} + \beta_{21j} \cdot T_{ij} + \xi_{2ij},\ \xi_{2ij} \sim N(0,\ \tau_{22}^2)$$

where β_{20j} is the average quadratic growth rate in school *j*, and β_{21j} is the average difference of quadratic growth rate between students in small classes and students in regular classes in school *j*. All the other terms has been defined previously.

The third level model (school level) is

$$\beta_{00j} = \gamma_{000} + \eta_{00j},\ \eta_{00j} \sim N(0,\ \omega_{00}^2)$$
$$\beta_{01j} = \gamma_{010} + \eta_{01j},\ \eta_{01j} \sim N(0,\ \omega_{T00}^2)$$
$$\beta_{10j} = \gamma_{100} + \eta_{10j},\ \eta_{10j} \sim N(0,\ \omega_{10}^2)$$
$$\beta_{11j} = \gamma_{110} + \eta_{11j},\ \eta_{11j} \sim N(0,\ \omega_{T11}^2)$$
$$\beta_{20j} = \gamma_{200} + \eta_{20j},\ \eta_{20j} \sim N(0,\ \omega_{20}^2)$$
$$\beta_{21j} = \gamma_{210} + \eta_{21j},\ \eta_{21j} \sim N(0,\ \omega_{T22}^2)$$

where γ_{200} is the average quadratic rate of change, and γ_{210} is the main effect of treatment for the quadratic change rate, which is my primary interest. The variance estimates are

$$\sigma_e^2 = 0.24239,\ \tau_{22}^2 = 0.00943,\ \omega_{T22}^2 = 0.07524$$

To calculate power, I assumed a standardized effect size of 0.40 and a significance level of 0.05. I also assumed sample sizes $M = 40$ and $N = 40$, which indicates there are 20 students in the treatment or control condition (40 students in total) in each school and there are 40 schools.

According to equation (3.8) and equation (3.9) with $G = 4$, $p = 1$ and $k_2 = \frac{1}{2}$, first I compute

$$\sigma_2^2 = \frac{720 \cdot 0.24239}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2} = 0.24239.$$

Then, I calculate the non-centrality parameter of the t-test based on equation (3.31)

$$\lambda = \sqrt{\frac{Mn}{2}} ES \sqrt{\frac{\omega_{Tpp}^2 + \tau_{pp}^2}{n w_3 \omega_{Tpp}^2 + w_2 \tau_{pp}^2 + \sigma_p^2}} = \sqrt{\frac{40 \cdot 40}{2}} \cdot 0.40 \cdot \sqrt{\frac{0.07524 + 0.00943}{150 \cdot 0..07524 + 0.00943 + 0.24239}} \approx 1.756$$

The critical value of the test using the *t*-distribution with 40 - 1 = 39 degrees of freedom is $c(0.25, 49) \approx 2.023$.Finally, I computed power as

$$P = 1 - \mathrm{H}\,[2.023, 39, 1.756] + \mathrm{H}\,[-2.023, 39, 1.756] \approx 0.403.$$

Table 3.11: Effect of Study Duration (*D*) and Number of Schools (*M*) on Power Holding Number of Students (*N*) in Each School Constant at 40: BRD, Quadratic Rate of Change

D	*M*									
	10	20	30	40	50	60	70	80	90	100
3	0.093	0.148	0.205	0.261	0.316	0.370	0.422	0.471	0.518	0.562
4	0.124	0.218	0.313	0.403	0.486	0.562	0.630	0.689	0.741	0.785
5	0.132	0.237	0.341	0.438	0.527	0.606	0.675	0.734	0.784	0.825
6	0.134	0.243	0.349	0.448	0.538	0.618	0.687	0.745	0.795	0.836
7	0.135	0.244	0.351	0.452	0.542	0.622	0.691	0.749	0.798	0.839
8	0.135	0.245	0.353	0.453	0.544	0.624	0.692	0.751	0.800	0.840
9	0.135	0.245	0.353	0.454	0.544	0.624	0.693	0.752	0.801	0.841

Note. Effect size is 0.4 with a significance level of 0.05.

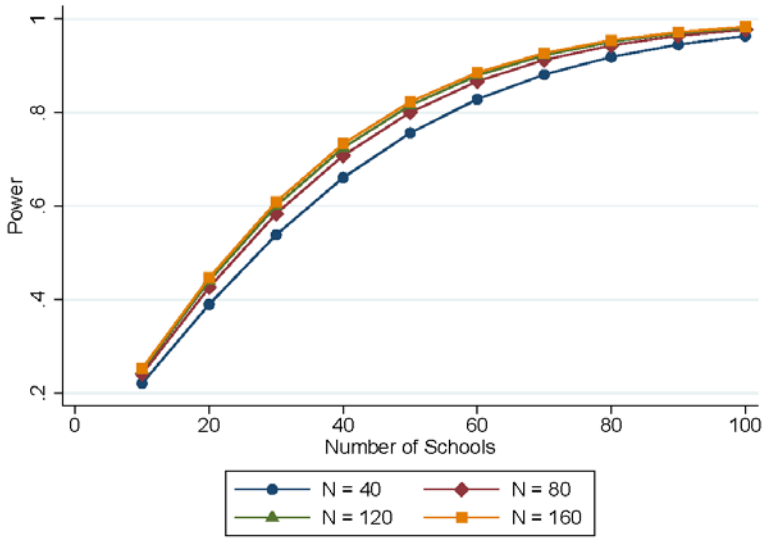

Figure 3.11: Effect of Study Duration (*D*) and Number of Schools (*M*) on Power Holding Number of Students (*N*) in Each School Constant at 40: BRD, Quadratic Rate of Change

Note. Effect size is 0.4 with a significance level of 0.05.

Table 3.12: Effect of Study Duration (*D*) and Number of Students (*N*) on Power Holding Number of Schools (*M*) Constant at 40: BRD, Quadratic Rate of Change

D	*N*									
	20	40	60	80	100	120	140	160	180	200
3	0.190	0.261	0.302	0.330	0.349	0.363	0.374	0.382	0.389	0.395
4	0.360	0.403	0.419	0.428	0.433	0.437	0.440	0.442	0.443	0.445
5	0.421	0.438	0.444	0.447	0.449	0.450	0.451	0.452	0.452	0.453
6	0.440	0.448	0.451	0.452	0.453	0.454	0.454	0.454	0.455	0.455
7	0.447	0.452	0.453	0.454	0.455	0.455	0.455	0.455	0.455	0.456
8	0.449	0.453	0.454	0.455	0.455	0.455	0.456	0.456	0.456	0.456
9	0.451	0.454	0.455	0.455	0.455	0.456	0.456	0.456	0.456	0.456

Note. Effect size is 0.4 with a significance level of 0.05.

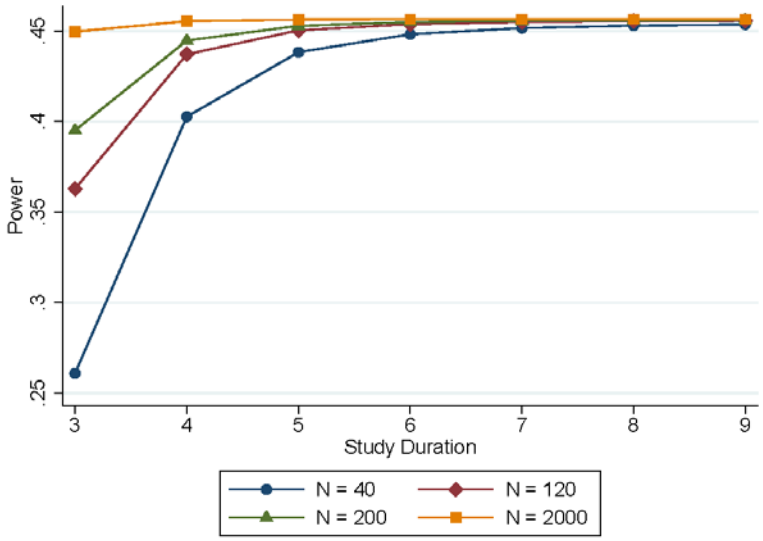

Figure 3.12: Effect of Study Duration (*D*) and Number of Students (*N*) on Power Holding Number of Schools (*M*) Constant at 40: BRD, Quadratic Rate of Change

Note. Effect size is 0.4 with a significance level of 0.05.

Table 3.13: Effects of Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 4: BRD, Quadratic Rate of Change

M	*N*									
	20	40	60	80	100	120	140	160	180	200
10	0.114	0.124	0.127	0.129	0.131	0.132	0.132	0.133	0.133	0.133
20	0.197	0.218	0.227	0.232	0.235	0.237	0.238	0.239	0.240	0.241
30	0.280	0.313	0.326	0.333	0.337	0.340	0.342	0.344	0.345	0.346
40	0.360	0.403	0.419	0.428	0.433	0.437	0.440	0.442	0.443	0.445
50	0.437	0.486	0.505	0.515	0.521	0.526	0.529	0.531	0.533	0.534
60	0.508	0.562	0.583	0.593	0.600	0.605	0.608	0.610	0.612	0.614
70	0.572	0.630	0.651	0.662	0.669	0.673	0.677	0.679	0.681	0.683
80	0.631	0.689	0.710	0.721	0.728	0.732	0.736	0.738	0.740	0.741
90	0.683	0.741	0.761	0.772	0.778	0.782	0.785	0.788	0.790	0.791
100	0.730	0.785	0.805	0.814	0.820	0.824	0.827	0.829	0.831	0.832

Note. Effect size is 0.4 with a significance level of 0.05.

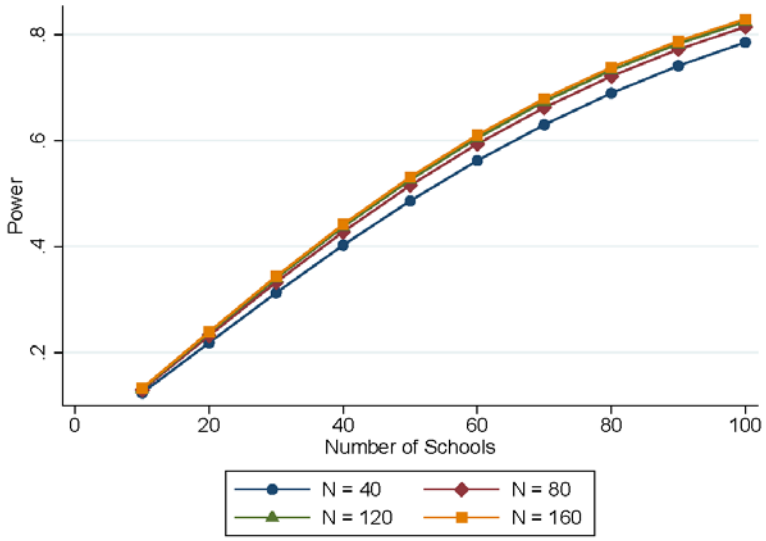

Figure 3.13: Effects of Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 4: BRD, Quadratic Rate of Change

Note. Effect size is 0.4 with a significance level of 0.05.

Table 3.11 to 3.13 and Figure 3.11 to 3.13 show how variations of study duration and sample sizes affect the power to detect the treatment effect of the quadratic rate of change in block designs, assuming two-tailed *t*-tests at the 0.05 significance level and effect size as 0.40. Table 3.11 and Figure 3.11 show how power changes as study duration (*D*) and the number of schools (*M*) changes, holding the number of students (*N*) in each school constant at 40. Please note that there should be at least three repeated measures (*D* = 3) to estimate a quadratic growth model. As the duration of study increases, the power of detecting a quadratic rate of change increases slightly when the study duration increases from three to six; and remains virtually unchanged as the study duration increases from six to nine. However, as the number of schools increases, power increases significantly more. It should be noted that it requires more schools and longer study duration for power to reach to 0.80 comparing the results from linear growth model. That is mainly because the ratio between the level-2 random effects and the variance of treatment by school random effect (i.e., $\tau_{22}^2 / \omega_{T22}^2$) in the quadratic growth model was smaller than the ratio between the level-2 random effects and the variance of treatment by school random effect (i.e., $\tau_{11}^2 / \omega_{T11}^2$) and in the linear growth model. In particular, when *M* = 90 and *D* = 8, or *M* = 100 and *D* = 7, power reaches to 0.80.

Table 3.12 and Figure 3.12 provides power estimates for designs that vary the duration of study (*D*) and the number of students (*N*) in each school, holding the number of schools (*M*) at 40. The results were quite similar to those in Table 3.7. Both the study duration and the number of students in each school have quite limited influence on the power, especially when the study duration or the number of students in each school is beyond a certain

Table 3.14: Effect of Covariates on Power: BRD, Quadratic Rate of Change

W3	*W2*								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.950	0.949	0.949	0.948	0.948	0.948	0.947	0.947	0.946
0.2	0.865	0.865	0.864	0.864	0.863	0.863	0.862	0.861	0.861
0.3	0.774	0.774	0.773	0.773	0.772	0.772	0.771	0.770	0.770
0.4	0.691	0.691	0.691	0.690	0.690	0.689	0.689	0.688	0.688
0.5	0.620	0.620	0.620	0.619	0.619	0.618	0.618	0.618	0.617
0.6	0.561	0.560	0.560	0.560	0.559	0.559	0.559	0.558	0.558
0.7	0.511	0.510	0.510	0.510	0.509	0.509	0.509	0.509	0.508
0.8	0.468	0.468	0.468	0.468	0.467	0.467	0.467	0.467	0.466
0.9	0.432	0.432	0.432	0.432	0.432	0.431	0.431	0.431	0.431

Note. The study duration is 4 with 40 schools and 40 students in each school; significance level is 0.05.

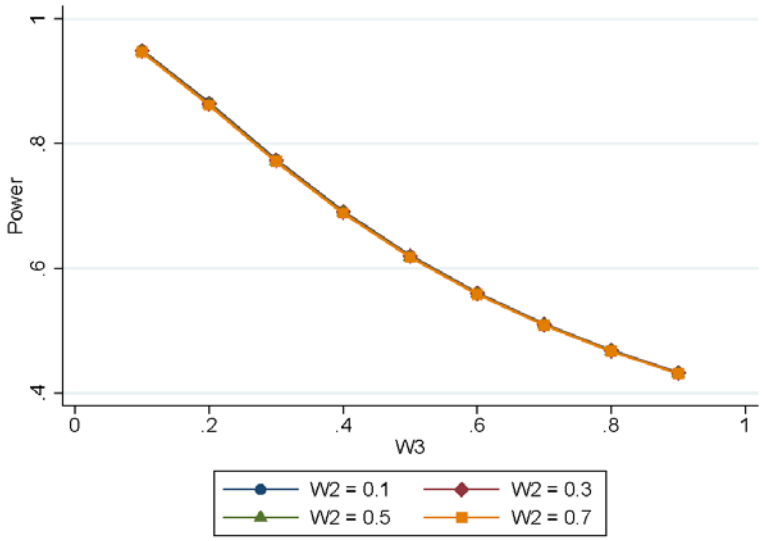

Figure 3.14: Effect of Covariates on Power: BRD, Quadratic Rate of Change
Note. The study duration is 4 with 40 schools and 40 students in each school; significance level is 0.05.

Table 3.15: Effects of Covariates, Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 4, w_2 = 0.6 and w_3 = 0.6: BRD, Quadratic Rate of Change

M	*N*									
	20	40	60	80	100	120	140	160	180	200
10	0.120	0.135	0.142	0.146	0.148	0.150	0.151	0.152	0.153	0.153
20	0.254	0.298	0.317	0.328	0.335	0.339	0.343	0.345	0.347	0.349
30	0.373	0.438	0.465	0.480	0.489	0.496	0.500	0.504	0.507	0.509
40	0.481	0.559	0.590	0.607	0.618	0.625	0.630	0.634	0.637	0.640
50	0.576	0.660	0.692	0.710	0.720	0.727	0.732	0.736	0.739	0.742
60	0.658	0.742	0.773	0.789	0.799	0.805	0.810	0.813	0.816	0.818
70	0.727	0.807	0.835	0.849	0.858	0.863	0.867	0.870	0.873	0.874
80	0.784	0.857	0.882	0.894	0.901	0.905	0.909	0.911	0.913	0.915
90	0.831	0.896	0.916	0.926	0.932	0.935	0.938	0.940	0.941	0.943
100	0.869	0.924	0.941	0.949	0.953	0.956	0.958	0.960	0.961	0.962

Note. Effect size is 0.4 with a significance level of 0.05.

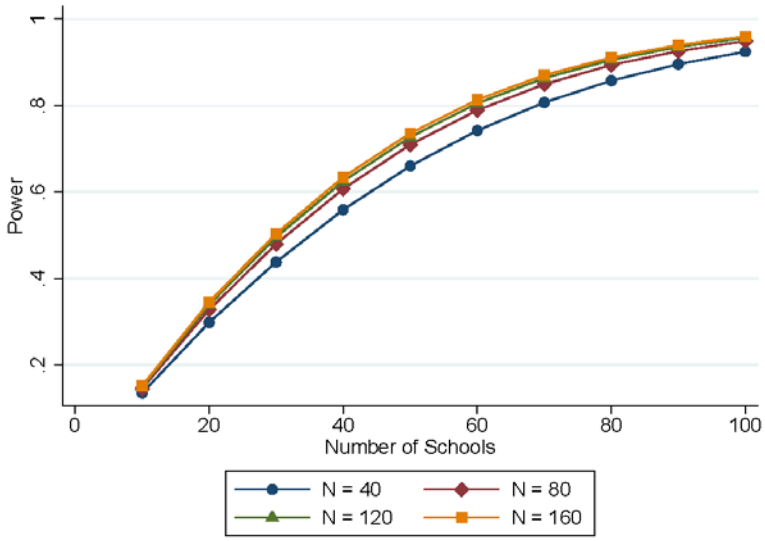

Figure 3.15: Effects of Covariates, Number of Schools (*M*) and Number of Students (*N*) on Power Holding Study Duration (*D*) Constant at 4, w_2 = 0.6 and w_3 = 0.6: BRD, Quadratic Rate of Change

Note. Effect size is 0.4 with a significance level of 0.05.

number. Table 3.13 and Figure 3.13 provides power estimates for designs that vary the number of students (*N*) in each school and the number of schools (*M*), holding study duration constant at four. As the number of schools increases, power increases consistently as expected. When $M = 100$ schools and $N = 60$ students, power becomes 0.80.

Table 3.14 and Figure 3.14, and Table 15 and Figure 15 show how the power of detecting a quadratic rate of change is influenced by the proportion of unexplained variance at the second and third levels when $D = 4$, $ES = 0.40$ and $q = 5$. The results are quite similar to the results in Table 3.9 and Table 3.10. Power increases as the proportion of variances explained increases at the second or the third level. In particular, fewer schools or fewer students per school are needed for power to reach to 0.80. For instance, with $N = 40$, only 70 schools are need to boost power to 0.80, which is more than 30 schools fewer comparing to the design without covariates included. In addition, covariates at the third level have more impacts on power than covariates at the second level.

Conclusion

Multilevel experimental designs are becoming more common in education. Frequently these designs assign individuals (e.g., students) or entire clusters such as schools randomly to a treatment or a control group and follow them over time. In such designs, researchers face the challenge of choosing study duration and sample sizes to ensure that treatment effects will be detected. The present study extended previous work on power analysis for two-level models in studies of polynomial change and presented methods for three-level models.

The power of the test of the treatment effect in studies of polynomial change with two-levels of nesting is a function of the magnitude of the treatment effect, the study duration, the sample size of individuals, the sample size of clusters, and the proportion of the variances that covariates at the second or third levels explain.

Several findings emerged from this study that applies to both CRD and BRD. First, power increases as the study duration, the number of students in each school, or the number of schools increases. Other things being equal, the number of level-3 units (clusters) influences power more than the number of level-2 units (individuals) or the duration of the study. In particular, the number of students and the study duration have limited influence on power. This indicates clearly that researchers should sample more schools rather than students within schools to maximize power. Note that the number of schools impacts power through the degrees of freedom of the *t*-test. It also should be noted that the higher order polynomials a growth model includes, the longer the study duration is needed. For example, to fit a linear rate of change model, the minimum study duration is two; to fit a quadratic rate of change model, the study duration should be at least three.

Second, covariates that explain a proportion of variances at the second or third level could increases powers and thus reduce the study duration or sample sizes needed to boost power to a certain level. For instance, in the first illustrative example with a CRD, when covariates could explain 40% of variances at both the second and the third level, the required number of schools for power reaching to 0.80 drops from 30 to 20, holding the number of students in each school constant at 60. Because the number of covariates at the third level reduces the degrees of freedom for the *t*-test researchers should use a small

number of third level covariates that are strongly related to the outcome, especially when the number of schools is not large.

Third, the effects of covariates on powers depend on the ratio between the variance of the random effects at the second and the third level. For instance, in my three illustrative examples, since the ratio between the variance of the random effects at the second and the third level is small, covariates at the third level affect powers more significantly than covariates at the second level. In addition, comparing the results from the second and the third illustrative sample, powers are larger in the second sample, which is mainly because the ratio between the variance of the random effects at the second and the third level is larger in the second example than that in the third example.

APPENDICES

Appendix A: Variable Description

Table A.1: Variable Names and Coding Methods using Data from TIMSS 2011

Variables:	Description (TIMSS Variable Name)
Student Variables	
Mathematics Achievement	Set of five overall mathematics score plausible value variables
Female	Binary indicator for the student whose gender is female
Age	Student age at the time of testing
Speaking the Tested Language at Home	Binary indicator for the student who spoke the tested language at home "always or almost always"
SES: Books in the Home	Number of books in the home
SES: Items in the Home	Sum of eleven wealth-related household possessions variables
Positive Affect to Mathematics	Average of five self-reported student's affect to mathematics variables, with negatively-worded items reverse-coded
Parents Asked What the Student was Learning in School	Binary indicator for the parents asking the student what he/she is learning in school every day or almost every day
Student Talked about the Schoolwork with Parents	Binary indicator for the student talking about the schoolwork with parents every day or almost every day
Parents Made Sure the Student Set Aside Time for the Homework	Binary indicator for the parents making sure that the student sets aside time for the homework every day or almost every day
Parents Checked if the Student Did the Homework	Binary indicator for the parents checking if the student does the homework every day or almost every day
Teacher/Classroom Variables	
Class Size	Number of students in the classroom
Classroom SES: Books	Average number of books in the home
Classroom SES: Items	Average number of items in the home
Proportion Female	Proportion of female students in a class
Average Students' Positive Affect to Mathematics	Average self-reported student's affect to mathematics in a class
Teacher Experience in Years	Teacher's year of teaching
Teacher Completing Post-Secondary Education	Binary indicator for the teacher who completed post-secondary education
Female	Binary indicator for the teacher who is female
Instruction Time	Time spending teaching mathematics to the students in the class per week
School Variables	
Percent Disadvantaged Students	Set of four indicators for categorical percentage of economically-disadvantaged students
Percent of Students Having Tested Language as Native Language	Binary indicator for categorical percentage of the students having tested language as their native language more than 90%
Students Having Early Numeracy Skills	Set of four indicators for categorical percentage of the students entering the primary grades with early numeracy skills
City Size	Set of six indicators for categorical city population (labels = 0–3,000, 3,001–15,000, 15,001-50,000, 50,001-100,000, 100,001–500,000, greater than 500,000)
Income Level of the School's Immediate Area	Set of three indicators for the income level of the school's immediate area
Grade 4 Enrollment	Total enrollment of fourth graders in the school

Appendix B: Control Function Approach for Quantile Regression

A quantile regression model with endogenous variables can be written as

$$\begin{aligned} y &= x'\beta(\tau) + z'_1\gamma(\tau) + u \\ x &= z'\pi(\tau_r) + v \end{aligned} \tag{B.1}$$

where x is a vector of endogenous variables, and $z=(z_1, z_2)$ are exogenous variables, and our interest is to estimate $\beta(\tau)$, the coefficients of x at τ-*th* quantile.

There are three ways to estimate $\beta(\tau)$ in quantile regression literature. Amemiya (1982) and Powell (1983) first proposed a two-stage absolute value (2LAD) approach, which specifically focused on the median and is quite similar to the 2SLS estimation procedure. However, the required assumption for this approach is difficult to interpret and thus it was not been used widely for empirical studies.

Chernozhukov and Hansen (2006) proposed an Instrumental variable quantile regression (IVQREG) approach that assume $Q_u(\tau \mid z) = 0$, which means the τ-*th* quantile of u –one of the error terms in equation (A2.1) equals to zero, conditional on the other error term (z) in in equation (A2.1).

Chernozhukov and Hansen (2008) developed inference procedures that are fully robust to weak instruments based on the IVQREG estimator. However, there is only Matlab codes available to their approach. In addition, it is not clear how to incorporate sampling weights and how to adjust the clustering effects (e.g., students nested in schools) using their methods.

Lee (2007) proposed a control function approach deal with the endogenous variables in quantile regression. According to equation (B.1) we have

$$Q_u(\tau \mid x, z) = Q_u(\tau \mid v, z)\,. \tag{B.2}$$

This approach assumes that the instrument variables *z* is independent of (*u*, *v*), therefore we have

$$Q_u(\tau \mid v, z) = Q_u(\tau \mid v, z)\,. \tag{B.3}$$

Substitute equation (A2.3) to equation (A2.1), we have

$$\begin{aligned} Q_y(\tau \mid x, z, v) &= x'\beta(\tau) + z'_1\gamma(\tau) + Q_u(\tau \mid z, v) \\ &= x'\beta(\tau) + z'_1\gamma(\tau) + Q_u(\tau \mid v). \end{aligned} \tag{B.3}$$

Therefore, to estimate $\beta(\tau)$, we must know $Q_u(\tau \mid v)$, which is a function of *v*. Since *v* is not observed, we must estimate it through regressing *x* on *z* using OLS or quantile regression. Also, we have no idea if the correlation between *u* and *v* is linear or non-linear, Lee (2007) suggest using a series or kernel of *v* to better model the relationship between *u* and *v*.

To sum up, Lee's (2007) control function approach is also a two-stage estimation approach: (1) regression *x* on *z* using OLS or quantile regression and get $\hat{v} = x - z'\pi(\tau_r)$; (2) regress *y* on *x*, z_1, and a series or kernel of $\hat{v}$ through quantile regression to get $\beta(\tau)$.

Appendix C: Proof of Equation (3.6)

According to Randenbush and Liu (2001), Y_{gi} is an outcome for person i (i=1, 2, …, n) at occasion g (g=1, 2, ..., G) and thus $\sum_{g=1}^{G} m = \frac{G \times (1+G)}{2}$. According to the equation (5) in Randenbush and Liu (2001), the equation (2) in Randenbush and Liu (2001) could be simplified as

$$
\begin{aligned}
C_{0g} &= 1 \\
C_{1g} &= g - \sum_{g=1}^{G} \mathrm{g}/\,\mathrm{G} = g - \bar{g} \\
C_{2g} &= \frac{1}{2}(C_{1g}^2 - \sum_{g=1}^{G} C_{1g}^2 / G) = \frac{1}{2}\left[(g-\bar{g})^2 - \frac{(\mathrm{G}+1)\cdot G \cdot (G-1)}{12 \cdot G} \right] \\
&= \frac{1}{2}\left[(g-\bar{g})^2 - \frac{G^2-1}{12} \right] \\
C_{3g} &= \frac{1}{6}\left(C_{1g}^3 - \frac{\sum_{g=1}^{G} C_{1g}^4}{\sum_{g=1}^{G} C_{1g}^2} \cdot C_{1g} \right) = \frac{1}{6}\left[(g-\bar{g})^2 - \frac{3G^2-7}{20} \cdot (g-\bar{g}) \right]
\end{aligned}
\qquad \text{(C.1)}
$$

where $\bar{g} = \frac{\sum_{g=1}^{G} g}{G} = \frac{1+G}{2}$.

According to the equation provided in Fisher (1936, P149), we have

$$
\begin{aligned}
\tilde{C}_{0g} &= 1 \\
\tilde{C}_{1g} &= \mathrm{g} - \bar{g} \\
\tilde{C}_{2g} &= C_{1g}^2 - \frac{G^2-1}{12} \\
\tilde{C}_{3g} &= C_{1g}^3 - \frac{3G^2-7}{20} \cdot C_{1g}
\end{aligned}
\qquad \text{(C.2)}
$$

When $k_0 = 1,\ \ k_1 = 1,\ \ k_2 = \frac{1}{2},$ and $k_3 = \frac{1}{6}$, we have

$$\begin{cases} C_{0g} = \tilde{C}_{0g} \times k_0 \\ C_{1g} = \tilde{C}_{1g} \times k_1 \\ C_{2g} = \tilde{C}_{2g} \times k_2 \\ C_{3g} = \tilde{C}_{3g} \times k_3 \end{cases} \text{ or } C_{pg} = \breve{C}_{pg} \times k_p \qquad \text{(C.3)}$$

According to equation in page 30 of Fisher(1957) and equation (1) from Jennrich and Sampson(1971), we have a recurrence formula:

$$\begin{aligned} &\breve{C}_{p+1,\mathrm{g}} = \breve{C}_{1g} \cdot \breve{C}_{pg} - \alpha_p \breve{C}_{p-1,\mathrm{g}} \\ &\breve{C}_{0g} = 1 \\ &\breve{C}_{1g} = g - \overline{g} \end{aligned} \qquad \text{(C.4)}$$

where $\alpha_p = \frac{p^2(G^2 - p^2)}{4 \cdot (4p^2 - 1)}$ and g is number of occasions ($g=1, 2, ..., G$); p is the degree of the orthogonal polynomial; and $\breve{C}_{pm}$ the orthogonal polynomial coefficient of degree p at occasion g.

According to equation (C.4), we have

If $p = 1,$

$$\breve{C}_{2g} = \breve{C}_{1g} \cdot \breve{C}_{1g} - \frac{1 \cdot \left(G^2 - 1\right)}{4 \cdot (4 - 1)} \cdot \breve{C}_{0g} = \breve{C}_{1g}{}^2 - \frac{G^2 - 1}{12} = \left(g - \overline{g}\right)^2 - \frac{G^2 - 1}{12}$$

If , $p = 2$,

$$\breve{C}_{3g} = \breve{C}_{1g} \cdot \breve{C}_{2g} - \frac{4 \cdot (\mathrm{G}^2 - 4)}{4 \cdot (4 \cdot 4 - 1)} \cdot C_{1g}$$

$$= \left(g - \bar{g}\right) \cdot \left[\left(g - \bar{g}\right)^2 - \frac{G^2 - 1}{12}\right] - \frac{G^2 - 4}{15} \cdot \left(g - \bar{g}\right)$$

$$= \left(g - \bar{g}\right)^3 - \left[\frac{G^2 - 1}{12} + \frac{G^2 - 4}{15}\right] \cdot \left(g - \bar{g}\right)$$

$$= \left(g - \bar{g}\right)^3 - \frac{3G^2 - 7}{20} \cdot \left(g - \bar{g}\right)$$

To sum up, we have

$$\begin{cases} C_{0g} = \tilde{C}_{0g} \times k_0 = \breve{C}_{0g} \times k_0 \\ C_{1g} = \tilde{C}_{1g} \times k_1 = \breve{C}_{1g} \times k_1 \\ C_{2g} = \tilde{C}_{2g} \times k_2 = \breve{C}_{2g} \times k_2 \\ C_{3g} = \tilde{C}_{3g} \times k_3 = \breve{C}_{3g} \times k_3 \end{cases} \text{ or } C_{pg} = \tilde{C}_{pg} \times k_p = \breve{C}_{pg} \times k_p \qquad \text{(C.5)}$$

Appendix D: Proof of Equation (3.9)

Based on the equation (2) from Jennrich and Sampson (1971), we have:

$$\begin{cases} \sum_{g=1}^{G} \breve{C}_{pg}^2 = \alpha_p \sum_{g=1}^{G} \breve{C}_{p-1,g}^2, \text{ where } \alpha_p = \dfrac{p^2(G^2 - p^2)}{4 \cdot (4p^2 - 1)} \\ \sum_{g=1}^{G} \breve{C}_{0g}^2 = G \end{cases} \tag{D.1}$$

Therefore we have

$$\begin{cases} If\ p = 1,\ \sum_{g=1}^{G} \breve{C}_{1g}^2 = \dfrac{1 \cdot (G^2 - 1)}{4 \cdot 3} \cdot G = \dfrac{(G+1) \cdot G \cdot (G-1)}{12} \\ If\ p = 2, \sum_{g=1}^{G} \breve{C}_{2g}^2 = \dfrac{4 \cdot (G^2 - 4)}{4 \cdot (4 \cdot 4 - 1)} \cdot \dfrac{(G+1) \cdot G \cdot (G-1)}{12} \\ \qquad = \dfrac{(G+2)(G+1) \cdot G \cdot (G-1)(G-2)}{180} \\ If\ p = 3, \sum_{g=1}^{G} \breve{C}_{3g}^2 = \dfrac{9 \cdot (G^2 - 9)}{4 \cdot (4 \cdot 9 - 1)} \cdot \dfrac{(G+2)(G+1) \cdot G \cdot (G-1)(G-2)}{180} \\ \qquad = \dfrac{(G+3)(G+2)(G+1) \cdot G \cdot (G-1)(G-2)(G-3)}{2800} \end{cases}$$

Let $k_0 = 1,\ \ k_1 = 1,\ \ k_2 = \dfrac{1}{2},$ and $k_3 = \dfrac{1}{6}$, we have $\sum_{g=1}^{G} C_{pg}^2 = k_p^2 \times \sum_{g=1}^{G} \breve{C}_{pg}^2 \quad (p = 1, 2, 3)$

Also, according to equation (D.1), we have:

$$\sum_{g=1}^{G} \breve{C}_{pg}^2 = \frac{p^2(M^2 - p^2)}{4 \cdot (4p^2 - 1)} \cdot \sum_{g=1}^{G} \breve{C}_{p-1,g}^2 = \frac{p^2}{4(4p^2 - 1)} \cdot (G^2 - p^2) \cdot \sum_{g=1}^{G} C_{p-1,g}^2$$

Let $\xi_p = \dfrac{p^2}{4(4p^2 - 1)}$, we have

$$\sum_{m=1}^{M}\breve{C}_{pg}^{2}=\xi_{p}\cdot(\mathrm{G}^{2}-p^{2})\cdot\sum_{g=1}^{G}C_{p-1,\mathrm{g}}^{2}$$

$$=\xi_{p}\cdot(\mathrm{G}+p)\cdot(\mathrm{G}-p)\cdot\xi_{p-1}\cdot\left[G+(p-1)\right]\cdot\left[G-(p-1)\right]\cdot\sum_{g=1}^{G}C_{p-2,\mathrm{g}}^{2}$$

$$=\xi_{p}\cdot(\mathrm{G}+p)\cdot(\mathrm{G}-p)\cdot\xi_{p-1}\cdot\left[G+(p-1)\right]\cdot\left[G-(p-1)\right]\cdot\ldots$$

$$\cdot\xi_{p-(p-1)}\cdot\left[G+(p-p+1)\right]\cdot\left[G-(p-p+1)\right]\cdot\sum_{m=1}^{M}\breve{C}_{0g}^{2}$$

$$=\xi_{p}\cdot\xi_{p-1}\cdot\xi_{p-2}\cdot\ldots\xi_{1}\cdot(\mathrm{G}+p)\cdot(\mathrm{G}+p-1)\cdot(\mathrm{G}+p-2)\cdots G\cdot(\mathrm{G}-1)\cdots(\mathrm{G}-\mathrm{p})$$

Let $\mathrm{H}_{p}=\xi_{p}\cdot\xi_{p-1}\cdot\cdots\cdot\xi_{1}$, we have

$$\sum_{g=1}^{G}\breve{C}_{pg}^{2}=H_{p}\cdot\frac{(M+p)!}{(M-p-1)!}. \tag{D.2}$$

Also, Let $\mathrm{K}_{p}=\mathrm{H}_{p}\cdot k_{p}^{2}$, we have

$$\sum_{g=1}^{G}C_{pg}^{2}=\lambda_{p}^{2}\cdot\sum_{g=1}^{G}\breve{C}_{pg}^{2}=\lambda_{p}^{2}\cdot H_{p}\cdot\frac{(G+p)!}{(G-p-1)!}=K_{p}\cdot\frac{(G+p)!}{(G-p-1)!}. \tag{D.3}$$

In addition, according to equation (8) in P. 104 of Plackett (1960), we have:

$$\sum_{G=1}^{G}\breve{C}_{pg}^{2}=\frac{(p!)^{2}\cdot\binom{G+p}{2p+1}}{\binom{2p}{p}}=\frac{(p!)^{2}\cdot\dfrac{(G+P)!}{(G-p-1)!\cdot(2p+1)!}}{\dfrac{(2p)!}{(p!)^{2}}} \tag{D.4}$$

$$=\frac{(p!)^{4}\cdot}{(2p)!\cdot(2p+1)!}\cdot\frac{(G+P)!}{(G-p-1)!}$$

Therefore we can write

$$H_{p}=\frac{(p!)^{4}}{(2p)!\cdot(2p+1)!}$$

$$\mathrm{K}_{p}=k_{p}^{2}\cdot\frac{(p!)^{4}}{(2p)!\cdot(2p+1)!} \tag{D.5}$$

REFERENCES

REFERENCES

Akerhielm, K. (1995). Does class size matter? *Economics of Education Review, 14*(3), 229-241.

Amemiya, T. (1982). Two Stage Least Absolute Deviations Estimators. *Econometrica, 50(3),* 689-711.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics, 114*(2), 533-575.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29(1),* 30-59.

Bloch, D. A. (1986). Sample size requirements and the cost of a randomized clinical trial with repeated measurements. *Statistics in Medicine, 5*(6), 663.667.

Bonesronning, H. (2003). Class size effect on student achievement in Norway: Patterns and explanations. *Southern Economic Journal, 69*(4), 952-965.

Boruch, R. F., & Gomez, H. (1977). Sensitivity, bias, and theory in impact evaluations. *Professional Psychology, 8*(4), 411.

Cho, H., Glewwe, P., & Whitler, M. (2012). Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review, 31*(3), 77-95.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*: Routledge.

Ding, W., & Lehrer, S. F. (2011). Experimental estimates of the impacts of class size on test scores: robustness and heterogeneity. *Education Economics, 19(3),* 229-252.

Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute, 22*(1/3), 23-32.

Dufour, J. M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics, 36(4),* 767-808.

EACEA Eurydice. (2012). *Key data on education in Europe 2012*. Brussels: Eurydice.

Education Week. (2008). *Quality counts 2008: Tapping into teaching*. Bethesda, MD: Editorial Projects in Education.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*(3), 557-577. Fisher, R. A.

(1928). *Statistical methods for research workers* (2d ed.). Edinburgh, London,: Oliver and Boyd.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1*(1), 2-16.

Hausman, J. A. (1978). Specification tests in Econometrics. *Econometrica, 46*(6), 1251-1271.

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics, 24*(1), 70-93.

Hojo, M. (2013). Class-size effects in Japanese schools: A spline regression approach. *Economics Letters, 120*(3), 583-587.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics, 115*(4), 1239-1285.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*(2), 236.

Jackson, E., & Page, M. E. (2013). Estimating the distributional effects of education reforms: A look at Project STAR. *Economics of Education Review, 32,* 92-103.

Jennrich, R. I., & Sampson, P. I. (1971). Remark as R3: A remark on algorithm AS 10. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 20*(1), 117-118.

Jong, K., Moerbeek, M., & Van der Leeden, R. (2010). A priori power analysis in longitudinal three-level multilevel models: an example with therapist effects. *Psychotherapy Research*, *20*(3), 273.284

Kirk, R. E. (2012). *Experimental design : procedures for the behavioral sciences* (4th ed.). Thousand Oaks: Sage Publications.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data.*Stata Journal*, *10*(2), 165-199.

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*(1), 66-88.

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, 1(4), 265-288.

Konstantopoulos S., & Chung, V. (2009). What are the long-term effects of small classes on the achievement gap? Evidence from the Lasting Benefits Study. *American Journal of Education, 116 (1),* 125-154.

Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized

designs: Which level matters more? *Multivariate Behavioral Research, 47*, 392-420.

Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). The impact of indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis, 35(4),* 481-499.

Konstantopoulos, S., & Traynor, A. (2014). Class size effects on reading achievement using PIRLS data: Evidence from Greece. *Teachers College Record, 116*(2), 1-29.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114*(2), 497-532.

Lee, S. (2007). Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics, 141*(2), 1131-1158.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research* (Vol. 19): Sage.

Levin, J. (2001). For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement. *Empirical Economics*,*26*(1), 221-246.

Leuven, E., Oosterbeek, H., & Ronning, M. (2008). Quasi-experimental estimates of the effect of class size on achievement in Norway. *Scandinavian Journal of Economics, 110*(4), 663-693.

Ma, L., & Koenker, R. (2006). Quantile regression methods for recursive structural equation models. *Journal of Econometrics, 134*(2), 471-506.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Moerbeek, M. (2008). Powerful and cost-efficient designs for longitudinal intervention studies with two treatment groups. *Journal of Educational and Behavioral Statistics, 33*(1), 41-61.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., & Ehrle, K. (1999). Evaluating the SAGE program: A pilot program in targeted pupil-teacher reduction in Wisconsin. *Educational Evaluation and Policy Analysis, 21*(2), 165-177.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *Future of Children, 5*(2), 113-127.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*(4), 371-402.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37*(1), 123-151.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2002). Do low-achieving students benefit more from small classes? Evidence from the Tennessee class size experiment. *Educational Evaluation and Policy Analysis, 24*(3), 201-217.

Plackett, R. L. (1960). *Principles of regression analysis.* Oxford: Clarendon Press.

Pong, S. L., & Pallas, A. (2001). Class size and eighth-grade math achievement in the United States and abroad. *Educational Evaluation and Policy Analysis, 23*(3), 251-273.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models : applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Raudenbush, S. W., & Liu, X. F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29(1),* 5-29.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). Hoboken, N.J.: Wiley-Interscience.

Shafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8,* 3-15.

Slavin, R. E. (1989). Class size and student achievement: Small effects of small classes. *Educational Psychologist, 24*(1), 99-110.

Snijders, T., & Bosker, R. J. (1999). Multilevel analysis: an introduction to basic and advanced multilevel modeling. Sage Publications. *Thousand Oaks, CA.*

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. Econometrica, 65(3), 557-586.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics, 20*(4), 518-529.

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural Bolivia. *The Review of Economics and Statistics, 88*(1), 171-177.

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review, 99*(1), 179-215.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wossmann, L. (2005). Educational production in Europe. *Economic Policy,* (43), 445-504.

Wossmann, L., & West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review, 50*(3), 695-736.

Wu, D. M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica, 41* (4), 733-750.