TWO APPLICATIONS OF QUANTITATIVE METHODS IN EDUCATION: SAMPLING DESIGN EFFECTS IN LARGE-SCALE DATA AND CAUSAL INFERENCE OF CLASS-SIZE EFFECTS

By

Ting Shen

A DISSERTATION

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

Measurement and Quantitative Methods — Doctor of Philosophy

2018

ABSTRACT

TWO APPLICATIONS OF QUANTITATIVE METHODS IN EDUCATION: SAMPLING DESIGN EFFECTS IN LARGE-SCALE DATA AND CAUSAL INFERENCE OF CLASS-SIZE EFFECTS

By

Ting Shen

This dissertation is a collection of four papers in which the former two papers address the issues of external validity concerning incorporating complex sampling design in model analysis in large-scale data and the latter two papers address issues of internal validity involving statistical methods that facilitate causal inference of class size effects.

Chapter 1 addressed whether, when and how to apply complex sampling weights via empirical, simulation and software investigations in the context of large-scale educational data focusing on fixed effects. The empirical evidences reveal that unweighted estimates agree with the weighted cases and two scaling methods make no difference. The possible difference between weighted single versus multi-level model may lie in the scaling procedure in the latter. The simulation results indicate that relative bias of the estimates in the models of unweighted single level, unweighted multilevel, weighted single level and weighted multi-level varies across different variables, but unweighted multilevel has the smallest root mean square errors consistently while weighted single model has the largest values for level-one variables. The software finding indicates that STATA and Mplus are more flexible and capable especially for weighted multi-level models where scaling is required. Chapter 2 investigated how to account for informative design arising from unequal probability of selection in multilevel modeling with a focus of the multilevel pseudo maximum likelihood (MPML) and the sample distribution approach (SDA). The Monte Carlo simulation evaluated the performance of MPML considering

sampling weights and scaling. The results indicate that unscaled estimates have substantial positive bias for estimating cluster- and individual-level variations, thus the scaling procedure is essential. The SDA is conducted using empirical data, and the results are similar to the unweighted case which seems that the sampling design is not that informative or SDA is not working well in practice.

Chapter 3 examined the long-term and causal inferences of class size effects on reading and mathematics achievement as well as on non-cognitive outcomes in early grades via applying individual fixed effects models and propensity scores methods on the data of ECLS-K 2011. Results indicate that attending smaller class improves reading and math achievement. In general, evidence of class size effects on non-cognitive outcomes is not significant. Considering potential measurement errors involved in non-cognitive variables, evidence of class size effects on non-cognitive domain is less reliable. Chapter 4 applied instrumental variables (IV) methods and regression discontinuity designs (RDD) on TIMSS data in 2003, 2007 and 2011 to investigate whether class size has effects on eighth grader's cognitive achievement and non-cognitive outcomes in math and four science subjects across four European countries (i.e., Hungary, Lithuania, Romania and Slovenia). The results of the IV analyses indicate that in Romania smaller class size has significant positive effects on academic scores for math, physics, chemistry and earth science as well as for math enjoyment in 2003. In Lithuania, class size effects on non-cognitive skills are not consistent between IV and RDD analyses in 2007. Overall, the small class size benefit on achievement scores is only observed in Romania in 2003 while evidence of class-size effects on non-cognitive skills may lack of reliability.

ACKNOWLEDGEMENTS

I have been fortunate to have an opportunity of intellectual growth at MSU, particularly for the quantitative skills that I have acquired at MQM program. Nevertheless, the lack of this prerequisite skill and the decision of being a parent of two makes the journey more challenging for me to go through. I would like to acknowledge many great, kind people who have helped me in numerous ways and finally make the completion of this journey possible and rewarding, which I am forever grateful.

I would like to thank the members of my dissertation committee. First and foremost, I would like to express my deepest gratitude to my advisor and dissertation chair, Dr. Spyros Konstantopoulos, who has been an excellent mentor to nurture and inspire me to be a thoughtful researcher. He has been my role model in pursuing scholarly adventures. I thank him for providing superb guidance to me throughout my PhD study with his rich knowledge and exceptional insights. This thesis alongside my practicum paper would not have been possibly done without his great ideas and constructive feedback. Working with him has been an amazing and inspirational experience. I also thank Dr. Konstantopoulos for supporting and helping me in solving all kinds of problems with his practical advice and warm note. It has been my great privilege to be his advisee. I look forward to having more research exploration and learning experience in education with him.

Next, I would like to thank Dr. Amita Chudgar for providing a two-year research assistantship to me with very flexible working schedule and condition. It has been a great experience to work on data in developing countries and have a sense of different educational problems worldwide from education policy and gender perspective. Personally, I am truly indebted to Dr. Chudgar for her special care and kindness offered to me when I encountered

crisis and difficulties. I also very much appreciate her continued guidance, support and help on my dissertation study and job hunting. In addition, I thank Dr. Kenneth Frank and Dr. Barbara Schneider for providing service and time, and for contributing insightful feedback and helpful comments. I also thank Dr. Frank for serving in my guidance committee before and for frequently spreading news of academic activity in MQM.

Moreover, I am very thankful to Dr. John Carlson for offering me a one-year research assistantship to work on the Wraparound project through which I have an opportunity of knowing special education a little bit. Thanks also go to Dr. Gary Troia and the Statistical Consulting Center for a short research and work experience provided with assistantship. I thank Dr. Kimberly Maier, Dr. Rand Spiro and Dr. Raoul LePage for their guidance and help provided beyond course instruction. I would like to extend thanks to Dr. Guan Saw for his service and guidance in my practicum study. I thank administrative staff in the College of Education at large for helping deal with paperwork, especially I thank former and current MQM secretaries, Erin Johnson, Adam Rafalski and Brette Smith. I acknowledge the four-year assistantship in MQM and a summer research fund from the College of Education and the Graduate School.

In China, I am very grateful to three teachers, Drs. Mingyan Deng, Feng Chen and Baoxing Wang at East China Normal University for sharing their thoughts and views with me when I have education-related questions to ask. Particularly, I thank my former mentor, Mingyan Deng, who has cared for me and encouraged me when I felt frustrated. I also appreciate his good parenting advice.

I am so thankful to the support and help coming from my classmates and dear friends. I thank Brooke Quisenberry, Tara Kilbride and Dr. Jinyoung Koh for many study-related activities in and outside campus. I also thank Forest Young for being my buddy in learning statistics. I am

thankful to Tom Almer for his fun English class. I thank my former neighbors, Dr. Yanhui Zhao and Na He, for helping me look after my kids in emergency. I am also indebted to Dr. Bin Fan for always offering a helping hand when I got involved in car issues. In China, I am thankful to Lixiao Zheng and Yan Xin for their long-term friendship as my undergraduate buddies since the year of 2001. I am extremely grateful to Hui Yu for taking good care of me with great tenderness as a big brother and for being excellent company of seven years. I am indebted to Hui's parents, Yinliang Yu and Linzhen Yu, for treating me as a daughter.

My deep gratitude goes to my relatives and immediate family. Thanks are due to my aunt, Mingxing Shen, and cousin, Feng Wen, for being close relatives over the years. I am truly thankful to my husband, Haiming Wen, for his friendship and love for a decade. I thank him for his understanding and support in countless occasions throughout my many years of graduate study. I also appreciate that he listens to my ideas of early childhood education and helps me put them into practice as a partner. I thank Lu-Ke and Lu-Na for always being the source of great happiness, for bringing me into their very big little world, and for driving me to seek useful information from research literature to provide a better environment for each of their growth. I own special thanks to my mom, Mingfang Dai, and my dad, Rongguan Shen, for supporting my decision whatever it is and for helping me take care of kids whenever I need them. Finally, I dedicate this thesis to daddy for your extreme patience and unconditional love to me, to mommy, and to Luke and Luna.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

PML	Pseudo Maximum Likelihood
MPML	Multilevel Pseudo Maximum Likelihood
IGLS	Iterative Generalized Least Squares
PWIGLS	Probability Weighted Iterative Generalized Least Squares
ICC	Intra-Class Correlation
SDA	Sample Distribution Approach
MCMC	Monte Carlo Markov Chain
NCES	National Center for Education Statistics
IEA	International Association for the Evaluation of Educational Achievement
OECD	Organization for Economic Co-operation and Development
NAEP	National Assessment of Educational Progress
ECLS-K	Early Childhood Longitudinal Study - Kindergarten
TIMSS	The Trend in International Mathematics and Science Study
PIRLS	The Progress in International Reading Literacy Study
PISA	The Program for International Student Assessment
PSM	Propensity Score Methods
IV	Instrumental Variable
RDD	Regression Discontinuity Design

CHAPTER 1
INCORPORATING COMPLEX SAMPLING WEIGHTS IN MODEL ANALYSIS

1.1 Introduction

Institutions such as the National Center for Education Statistics (NCES), the International Association for the Evaluation of Educational Achievement (IEA) and the Organization for Economic Co-operation and Development (OECD) have invested tremendous resources to conduct large-scale surveys and collect large-scale data in the field of education. Educational researchers have been increasingly encouraged to utilize these high-quality data to inform education research, policy and practice. In the era of data-driven based research, the advantages of using large-scale data sets are well-recognized. For example, large-scale data sets (e.g., the National Assessment of Educational Progress – NAEP and the Program for International Student Assessment – PISA) provide reliable measures of student academic achievement which have been used to identify comparatively high and low achieving students and schools. Moreover, such data include rich information about student characteristics and family background as well as school characteristics and the learning environment in schools that has allowed researchers to investigate various important questions in educational research. Further, the complex sampling design employed in these surveys creates national probability samples of students that represent well-defined populations (e.g., 4^{th} graders in the U.S.). This allows the projection of inference obtained from a sample to its national population.

Although large-scale data sets have enormous potential to advance knowledge and guide research, policy and practice in education, there are some concerns and challenges that need special attentions (Saw & Schneider, 2015; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). For instance, large-scale data sets are non-experimental data, and thus addressing the

issue of potential confounding or omitted variable bias to achieve high internal validity of results (i.e., causal inference) would require the use of special statistical methods and research designs (e.g., Instrumental Variables or Regression Discontinuity Design). In addition, the high external validity of results (i.e., generalizability) is another matter that needs to be considered. Generalizing estimates from a sample to a population, for instance, would require the appropriate use of complex sampling weights. This is an issue of great importance because inappropriate usage of sampling weights may result in erroneous standard errors, possibly inaccurate statistical inference for variables of interest, and consequently misleading evaluation and policy implications. Unfortunately, although the causal inferences have received great attention in the literature, the issue of applying sampling weights in model analyses appropriately, especially for multilevel models has been neglected by both large-scale survey providers and data analysts in education.

Specifically, there are several literature voids concerning weighted analyses that have been observed in educational research, to which this study aims to address and hopefully to resolve to some extent. One issue is that there are many different kinds of weights variables provided in large-scale datasets, which often would be a source of confusion to begin with. For instance, the Trends in International Mathematics and Science Study (TIMSS) 2011 provides many sample weight variables in this cross-sectional data set, which include: (1) weighting factor and weights nonresponse adjustment at each of school, classroom and student levels (i.e., WGTFAC1, WGTADJ1, WGTFAC2, WGTADJ2, WGTFAC3 and WGTADJ3); (2) school and student overall weights (i.e., SCHWGT and TOTWGT); (3) senate weights (i.e., SENWGT) and house weights (i.e., HOUWGT); (4) replicate weights (i.e., Jackknife zone and replicate code). Subsequently, selecting appropriate weights in a model analysis could be quite challenging when

currently there is a lack of practical guidance in data user's manual regarding how to apply these sampling weights appropriately from model perspective. For instance, in a single level model, it is unclear whether researchers should utilize the sampling weights based on the analysis unit level or the replicate weights to adjust standard errors as recommended by the data user's manual. Furthermore, in a typical two-level model (i.e., students and schools), it is very likely that analysts would encounter difficulty in choosing appropriate weights among those aforementioned.

Second, unlikely the unweighted analysis where computational application is pretty robust and consistent across different software programs, for sampling weights analysis, there is much variation especially for weighted multilevel models. Although statistical software is crucial to obtain reliable results, previous information about software applications in this respect has been very few plus outdated given the fast updating speed of statistical software. For example, it was mentioned that SAS, SPSS and R did not treat sampling weights correctly in the multilevel model analysis (Carle, 2009; Chantala, Blanchette, & Suchindran, 2006). However, currently, since the version of the SAS 13.1, the "PROC GLIMMIX" command allows users to incorporate sampling weights in two-level models (SAS Institute Inc., 2013). Therefore, it is not uncommon that educational researchers may have questions about whether their routine software programs support sample weighted model analysis. If not, how to choose appropriate statistical software and whether the results would vary when different software programs are utilized.

Third, there is no gold standard on weighted estimation methods and it is unclear which weighted estimation method(s) would be preferred and under what conditions. Particularly, when both weighted single- and multi-level models are feasible and the research interest is purely on the inference of the fixed effects (i.e., regression coefficients), it is unknown which approach

would be preferred. In the literature, the comparison of weighted single- and multi-level model under complex sampling design is very rare with the exception of one recent study (Koziol, Bovaird, & Suarez, 2017). Therefore, investigating when and how to incorporate sampling weights to model analysis using large-scale educational data via appropriate software programs is timely and essential.

1.2 Theoretical background and literature review

Theoretically, there is disagreement and inconsistency about applying sampling weights due to two fundamentally opposite schools of thought on making inference from survey data: the design-based approach and the model-based approach (D. A. Binder & Roberts, 2009; D. A. Binder & Roberts, 2003; Little, 2004; Rao & Bellhouse, 1990; T. F. M. Smith, 1984). The design-based (or randomization sampling) approach is traditionally adopted to conduct descriptive analysis of finite population quantities and produce design-unbiased estimates of population values (e.g., mean, ratio, and total). The assumption of this approach is that the estimates for the finite population are fixed and the uncertainty comes exclusively from sampling error. In the estimation of sampling variance, conventional procedures include Taylor series linearization, balanced repeated replication (BRR), jackknife repeated replication (JRR) and bootstrap (Cohen, Burt, & Jones, 1986; K. Rust, 1985; K. Rust, 2013; K. F. Rust & Rao, 1996). In contrast, the model-based approach is typically used to carry out analyses for statistical inference that produce estimates of coefficients, corresponding standard errors and confidence intervals for population relationships assuming data come from simple random sample (SRS). The emphasis has been put on specifying a correct model while ignoring sampling design or its effect could be controlled by including some design variables as model covariates. In theory, the

design-based and the model-based approaches are incompatible. However, in practice, data structures need to be considered in model analysis when the research interests go beyond knowing descriptive statistics while data come from complex sampling designs with special features rather than SRS. To deal with this challenge, currently, an increasing number of researchers have proposed and adopted a hybrid approach, which could be referred to as a design-model-based approach that combines these two approaches together.

Concerning the controversy, the focally debatable question when using large-scale data is "whether to weight or not to weight" (Bertolet, 2008; Kish, 1992; C. Skinner, 1994; T. M. F. Smith, 1988; Xia & Torian, 2013). Korn & Graubard found that weighted and unweighted estimators can be quite different in empirical data and they stated that unweighted estimators from sample data can be badly biased whereas weighted estimators are approximately unbiased (Korn & Graubard, 1995). Lohr & Liu also recommended weighted analyses although weights did not make a difference in their study (Lohr & Liu, 1994). However, other researchers (e.g., Winship & Radbill, 1994) suggested that unweighted estimators are preferred because they are unbiased, consistent, and have smaller standard errors than the weighted estimates when sampling weights are solely a function of the independent variable included in the model (Winship & Radbill, 1994). Overall, there is no consensus about the application of complex sampling weights on statistical models, especially for multilevel models (Graubard & Korn, 1996; Korn & Graubard, 2003; Pfeffermann, 1993, 2011).

The difficulty starts with the conceptual confusion of sampling weights for a couple of reasons. Firstly, weights have different or even opposite meanings in the design-based approach versus the model-based approach. Weights in the design-based approach are the inverse of the unequal probability of inclusion, so large weights represent a small selection probability, which

means we know less about these data. However, weights in the model-based approach are frequency weights, which have typically been used to correct for non-constant error variance in model analyses. Thus, large weights correspond to smaller error variances, which indicate that we know more about these data (e.g., weights used in meta-analysis). Secondly, weights can be referred to various things such as purely inverse-selection probability weights or weight components that include additional adjustment for non-response and post-stratification. Additionally, there are cluster weights (e.g., school level weights) and individual level weights (e.g., conditional probability weights or joint probability weights) in a two-level case (e.g., students nested within schools). Stapleton (2013) provided a nice summary review about different types of sampling weights appeared in large-scale educational data (Stapleton, 2013).

In addition, it is unknown whether there exists one preferred weighted estimation method that would work for all conditions. Researchers may be indecisive about whether to choose weighted single level versus weighted multilevel analysis. There are several key issues to be consider. First, it is necessary to verify that weights are designed to be used for multi-level analysis. When only single-level weights are available, the recommendation is to use weighted single-level analysis (Asparouhov, 2006). Second, if the research objective is to examine the level specific effects (e.g., teacher or school effects) or to decompose the variance to estimate the cluster level variance (e.g., school level variability), then the use of multi-level models is more appropriate. Third, when both single-level and multilevel sampling weights are all available, which is usually the case in large-scale education datasets and also when the research interest is just the fixed effect (i.e., regression coefficient and its standard error), currently, it is unclear which approach would outperform the other and under what conditions. One literature gap is that relevant empirical and simulation evidence of the performance of sample weighted estimation

methods under unequal probabilities of selection is very limited. To my knowledge, there is only one study that compared weighted single-level and multilevel models under simulation conditions of ICC (0.05 and 0.25) and cluster size (5 and 20). It was found that unweighted analyses for single and multilevel model generated similar estimates across various conditions while weighted single-level models had a better performance than weighted multi-level models when the design is informative (see Koziol et al., 2017). Nevertheless, when working with real large-scale education data with particular design features, it is still unknown which approach would have a better performance.

From a model-based perspective in single-level models, the pseudo maximum likelihood (PML) has been utilized as a relatively well-established approach to deal with unequal probability of selection and produce consistent estimates (D. A. Binder, 1983; Krieger & Pfeffermann, 1992; C. J. Skinner, 1989). In multi-level models, Graubard and Korn (1996) proposed the weighted ANOVA estimators. However, their approach is limited to a specific simple model and without support of software application (Graubard & Korn, 1996; Jia, Stokes, Harris, & Wang, 2011; Korn & Graubard, 2003). In addition, two general methods have been proposed to produce possibly the least biased estimates. Rabe-Hesketh and Skrondal, and Asparouhov (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006) have concurrently proposed the estimation method of the Multilevel Pseudo Maximum Likelihood (MPML) based on the PML, while Pfeffermann et al. have proposed the Probability Weighted Iterative Generalized Least Square (PWIGLS) method based on the iterative generalized least squares (IGLS) method introduced by Goldstein in 1986 (Goldstein, 1986; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). Compared with PML and MPML, the IGLS algorithm involves iteration between estimation of fixed and random effects and in the PWIGLS the population quantities are

replaced by weighted sample statistics. These two general estimation methods have been adopted in different software packages. Particularly, MPML has been adopted in STATA, Mplus and SAS while PWIGLS is implemented in LISREL, HLM and MLwiN.

In general, it is unknown which estimation method should be preferred because the performance is related to several factors such as design informativeness (i.e., the degree to which the outcome variable is related to sample selection probabilities conditioning on model covariates), cluster sample size, variability of sampling weights, intra-class correlation (ICC) and scaling methods (Asparouhov, 2006; Bertolet, 2008; Cai, 2013; Grilli & Pratesi, 2004; Kovačević & Rai, 2003; Pfeffermann, Skinner, et al., 1998; Rabe-Hesketh & Skrondal, 2006).

Specifically, previous research studies have shown that first of all design informativeness is of the utmost importance to decide whether sampling weights should be used. An informative design is a prerequisite for weighted analysis to account for the difference between the sample distribution and the population distribution due to unequal probabilities of selection.

With regard to the informativeness, it is a model concept (Binder et al, 2005). The term "informativeness" and "ignorability" have been used loosely to refer to the interaction between model and survey design when explaining the possible effect of ignoring the sample design features in model analysis on survey data, but a formal definition was not provided. Pfeffermann (1993) referred the sampling design is "ignorable" when data is selected from simple random sample and "informative" when using proportional probability of selection for the sample data. Binder et al (2015) mentioned that the sampling is informative when the distribution of the sampled unit is different from that in the population and otherwise it would be non-informative. In general, informativeness and ignorability have much in common. It is not uncommon that the design is referred to as "informative" or "non-informative which is used interchangeably with

"non-ignorable" or "ignorable" without clear differentiation and clarification. Nevertheless, there could be a slight difference in terms of meaning. Binder and Roberts (2001) indicated that non-informative design leads to ignorability, but not vice versa. That is, being non-informative is sufficient for ensuring ignorablility while being ignorable cannot determine non-informativeness.

Conceptually, informativeness is not difficult to understand. However, in practice, it is quite difficult to determine whether sample design is informative for a particular model analysis. It is even more challenging to quantify the magnitude of informativeness. Therefore, although design informativeness is a key factor of deciding whether to weight or not to weight, traditional practice still plays a dominant role in a discipline. For example, in biostatistics and public health, researchers use weights whereas in social sciences (e.g., econometrics) researchers do not apply weights in general (Bollen et al, 2016). In the literature, there are some simple diagnostic tests (e.g., t-test and chi-square test) which often ignore possible effects clustering and unequal probability of selection have been used to determine the effect of weighting. More research of the diagnostic tests on design informativeness is needed to provide scientific criterion about whether weights would be necessary for a particular model analysis.

Second, what is also important to parameter estimation is which sampling stage is informative (Cai, 2013; Pfeffermann, Skinner, et al., 1998). Third, simulation studies showed that bias is associated with small ICC values and cluster sizes (Asparouhov, 2006; Jia et al., 2011). Fourth, although the scaling of the lower level sampling weights has been regarded as the primary tool for bias reduction, there is a lack of agreement about the best scaling method (Pfeffermann, Skinner, et al., 1998; Stapleton, 2002). Additionally, even with rescaled weights, survey weighted estimators could still be grossly biased for estimating variance components (Korn & Graubard, 2003).

The literature review reveals several practical issues that need to be addressed concerning the incorporation of complex sampling weights in model analysis when using large-scale educational datasets. First, findings from previous simulation studies are likely to be inapplicable to real large-scale educational data because it is not uncommon that they were conducted with small-scale data or under extreme simulation conditions. For instance, the sampling design in large-scale data would have design informativeness appeared at different sampling stages rather than either overall informative or non-informative design appeared in past simulation studies (Koziol et al., 2017). In addition, large-scale data typically have decent sample sizes which may prevent it from suffering the possible substantial bias attributed to small sample bias. Second, to educational researchers who typically have interest in knowing the statistical inference of particular variables (e.g., teacher, classroom and school variables), it is indecisive whether to use weighted single level or multilevel analysis and it is unclear how software programs would support specific model analyses. By and large, the current literature fails to provide sufficient information in this respect.

1.3 The present study

This study aims to examine when and how to apply complex sampling weights of large-scale educational data in single- and multi-level models via empirical and simulation investigation. It will shed some new light on the practical guidance concerning the incorporation of sampling weights in model analysis along with corresponding software usage. Specifically, there are three research questions:

(Q1) When and how to apply sampling weights in single- and multi-level models using large-scale educational data and what would be the appropriate software programs to use?

(Q2) What factors (e.g., informative index and design effect) might be associated with potential divergent results between weighted and unweighted for some variables?

(Q3) Which statistical model (i.e., unweighted single, unweighted multi-level model, weighted single and weighted multiple) would have a better performance?

This thesis consists of two components: an empirical examination to address the research questions (1) and (2) and a simulation investigation for question (3). The empirical component will demonstrate the use of complex sampling weights via small practical examples using PISA U.S. data in 2012 and Early Childhood Longitudinal Study-Kindergarten Class of 2010-11 (ECLS-K 2011). Meanwhile it explores how to conduct the analysis using various software programs, which include STATA 14, Mplus 7, SAS 9.4, LISREL 9.30 and HLM 7 and for each software, user's manuals or online resources have been served as useful guidance (Muthén & Muthén, 2010; Randenbush, Bryk, Cheong, Congdon, & Toit, 2011; SAS Institute Inc., 2013; Scientific Software International, 2005-2012; StataCorp, 2013; Zhu, 2014). HLM instead of MLwiN is chosen because both are special software for multilevel modeling but HLM is widely used among researchers in the U.S. while MLwiN may be more popular in UK or in Europe. The results of different models with or without weights across varied software tools will be compared and divergent findings will be examined and related to possible reasons. Mimicking the real sampling design in ECLS-K where fixing the student-level design informativeness while varying that at the school level, the simulation component will focus on examining the performance of PML and MPML in single and multilevel models and compared with their unweighted counterparts to determine which method would generate the least biased estimates for the fixed effects.

The contribution of this study is two-fold. In practice, it discusses and summarizes the controversies and challenges of applying sampling weights in the context of analyzing large-scale educational data, which helps to clarify whether and when it is appropriate to incorporate complex sampling weights. More importantly, it demonstrates how to apply complex sampling weights using large-scale data sets with illustrative empirical examples. Additionally, this research will provide an informative update about software development on sample-weighted model analysis. This addresses an urgent need of many educational researchers, who may want to be informed about the capabilities of various software programs on incorporating sampling weights in their practical research. In theory, the mechanism and underlying reasons for divergent results are unknown although it is not uncommon that weighted inference sometimes would differ from unweighted case. This research tries to link divergent results to some possible factors. Furthermore, the simulation component of this research will advance the methodological knowledge about the performance of the weighted estimation in single- versus multi-level models in the context of large-scale educational data.

1.4 Data and Methods

1.4.1 Empirical data and variables

This thesis will utilize U.S. data for PISA 2012 and ECLS-K 2011. The PISA study was implemented by OECD to provide the assessment of academic achievement of the 15-year-old on mathematics, science and reading literacy. In addition, information about students learning environment, educational experiences and attitudes towards education has been collected. PISA started to collect data cycles every three years since 2000. PISA 2012 is a recent data with information about sixty-five countries and economies participated. In general, the sampling

design involves a two-stage stratified sampling design where the first-stage sampling units consisted of individual schools that have targeted 15-year-old students with probabilities proportional to size and the second-stage sampling units were students selected with equal probabilities within sampled schools (OECD, 2012).

ECLS-K 2011 is the most recent longitudinal study which follows a U.S. national representative sample of kindergarten students of diverse socioeconomic and ethnic backgrounds from kindergarten through early elementary grades. ECLS-K provides information regarding children's early school experience. Data have been collected to study how students' cognitive, social and emotional development is related to various family, classroom and school environments that students have been exposed to. The ECLS-K has adopted a three-stage stratified sampling strategy in which 90 geographic regions serve as the primary sampling units (PSUs). Then, samples of public and private schools with 5-year-old children were collected within sampled PSUs with probabilities proportional to measures of population size at first and second sampling stages. Finally, students were randomly selected within sampled schools (Tourangeau et al., 2015).

Table 1 below presents the variables. In PISA, the outcome is the math achievement and the covariates include gender, economic social and cultural status (ESCS), father full-time work status, school sector, and school ESCS. Final student and school weights are "W_FSTUWT" and "W_FSCHWT" respectively. There are 4978 students in 162 schools and the average number of students per school is 31 approximately. In ECLS-K, the kindergarten spring data is used in which the outcome measure is children's reading gain scores and the independent variables include students' gender, race, SES, school location, sector, enrollment as well as free and reduced lunch. The variable named W2SCH0 is the school level base weight adjusted for

nonresponse and the W12AC is the child base weight adjusted for nonresponse associated with the spring kindergarten teacher-level questionnaire and the fall kindergarten child assessment. The sample size is 10349 students in 678 schools, so about 15 students per school were selected on average. It should be noted that these commonly used variables at both the student and school levels are selected as illustrative examples to study the sample weights issue, so measurement error is ignored and missing data problem is replaced with median values for continuous variables and with zero for binary variables.

1.4.2 Statistical models

The single-level model for the i^{th} student can be expressed as:

$$Y_i = \beta_0 + \boldsymbol{COV_i B_1} + e_i \quad e_i \sim N(0, {\sigma_e}^2) \tag{1.1}$$

where Y is the outcome variable, $\boldsymbol{COV}$ refers to a row vector of covariates at student and school levels as listed in Table 1, Greek letter $\boldsymbol{B_1}$ represents a column vector of covariate coefficients, e is the residual terms which is assumed to follow normal distribution with mean zero and constant variance ${\sigma_e}^2$.

For simplicity, a two-level random intercept model is used to represent multilevel model in which individual student i in school j can be written as:

$$Y_{ij} = \beta_0 + \boldsymbol{COV_{(i)j} B_1} + u_j + \varepsilon_{ij};\ u_j \sim N(0, {\sigma_u}^2) \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \tag{1.2}$$

where the variance consists of two components: the variance (${\sigma_u}^2$) of the school random effect u and the variance (σ_ε^2) of the student random effect ε. Both errors u and ε follow normal distributions with zero means and variances σ_u^2 and ${\sigma_\varepsilon}^2$ respectively. Additionally, the errors at different levels are assumed to be independent of each other. All the other terms have been defined as in equation (1).

Comparing equations (1) and (2), it is clear that the difference between single- and multi-level model lies in the residual term. Specifically, the residual term *e* in the single level model is partitioned into a cluster residual u and an individual residual ε in the multi-level model, The estimation of the variance component of u and the variance ratio of σ_u^2 over ($\sigma_u^2+\sigma_\varepsilon^2$)(i.e., ICC value) are of particular interest.

1.4.3 Sample weighted estimation method

Concerning the estimation methods, the basic idea of the PML method is that assuming observations are independent to each other the population (or census) likelihood then can be obtained via a multiplication of the sample weighted likelihood. Since the census likelihood would produce consistent and unbiased estimates, sample-weighed estimates from PML should also be consistent and unbiased. However, regarding the MPML method in multi-level models, observations are dependent within each cluster, so cluster effects have to be integrated out before applying the PML. Another difference from PML is that the sampling weights at the lower level cannot be used as is and scaling needs to be done to reduce bias. Nevertheless, like PML, MPML could be defined as a general estimator that can be obtained via any optimization algorithm such as the EM-algorithm, the accelerated EM algorithm and the Quasi-Newton algorithm. Typically, there are no closed form solutions for estimated parameters in MPML, so approximation approaches need to be adopted in which the estimation varies depending on the algorithms and scaling procedures implemented in different software programs.

Let $\theta_1=(B,\sigma_e^2)$ and $\theta_2=(B,\sigma_u^2,\sigma_\varepsilon^2)$ represent parameters in single- and multi-level models respectively and the research interest in this study would be focused on estimating the fixed effects B, the regression coefficients, and the corresponding statistical inference based on the

estimation of error variance. The likelihood function can be expressed as in equations (3) and (4) below for a single-level and a multi-level model respectively. The subscript i is for individual unit and j is for cluster unit, which would be the default case throughout this thesis.

$$L(\hat{\theta}) = \prod_i f(y_i \,|X, B, \sigma_e^2) \tag{1.3}$$

$$L(\hat{\theta}) = \prod_j L_j = \prod_j [\int (\prod_i f(y_{ij} \,|X, B, u_j, \sigma_\varepsilon^2) \phi(u_j | \sigma_u^2) du_j] \tag{1.4}$$

Suppose data was sampled with unequal probability at each of the two sampling stages. The probability of selection at the first stage is p_j and the conditional probability of selection at the second stage is $p_{i|j}$. The corresponding weights at the first and second stages are w_j and $w_{i|j}$ respectively. The overall probability of sampling selection is p_{ij} , which is the multiplication of $p_{i|j}$ and p_j, and the overall sampling weights are w_{ij} (=$w_{i|j}$ x w_j), which is referred to as w_i in a single-level case. In corporation of sample weights, the likelihood functions become:

$$L(\hat{\theta}) = \prod_i f(y_i \,|X, B, \sigma_e^2)^{w_i} \tag{1.5}$$

$$L(\hat{\theta}) = \prod_j [\int \left(\prod_i f(y_{ij}|x, \beta, u_j, \sigma_\varepsilon^{2\,2})^{w_{i|j}}\right) \phi(u_j | \sigma_u^2) du_j]^{w_j} \tag{1.6}$$

Previous studies show that weighted multi-level procedures would produce substantial bias for estimating variance components especially for small cluster sample size. Therefore, scaling the individual-level weights is recommended as a primary tool of bias reduction, which is represented by λ_1 in equation (7) below.

$$L(\hat{\theta}) = \prod_j [\int (\prod_i f(y_{ij}|x, \beta, u_j, \sigma_\varepsilon^2)^{\lambda_1 w_{i|j}}) \phi(u_j | \sigma_u^2) du_j]^{w_j} \tag{1.7}$$

Two scaling methods have been proposed and received much attention in the literature (see Pfeffermann et at, 1998). The first scaling method scales the level-1 weights to the actual cluster sample size (n_j), which is referred to as “size” scaling. The second one scales the level-1 weights to its effective cluster size, which is referred to as “effective” scaling. Here the name

follows the terms of scaling command in STATA software. They are defined as in the equations (8) and (9) respectively.

$$\lambda_{1_{Size}} = \frac{n_j}{\sum_{i=1}^{n_j} w_{i|j}}, \tag{1.8}$$

$$\lambda_{1Effective} = \frac{\sum_{i=1}^{n_j} w_{i|j}}{\sum_{i=1}^{n_j} {w_{i|j}}^2}. \tag{1.9}$$

The motivation of the size scaling method is to represent the number of elements in a cluster to reduce bias. The motivation for the effective scaling may be to control for the design effect, which is the ratio of the variance with the sampling design over that with simple random sample. Since the variance is the second moment, the weights have a square term. There is no consensus about which scaling method works better and under what conditions. For example, Pfeffermann et al (1998) tentatively recommended the size scaling in their simulation study to reduce bias caused by informative sampling while Stapleton (2002) found that effective scaling provides unbiased estimates in multilevel SEM analyses. Asparouhov (2006) pointed out that different scaling methods may have different effect on different estimation techniques (p442). Previous study also suggested there are many factors (e.g., design informativeness, type of outcome, sample size) would affect scaling methods results separately or jointly. In general, simulation work suggested that when the interest is the point estimates use size scaling whereas for estimating cluster variance effective scaling would be preferred (Asparouhov, 2006; Carle, 2009). Relatively speaking, size scaling method is straightforward, so it has been used more often.

With regard to the asymptotic covariance matrix, the classical sandwich form has been adopted in both single and multi-level models, in which the latter can be expressed as

$$(l'')^{-1}(\sum_j (\lambda_2 w_j)^2 l_j' \, l_j'^T)(l'')^{-1} \tag{1.10}$$

where l' and l'' refer to the first and second derivatives of the log-likelihood.

1.4.4 Simulation setting

To evaluate the performance of the fixed effect estimators in a two-level case, a Monte Carlo simulation study is conducted when mimicking the sampling design in ECLS-K 2011. Specifically, the ECLS-K data sampled about 18,200 kindergarteners from 970 schools and on average 19 students per school. The school and conditional student selection probabilities are about 0.02 (p_j) and 0.25 ($p_{i|j}$) respectively and the overall student selection rate is about 0.005 (p_{ij}) (Mulligan, Hastedt, & McCarroll, 2012). It should be mentioned that although the data follows a three-stage sample design, the first sampling stage of geographic units was ignored and only school and student level sampling weights were provided. Using five public large-scale data sets from NCES, Stapleton and Kang found that ignoring the sampling design beyond the levels in the model would have minor effects on inference (Stapleton & Kang, 2016). Based on real data sample selection rate, the simulation sets a population of 50,000 schools and 4,000,000 students.

The finite population values Y_{ij} are generated from the following model:

$$Y_{ij} = 1 + 0.019 * female - 0.065 * SES - 0.001 * class\ size - 0.026 * private\ school\ + u_j + \varepsilon_{ij} \quad (1.11)$$

where female follows a Bernoulli distribution with probability of 0.49, SES follows a normal distribution with mean 0.06 and variance 0.80, class size follows a normal distribution with mean 20 and variance 16, private school follows a Bernoulli with probability of 0.16. Finally, ${\sigma_u}^2$ is set at 0.0625 (i.e. $\sigma_u = 0.25$), and σ_ε^2 at 0.25 (i.e.$\sigma_\varepsilon = 0.50$), and the ICC value is 0.20, which is a typical school level variability for U.S. data sets. All the coefficient values are determined according to real estimates using the data.

Following Asparouhov and Muthen (2006), the infinite target population approach was adopted to generate the samples (Asparouhov & Muthen, 2006). The index I equals one indicating that the observation is selected. The selection probability at cluster level is

$$prob_j(I=1) = \frac{1}{4+exp(1-\frac{u_j}{X})} \quad (1.12)$$

where the value of *X* varies in five numbers (i.e., 1/3, 1/2, 1, 2, and 3) with the mean is all around 0.02 while the range or variation narrows down gradually to represent the design informativeness decreases noted by I(1), I(2), I(3), I(4) and I(5) respectively. The selection probability at the individual level is

$$prob_{ij}(I=1) = \frac{1}{1.35+exp(1-\frac{\varepsilon_{ij}}{2})} \quad (1.13)$$

One finite population of 40,000,000 was generated according to equation (11) with preset parameter values and then five samples were generated based on the selection probabilities in equations (12) and (13) in which the school level sampling informativeness varies. For each sample, four analytic models were run: (1) a single level unweighted model with cluster robust standard error; (2) an unweighted two-level random intercept model; (3) a weighted single level model with overall weights; (4) a weighted two-level random intercept model with level-specific weights. The simulation procedure was repeated 1000 times via the software of STATA 14.

The quality of estimates were assessed via three criteria: (A) empirical relative bias; (B) empirical root mean square error (MSE) and (C) coverage rate for true parameter falls within the 95% confidence interval with t-test based standard errors, which have been used in previous simulation studies (Cai, 2013; Eideh & Nathan, 2009).

Specifically, the relative bias is defined as:

$$\text{RBias} = \frac{1}{\theta}\left[\frac{1}{1000}\sum_{t=1}^{1000}\left(\widehat{\theta_t} - \theta\right)\right] \quad (1.14)$$

Where θ is the true value and $\widehat{\theta_t}$ is the estimated value in each iteration.

The root MSE is expressed as:

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{[\frac{1}{1000}\sum_{t=1}^{1000}(\widehat{\theta_t}-\bar{\bar{\theta}})^2]} \quad (1.15)$$

where $\bar{\bar{\theta}} = \frac{1}{1000}\sum_{t=1}^{1000}\widehat{\theta_t}$

1.4.5 Software programs

Research on multilevel modeling in the context of complex sampling is quite recent, so its application on software programs are developing at different stages. Knowing the computational capability in this regard is of crucial importance because there is an increasing body of researchers using these analytical software programs for weighted multilevel model analysis, but they may not have a good understanding of the capability of their software in terms of incorporating sampling weights and performing weights scaling in this respect. For example, sampling weights used for multilevel model analysis would be different from single-level analysis. Particularly, analysts need to know how to construct weights scaling appropriately for individual level sampling weights to avoid potential estimates bias.

The specific contribution of this part is twofold. First, data analysts and practitioners may not be interested in the estimation quality of particular parameters, but it may be of their great interest to know whether their routinely used statistical program could handle sampling weights and scaling and also to know whether the results from their program would be different from other programs under the same conditions. Second, software programs are developing by leaps and bounds. What was known about the software application in multilevel modeling with sampling weights has been dated and incorrect, so providing an update would be needed. For instance, the scaling method has been applied differently in multilevel software programs.

Chantala & Suchindran (2006) summarized the software implementation of sampling weights in which Mplus 4.1 does not provide automatic scaling option so users have to do scaling of sampling weights by themselves, but it is no longer the case in Mplus 7 (Chantala & Suchindran, 2006).

This analysis attempts to study different statistical programs that have been the conventional tools to conduct multilevel model analysis. To my knowledge, SPSS and R have not provided technical support for complex sampling weights yet. Specifically, the MIXED command in SPSS provides a residual weights option, which is similar to the weights option in SAS PROC MIXED command. However, this option supports frequency weights but not sampling weights. This creates two issues. First, level-specific sampling weights cannot be fitted into the command and it would be inappropriate to choose either level 2 weights or level 1 weights. The second issue is that although sampling weights have been computed as replication weights (i.e., frequency weights) for the point estimates, the computation of the standard errors has not been adjusted. This study will focus on the following five statistical software programs that have been used by a large body of researchers: STATA, Mplus, HLM, SAS, and LISREL.

The software research component has some special features. First, this study provides real data analysis while some previous studies only presented the descriptive introduction of the software capability (West & Galecki, 2011). Second, this study provides a latest comparison among five software programs and six commands that have the capability of dealing with sampling weights in multilevel modeling, which is an improvement from previous work. For example Chantala & Suchindran (2006) compared the performance of four programs while Carle (2009) conducted analysis using three programs (Carle, 2009; Chantala & Suchindran, 2006).

a) Stata

In Stata 14, two commands are available to carry out the computation. In MIXED command (StataCorp, 2013), sampling weights can be specified at each level and scaling options of individual-level weights are available by specifying "pwscale" as "size" or "effective". For the MIXED commend, weighted estimation is achieved via incorporating w_j and $w_{i|j}$ into the matrix using the orthogonal-triangular (QR) decomposition and then maximizing the profile log-likelihood to obtain ML parameter estimates. Basically, the estimation is based on replicate clusters and individual observations, which is similar to the way of handling frequency weights but on a multilevel platform. The second program is "GLLAMM" (i.e., generalized linear latent and mixed models), which was developed by Rabe-Hesketh and Skrondal (Rabe-Hesketh, Skrondal, & Pickles, 2004a, 2004b; Skrondal & Rabe-Hesketh, 2003). The computational technique involves adaptive quadrature which was proposed to be more reliable and efficient than ordinary quadrature (Rabe-Hesketh, Skrondal, & Pickles, 2002, 2005) .

Comparing the command of MIXED and GLLAMM, the first difference is that the former is used only for linear models while the latter is in general used for non-linear models. GLLAMM could also be used for linear models when the conditional distribution is Gaussian, the link is identity and the marginal distribution of the random effects is Gaussian, which is the case in MIXED (Grilli & Pratesi, 2004). The second difference is in computation speed, that is, the GLLAMM is much slower than the MIXED command. The third difference is that the GLLAMM does not provide an automatic scaling option and users have to manually operate the weights scaling procedure (Rabe-Hesketh & Skrondal, 2006).

b) Mplus

The Mplus manual (Muthén & Muthén, 2010) provides a guidance on how to incorporate sampling weights and sampling scaling in a two-level model. Specifically, the individual level

sampling weights can be defined using “WEIGHT” and the cluster level weights using “BWEIGHT”, which represent the between-level sampling weights. For the individual level weights, there are three scaling options: “UNSCALED”, “CLUSTER” and “ECLUSTER”. The latter two are counterparts to “size” and “effective” scaling in STATA. The cluster level weights can be defined as “UNSCALED” and “SAMPLE”, but scaling is not carried out. In a three-level model, two upper level weights can be defined by “B2WEIGHT” and “B3WEIGHT”. However, it is not known how the scaling at two upper levels could be implemented.

c) SAS

In the SAS 13.1 user’s guide (Zhu, 2014), analysts are informed that weighted multilevel model for survey data can be conducted via the procedure “PROC GLIMMIX” by using the “OBSWEIGHT” and “WEIGHT” to represent level 1 and level 2 sampling weights respectively. SAS also works in the framework of MPML rather than PWIGLS and the “PROC GLIMMIX” procedure approximates the marginal log likelihood with an adaptive Gauss-Hermite Quadrature. To fit a model, the “METHOD=QUAD” and “EMPIRICAL=CLASSICAL” options need to be specified so that empirical (sandwich) variance estimators for the fixed effects and the variance components will be computed. SAS has not provided an automatic scaling option yet, so analysts are responsible for scaling the individual sample weights. Similar to other software, instruction on sampling weights in a three-level model is not available yet.

d) HLM

HLM user’s manual (Randenbush et al., 2011) informs the reader that design weights can be incorporated in HLM 7 based on the method in Pfeffermann et al. (1998), which is more appropriate than the method used in earlier versions of HLM. In a two-level model in school setting, suppose only student level or school level weights are available, HLM 7 will normalize

the student level or school level weights to a mean of 1. If both level weights are available, what HLM 7 does is to normalize the level 1 weights within level 2 units, so that the sum of the level 1 weights within a level 2 unit will be the cluster sample size n_j, which is for example the sample size of students within school j. With regard to the three-level model, it is unknown how the scaling will be implemented.

e) LISREL

In LISREL 9, sampling weights in either two-level or three-level models can be incorporated. However, the scaling option is not available as LISREL automatically applies cluster sampling weights for both levels only for the case of two-level model (Cai, 2013; Scientific Software International, 2005-2012).

1.5 Results

1.5.1 Research question one

In this section, the U.S. PISA data were used. In order to compare the performance of sample weights analysis across these software programs, it would be helpful to examine any differences beforehand. Table 1.2 presents the results of unweighted analyses. The results show among the output generated from STATA, Mplus and SAS, it can be observed that the estimates of fixed effects between single and multilevel models are a little different although the statistical inference remains the same. Third, the results are consistent across these five software programs for multi-level models.

Table 1.3 contains results of the weighted analyses. In the weighted analyses, in multi-level models, there are various scaling options. The “mixed” command in STATA provides the option of “size” and “effective” scaling which are corresponding to “cluster” and “ecluster” in

Mplus. In SAS, it is user's responsibility to do the scaling before employing the "proc glimmix" command. The LISREL and HLM automatically perform "size" scaling.

The results show that first STATA, Mplus and SAS are more capable in terms of providing a couple of scaling methods while LISREL and HLM only allows "size" scaling. Second, the statistical inference remains the same between the method of MPML employed in STATA, Mplus and SAS and the method of PWIGLS in LISREL and HLM although the estimates of the coefficients and standard errors have a slight difference. Third, the two scaling methods generated almost identical results although they are a little different from the unscaled weighted estimates. Fourth, "size" scaled estimates from MPML and PWIGLS are similar by and large, and within each estimation method, the results are identical across different software programs.

Overall, the statistical inference of weighted analyses is consistent for all the variables except for the variable of father's working status in which estimates from weighted single level models agree with weighted unscaled multi-levels, but they are different from the estimates in weighted scaled models.

It is noteworthy that sampling weights should be used for whole sample analyses, but not for subgroup analyses (e.g., gender and race) because the weights of subgroups no longer represent the observations in the population. Moreover, the scaling of the student level weights in multilevel models does not make sense in the subgroup analysis. However, for the purposes of software exploration, regardless of the inappropriateness, sample weights for subgroup analyses by gender and by school sector were also conducted. It was found that all the software packages run the analyses by selecting the target subgroup observations while deleting unused observations, among which STATA, Mplus and SAS still produce consistent output. LISREL

also produces output, but HLM failed to generate any results. Tables 1.4 and 1.5 present the software output by gender and school sector.

1.5.2 Research question two

In this section, ECLS-K 2011 data were used to explore how the results of statistical inference might be different using weighted and unweighted analyses in both single- and multilevel models given that the data have fixed ICC values and cluster sample sizes. An additional research interest is to identify factors that might be associated with producing divergent results for some variables but not for others. Specifically, four model analyses were conducted: (a) single-level unweighted model with robust standard errors; (b) unweighted multilevel model; (c) weighted single-level model; (d) weighted multi-level model.

Three indices were used to determine the degree to which design informativeness might affect the results: (i) informative index (I) based on the rationale that mean values are very sensitive to informative designs (Asparouhov, 2006) (see page 446 equation 6). Specifically, the design informative indices of I(A), I(C) and I(B) were computed based on overall student weights (A) or student specific (C) and school level weights (B). For each index the numerator is the absolute value of the difference between weighted and unweighted means and the denominator is the standard error of the unweighted mean. (ii) root design effect (RDE), that is, the square root of the design effect, which was used to measure the ratio of the actual variance of an estimate with the sampling design to the variance of the estimate with SRS, so the value of RDE above 1 and below 1 suggest that the sampling design leads to larger or smaller standard error respectively (Kish, 1967; Stapleton & Kang, 2016). Here RDE(A) was computed as the ratio of the standard error of each covariate in a single-level weighted model over that of SRS

and RDE(B) was the ratio of standard error in weighted multilevel model over that in SRS. (iii) the correlation between the covariates and the overall student weights (A), student level specific weights (C) and school level specific weights (B).

These results in Table 1.6. shows that the informative index may be linked with divergent results, but the indices of the design effect and the correlation do not seem to have detecting power. Therefore, when the informative indices are large, it is more likely to observe divergent results between weighted and unweighted analysis, but this is not true for all cases such as the variable of private school.

1.5.3 Research question three

By trying to vary the design informativeness at the school sampling stage while fixing the ICC value at 0.2, this simulation study evaluated the performance of an unweighted single level model, an unweighted multi-level model, a weighted single level model and a weighted multi-level model in terms of the estimation on the fixed effects of four variables (i.e., female and SES at the student level and class size and private school at the school level). Specifically, the quality of the parameter estimates was assessed via the empirical relative bias, the RMSE and the 95% coverage rate which were presented in the Figures 1, 2 & 3.

The figures show that first of all, the relative estimation bias varied greatly for different variables of interest regardless of degree of design informativeness and estimated models. In general, there is no clear pattern about which model should be preferred. Second, with regard to the RMSE, unweighted multilevel model consistently yielded smallest values with no exception. Weighted single level model has the largest value for student level variables. Third, in terms of coverage rate, all values are between 94% to 96% which are close to 95% with a slight variation.

To test whether design informativeness affect the fixed effects estimates, I also conducted a simulation which treats the selection probability at cluster level as a linear function of the residual rather than the exponential function that has been used. The results show that there is no clear pattern. This investigation illustrates the point that there is an evident gap between the sample selection in the simulation and in real data. In the simulation, the sample selection probability only depends on the relation with the residual, so it is very likely to form a trend pattern for the residual parameter therefore, but the selection probability does not seem to correlate with covariates. However, in real data, the sample selection probability is very likely to be related to some demographic variables such as gender, race and school geographic location that can be regarded as design variables for determining the way of sample selection. Nevertheless, it should be noted that in practice it is also difficult to include these covariates in the simulate setting because they may be too specific and too varying to have a general simulation form.

Following Pfeffermann et al. (1998), the standard errors of the estimates were examined by comparing the average value of the 1000 estimated standard errors with the standard deviation of the 1000 point estimates for each fixed-effect coefficient. The closeness of these two indices indicates good quality of the standard error estimation because by definition the standard error of the point estimate is the standard deviation of the sampling distribution of the point estimator. Table 1.7 contains results which show that the estimation of the standard errors performs very well. It is evident that the unweighted multi-level models have the smallest standard errors for the point estimates across the four variables and the five informative designs.

1.6 Discussion

Data collected via complex sampling designs typically have some special features such as unequal probabilities of selection, clustering, stratification and nonresponse (D. Binder & Roberts, 2006).

Here is a simple example to illustrate how the unequal probability of selection at one sampling stage may introduce bias and how the incorporation of sampling weights might help to reduce the bias from a design-based perspective. Suppose the population consists of 20 students with ten black (60, 60, 60, 65, 65,65, 70, 70, 70, 75) and ten white students (75, 75, 75, 80, 80, 80, 85, 85, 85, 90), and their math scores are included in the parentheses. Suppose two students (i.e., 1st and 6th) are selected from the first group and four students (i.e., 1st, 4th, 7th, 9th) are selected from the second group. Then the probability of selection is 0.20 = 2/10 for the black students and 0.4 = 4/10) for the white students. The corresponding weights are 5 = 1/0.2 and 2.5 =1/0.4 respectively. As sampling weights reflect the number of units in the population given sample observations, here each black and white student in the sample represents five black students and 2.5 white students respectively in the population. Suppose the research interest is to estimate the average math score in the population. Below is the computation for the sample mean, sample weighted mean and population mean.

Sample mean = (60+65+75+80+85+85)/6=75

Sample weighted mean = (5*(60+65)+2.5*(75+80+85+85))/20 ≈ 72

Population mean = ((60*3+65*3+70*3+75) + (75*3+80*3+85*3+90))/20=73.5

As black students have lower scores than white students on average, the unweighted descriptive statistics using the sample data would produce an upward-biased population mean due to the disproportional selection of more white students who have higher achievement score.

However, the incorporation of sampling weights helps to reduce the upward bias. In this demo example with very small sample size, the result seems to have a downward bias, but with large sample size the weighted estimate would yield unbiased population mean.

It is noteworthy that in design-based analysis in terms of generating finite population quantity, sampling weights are essential, but in model-based analysis, it is unclear whether and when sampling weights should be incorporated in model especially for multilevel model to take care of potential design effects arising from unequal probability of selection. The argument involves the trade-off between bias and efficiency in estimation. When the sampling design is informative, which means the sample distribution would be different from population, estimates could be substantially biased if the effect of unequal probability selection is ignored. However, applying sampling weights when unnecessary would create inefficient estimator with larger standard error.

In single-level model, researcher can use either design-based approach or model-based approach to account for sampling design effects. The former includes the Tayler linearization (i.e., delta method), replication method (e.g., balanced repeated replication and jackknife repeated replication) and bootstrap. The latter conventionally uses overall sampling weights to compute appropriate error variance. For both approach, various software programs are available for implementing and generating relatively consistent results. Nevertheless, in multilevel model, the design-based approach is not compatible and for the model-based approach, individual-level sampling weights cannot be used as is. Instead, scaling procedure needs to be employed. However, different scaling methods may have different effects depending on the specific estimation technique (Asparouhov, 2006). In addition, some software programs start to apply different estimation techniques as well as scaling procedure but the development level varies

greatly. For example, SPSS and R have not supported sample weighted multilevel model yet while STATA allows automatic scaling options for weighted two-level model.

It is worth noting that incorporating sampling weights itself will not account for the fact that the data were collected from complex sampling designs because it is extremely difficulty to evaluate the overall impact of the complex sampling designs on the statistical inference which may arise from each or combined aforementioned design features (Pfeffermann, 2011).

Previous studies have provided some discussion about the informative probability sampling (D. A. Binder, Kovacevic, & Roberts, 2005; D. A. Binder & Roberts, 2001; Pfeffermann, Krieger, & Rinott, 1998; Sugden & Smith, 1984), which is closely related to the debate about whether sampling weights should be incorporated in analyses. Some researchers support the notion that in non-informative designs, weights are not necessary as they would increase the error variance. However, when the sampling design is informative, sample weights should be considered so that the sample distribution would resemble that in the population. Therefore, in the presence of some unknown informative design, incorporating sampling weights could at least deal with the impact arising from the unequal probability selection and could yield consistent estimates (Pfeffermann, 1996). In terms of the modeling framework, multilevel models might be preferred because they naturally match the hierarchical data structure arising from multi-stage sampling and take into account the clustering effect. The advantage of single level models is that they could flexibly adjust the standard errors due to design features either in the model based approach (i.e., PLM) or in the design based approach using classical survey variance estimators calculated typically at the PSU level. Therefore, it is unclear whether one should choose weighted single-level models or weighted multi-level models to conduct analyses. This is a practical issue that many educational researchers may feel confused about alongside

some other questions. For example, whether sampling weights should be incorporated in the model analysis and also for subgroup analysis, and how to use appropriate software programs to carry out specific model analysis. By and large, relevant empirical and simulation evidences are very scarce and practical guidance is absent in education.

There are some findings in this study. First of all, the empirical investigation using PISA data shows that in general the statistical inferences of the regression coefficients are consistent across the conditions of unweighted vs. weighted and single vs. multi-level models. This is in line with previous findings that weights affect the estimates of population associations (e.g., coefficients) much less that of population mean values (Asparouhov, 2006; Korn & Graubard, 1995). Korn and Graubard (1995) found that for weighted and unweighted inference to be different, model must be grossly mis-specified or omitted variable has strong correlation with the independent variable as well as with sampling weights. For the variable of father's full-time working status, results revealed that the weighted single level model shares the same statistical inference with that of the weighted unscaled multi-level models, but is different from the weighted scaled multi-level model. This may indicate that the possible difference between weighted single level and weighted multi-level models lies in the scaling procedure in the latter.

In addition, within the weighted multi-level models, there is no difference between the "size" scaling and the "effective" scaling in terms of using U.S. data in PISA and ECLS-K. This finding is inconsistent with previous studies. For example, Pfeffermann et al (1998) found the "size" scaling (i.e., referred as method 2 in their paper) works better than "effective" scaling (Pfeffermann, Skinner, et al., 1998) in their simulation study while Stapleton (2002) found the "effective" scaling produced unbiased estimates while "size" scaling had negative bias in the

multilevel structural equation models (Stapleton, 2002). It seems that the effect of scaling method may depend upon both data and statistical models.

Second, the results of the simulation study indicate that regarding the relative bias, the quality of the parameter estimates generated from four statistical models (i.e., unweighted single level model, unweighted multi-level model, weighted single level model and weighted multilevel model) varies substantially across variables. Evidently, it is impossible to find one model that would produce the least biased estimate for all conditions. Instead, it is very likely that the preference of the model would vary across different covariates. Also statistical inference might appear divergent among these four models for different variables. Nonetheless, unweighted multilevel models seem to be superior to other models for having the smallest RMSE and standard errors.

One interesting question is whether the informativeness of the design is linked to the outcome only or to the predictors included in the model also? The informativeness of a design is a model concept which the model typically has outcome and predictors as a set of components to make inference about their relationship. Therefore, it makes more sense to investigate whether sampling design is related to the model inference given particular outcome and predictors rather than discuss the design effect on one particular component.

Take regression model as an example. Suppose the research interest is the coefficients from a specific model. If the observed values of the dependent variable satisfy the model universally regardless of which sample was actually selected, the design is non-informative as there is no additional information about the sample design beyond what is already specified in the model. For instance, suppose sample design only has one design variable such as stratum identifier in the case of a stratified sample, if this design variable is included in the model, this

informative design becomes ignorable as the inference of the predictors would not change once the design variable that contains all the design information is included in the model. However, suppose the sampling design also has the feature of clustering based on clusters of geographic areas. If this clustering feature was not included in the model, the regression error terms are correlated within these clusters. In this case, the design is still informative to the model inference for the predictors. To be more general, the answer varies from case to case because whether the informativeness of a design has an impact on the model inference depends on sample design, the available design information, the variables of interest and the assumed model (Binder & Roberts, 2001).

Third, the software investigation reveals that the method of MPML is superior to the method of PWIGLS because it has more flexibility and wider application in software programs. Among the software that use the MPML method, STATA and Mplus would be preferred to SAS for the advantages of scaling. This finding is in congruence with previous studies which showed that PML and MPML would be preferred to PWIGLS in terms of less computation intensity and more flexibility (Kovačević & Rai, 2003; Rabe-Hesketh & Skrondal, 2006).

One recommendation is that sampling weights analysis only works for the whole sample data while for the subgroup analysis, unweighted analysis would be more appropriate. In addition, since the design informativeness due to complex sampling design at multiple stages is hard to determine and its effect on statistical inference is unknown, researchers are encouraged to conduct both weighted and unweighted analyses. Convergent results will gain confidence for the finding while divergent results would also be informative, which scientifically showing that inference from the sample and population are likely to be different and researchers need to be aware of that. Moreover, researchers are encouraged to use STATA and Mplus for conducting

weighted model analyses especially for multilevel models. Furthermore, when using multi-level models, it should be noted that the scaling of first level sampling weights would be required to reduce bias. Although size and effective scaling produced similar results in large-scale educational data used in this study, the method of size scaling may have wider use as it appeared in all five software programs.

The empirical investigation in this study focused on PISA U.S. data which the basic sampling design is selecting schools at the first stage and then sample students within schools at the second stage. Therefore, findings may apply to other educational large-scale data sets that have design of similar sampling units at each level. For example, the High School Longitudinal study (HSLS) and Education Longitudinal Study (ELS) adopt a stratified two-stage sample design with primary sampling units defined as schools selected in the first stage and students then selected from the sampled schools within the second stage. In general, in terms of the common two-stage design, schools will be selected with probability that is proportional to measure of size and then students will be either randomly selected or with a certain selection rate.

Nevertheless, findings may not work in other data sets that have different sampling designs. For example, Early Childhood Longitudinal Study of Kindergarten (ECLS-K) adopt three stage sampling design. The TIMSS and PIRLS have two-stage sampling design in which schools are primary sampling units but within schools, one or more intact classrooms are sampled. With regard to the National Assessment of Educational Progress (NAEP), it has national sample and also state sample. In general, the state sample has similar two-stage sampling design as in PISA while in the national sample it is more similar to ECLS-K in which beyond school level, the geographic regions are the primary sampling units.

One caveat is in addition to the sampling layers and units, there are other factors such as sample size and over sampling particular ethnic groups that might need to be considered. There is further investigation work that needs to be done to make a more generalizable conclusion. Similarly, the findings and recommendations would be limited to the simulation conditions in this study which mainly based on the data structure and sampling condition in the large-scale data of ECLS-K. Therefore, it is possible that they may not be applicable to other simulation conditions.

This study has some limitations. First, as the simulation design tries to mimic the real large-scale data with specific sample selection rates, the design informativeness may not be varying sufficient enough to demonstrate alarming differences as appeared in previous simulation studies that used extreme conditions such as having cluster sample size less than 5 and ICC value less than 0.05 (i.e., 0.01). In addition, it aims to mirror the real condition of large-scale educational data, but still fails to fully capture the actual sampling selection procedures such as the neglected first sampling stage of geographic location, trimming of extreme large sampling variance and the variation of school size that occurred in real data. Second, the sampling weights provided in large-scale education data have already taken into account the adjustment for non-response and perhaps also for post-stratification. Therefore, post-stratification and non-response adjustment might have an effect on model analysis (Long, 1995), which was ignored in this simulation investigation.

The underlying causes of the divergence between unweighted and weighted estimators and the impact of informative design are still unclear and further investigation is needed. In addition, it would be interesting to focus on weighted multilevel modeling under informative probability sampling for estimating variance components (Asparouhov & Muthen, 2006; Jenkins,

2008). Moreover, extending the current research to non-linear model especially with binary outcome would be informative although it could be more challenging to deal with (Grilli & Pratesi, 2004). Moreover, longitudinal analysis with sampling weights need to be studied as it appears only in few studies (Stapleton, Harring, & Lee, 2016; Vieira & Skinner, 2008). Finally, the current literature and software application mainly works for two-level models, and future research may investigate the use of sampling weights in three-level models.

APPENDIX

SOFTWARE SYNTAX

/****************************** STATA 14 ******************************/

/******************** ***Unweighted analyses*** ************************/

- ---------- Single Level ----------

```
reg z1 female escs fafulltime private classsize schoolESCS, cluster(SCHOOLID)
```

- ---------- Multilevel ----------

```
mixed z1 female escs fafulltime private classsize schoolESCS||SCHOOLID:,var
```

/*************** ***Sampling weighted analyses*** ************************/

- ---------- Multi-Level ----------

```
svyset SCHOOLID [pweight = w_std_s]
svy: regress z1 female escs fafulltime private classsize schoolESCS
```

- ---------- Multilevel ----------

(1) Unscaled

```
mixed z1 female escs fafulltime private classsize schoolESCS [pw=w_std] ||SCHOOLID:,pweight(w_sch)
var
```

(2) Size scaled

```
mixed z1 female escs fafulltime private classsize schoolESCS [pw=w_std] ||SCHOOLID:,pweight(w_sch)
pwscale(size) var
```

(3) Effective scaled

```
mixed z1 female escs fafulltime private classsize schoolESCS [pw=w_std] ||SCHOOLID:,pweight(w_sch)
pwscale(effective) var
```

/****************************** Mplus 7.4 ****************************/

/******************** ***Unweighted analyses*** ************************/

- ---------- Single Level ----------

```
data: file = sw_pisa_mplus.csv;
variable: NAMES = SCH STI z1 fem escs fa pri cs schESCS w_s_al w1 w2 w1si w1ef;
        usevariables = SCH z1 fem escs fa pri cs schESCS;
        cluster = SCH;
analysis: type = complex;
        estimator = MLR;
model: z1 on fem escs fa pri cs schESCS;
output: stdyx;
```

- ---------- Multilevel ----------

```
data: file = sw_pisa_mplus.csv;
variable: NAMES = SCH STI z1 fem escs fa pri cs schESCS w_s_al w1 w2 w1si w1ef;
        usevariables = SCH z1 fem escs fa pri cs schESCS;
        within = fem escs fa;
        between = pri cs schESCS;
        cluster = SCH;
analysis: type = twolevel;
        algorithm=integration;
        estimator = ML;
model:  %WITHIN%
         z1 on fem escs fa;
        %BETWEEN%
        z1 on pri cs schESCS;
output: stdyx;
```

/*************** ***Sampling weighted analyses*** ************************/

- ---------- Single Level ----------

```
data: file = sw_pisa_mplus.csv;
variable: NAMES = SCH STI z1 fem escs fa pri cs schESCS w_s_al w1 w2 w1si w1ef;
        usevariables = SCH z1 fem escs fa pri cs schESCS;
        cluster = SCH;
        weight=w_s_al;
analysis: type = complex;
        estimator = MLR;
model: z1 on fem escs fa pri cs schESCS;
output: stdyx;
```

- ---------- Multi-level ----------

(1) Unscaled

```
data: file = sw_pisa_mplus.csv;
variable: NAMES = SCH STI z1 fem escs fa pri cs schESCS w_s_al w1 w2 w1si w1ef;
        usevariables = SCH z1 fem escs fa pri cs schESCS;
        within = fem escs fa;
        between = pri cs schESCS;
        cluster = SCH;
        weight=w1;
        bweight=w2;
        wtscale=unscaled;
```

```
        bwtscale=unscaled;
analysis: type = twolevel;
        algorithm=integration;
        estimator = MLR;
model:  %WITHIN%
         z1 on fem escs fa;
        %BETWEEN%
        z1 on pri cs schESCS;
output: stdyx;
```

(2) Size scaled

```
data: file = sw_pisa_mplus.csv;
variable: NAMES = SCH STI z1 fem escs fa pri cs schESCS
        w_s_al w1 w2 w1si w1ef;
        usevariables = SCH z1 fem escs fa pri cs schESCS;
        within = fem escs fa;
        between = pri cs schESCS;
        cluster = SCH;
        weight=w1;
        bweight=w2;
        wtscale=cluster;
        bwtscale=unscaled;
analysis: type = twolevel;
        algorithm=integration;
        estimator = MLR;
model:  %WITHIN%
         z1 on fem escs fa;
        %BETWEEN%
        z1 on pri cs schESCS;
output: stdyx;
```

(3) Effective scaled

```
data: file = sw_pisa_mplus.csv;
variable: NAMES = SCH STI z1 fem escs fa pri cs schESCS
        w_s_al w1 w2 w1si w1ef;
        usevariables = SCH z1 fem escs fa pri cs schESCS;
        within = fem escs fa;
        between = pri cs schESCS;
        cluster = SCH;
        weight=w1;
        bweight=w2;
        wtscale=ecluster;
        bwtscale=unscaled;
analysis: type = twolevel;
        algorithm=integration;
        estimator = MLR;
model:  %WITHIN%
         z1 on fem escs fa;
        %BETWEEN%
        z1 on pri cs schESCS;
output: stdyx
```

```
/***************************** SAS 9.4 ********************************/
```

- ---------- Single Level ----------

```
title1 'Single level unweighted analysis';
   proc surveyreg data=sw_pisa;
     cluster schid;
     model z1= female ses fawork private clsize schses / solution;
   run;

title1 'multi-level unweighted analysis';
   proc mixed data=sw_pisa method=ML;
     class schid;
     model z1= female ses fawork private clsize schses/ solution;
           random intercept/subject=schid;
   run;

title1 'Single level weighted analysis';
   proc surveyreg data=sw_pisa;
     cluster schid;
     model z1= female ses fawork private clsize schses/ solution;
           weight stdw;
   run;
```

- ---------- Multi-level ----------

```
title1 'multi-level weighted unscaled analysis';
   proc glimmix data=sw_pisa method=quadrature empirical=classical;
     class schid;
     model z1= female ses fawork private clsize schses/ obsweight= stdl1 solution;
           random intercept/subject=schid weight=schl2;
   run;


title1 'multi-level weighted size analysis';
   proc glimmix data=sw_pisa method=quadrature empirical=classical;
     class schid;
     model z1= female ses fawork private clsize schses/ obsweight= stdl1size solution;
           random intercept/subject=schid weight=schl2;
   run;

title1 'multi-level weighted effective analysis';
   proc glimmix data=sw_pisa method=quadrature empirical=classical;
     class schid;
     model z1= female ses fawork private clsize schses/ obsweight= stdl1eff solution;
           random intercept/subject=schid weight=schl2;
   run;
```

/*************************** LISREL 9.30 ******************************/
/******************** ***Unweighted analyses*** ************************/

- ---------- Single Level ----------

Not avaiable

- ---------- Multilevel ----------

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
 TITLE=;
 SY='C:\Ting\D\msu_study\Dissertation\LISREL9.30\SW_SPSS.LSF';
 ID2=SCHOOLID;
 RESPONSE=Z1;
 FIXED=intcept FEMALE ESCS FAFULLTI PRIVATE CLASSSIZ SCHOOLES;
 RANDOM1=intcept;
 RANDOM2=intcept;
```

/*************** ***Sampling weighted analyses*** ************************/

- ---------- Single Level ----------

```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
RefCatCode=-1 IterDetails=No
        Method=Fisher; ! (Netwon-raphson)
Title=;
SY='C:\Ting\D\msu_study\Dissertation\LISREL9.30\SW_SPSS.LSF';
Distribution=NOR;
Link=IDEN;
Intercept=Yes;
Scale=None; (deviance/pearson/ML)
DepVar=Z1;
CoVars=FEMALE ESCS FAFULLTI PRIVATE CLASSSIZ SCHOOLES;
Cluster=SCHOOLID;
Weight=W_STD_S;
```

not working

- ---------- Multilevel ----------

(1) Size scaled (automatic)

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
 TITLE=;
 SY='C:\Ting\D\msu_study\Dissertation\LISREL9.30\SW_SPSS.LSF';
 ID2=SCHOOLID;
 WEIGHT2=W_SCH;
 WEIGHT1=W_STD;
 RESPONSE=Z1;
 FIXED=intcept FEMALE ESCS FAFULLTI PRIVATE CLASSSIZ SCHOOLES;
 RANDOM1=intcept;
 RANDOM2=intcept;
```

FIGURES

Figure 1.1: Relative Bias

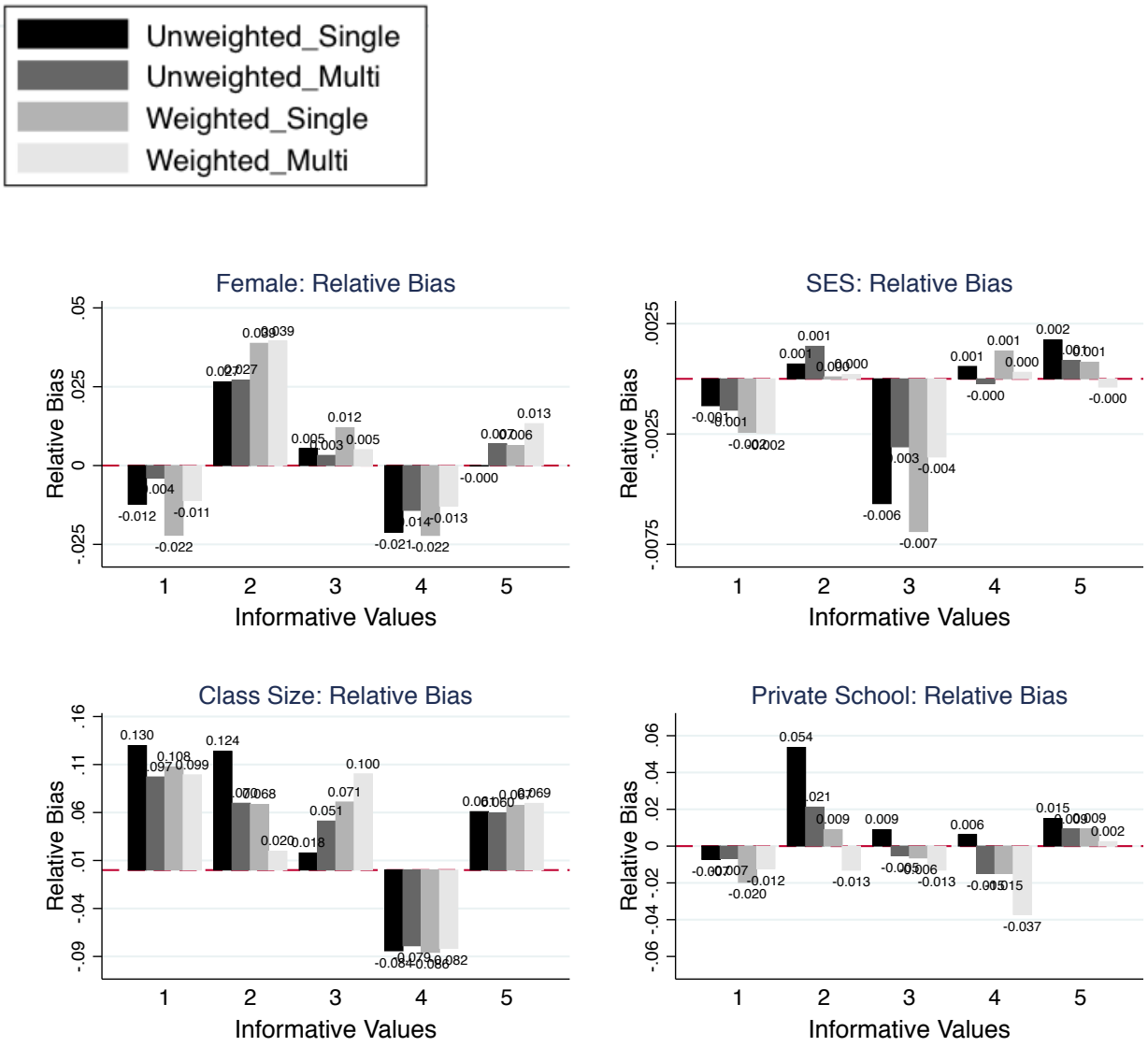


Figure 1.2: Root of Mean Square Error

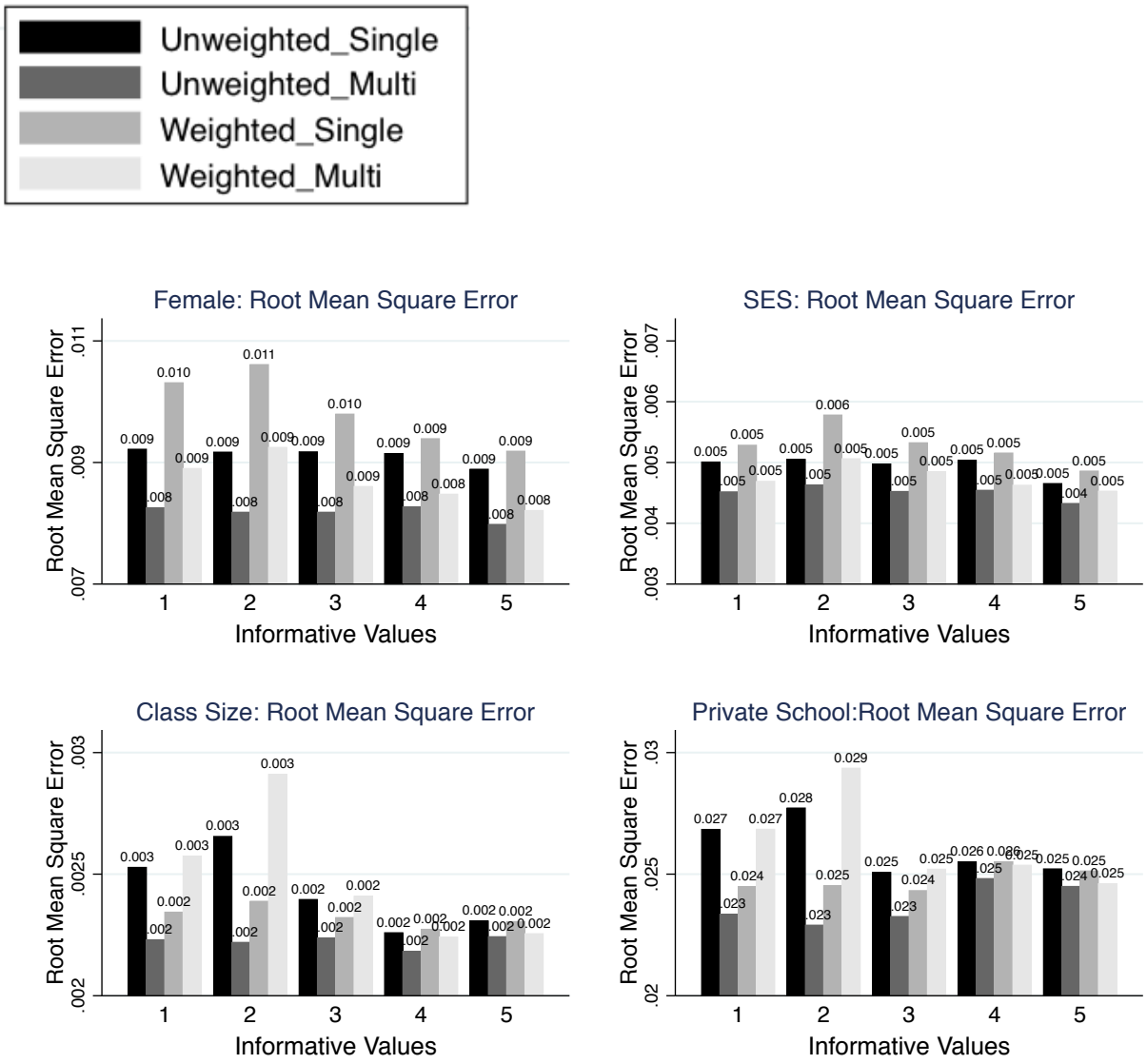
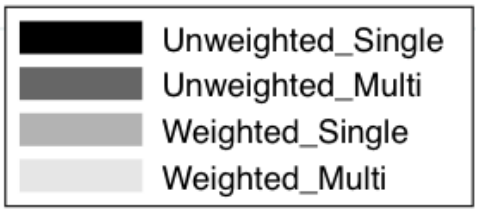
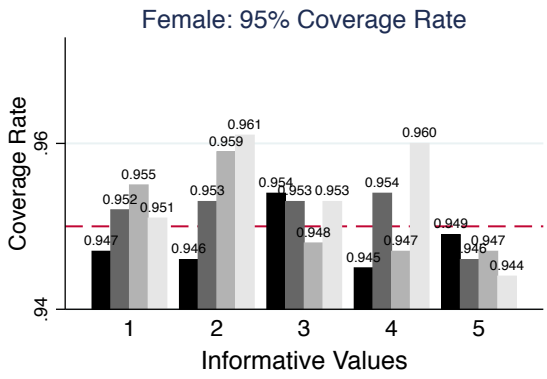
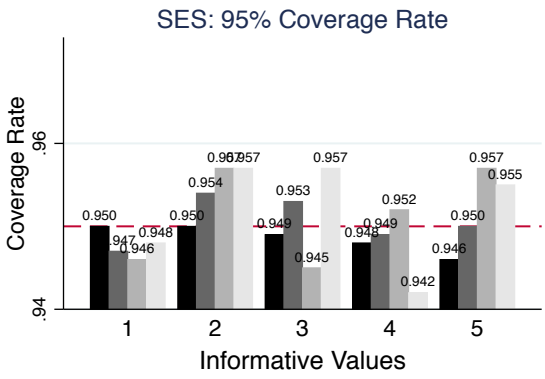
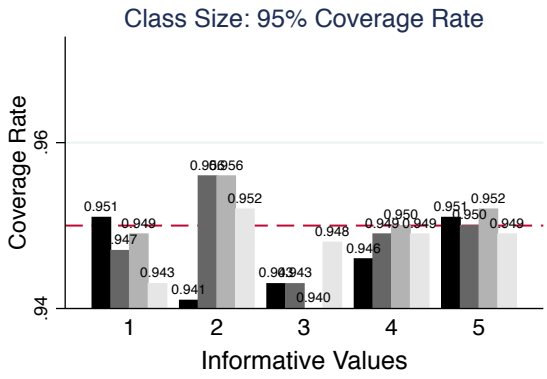


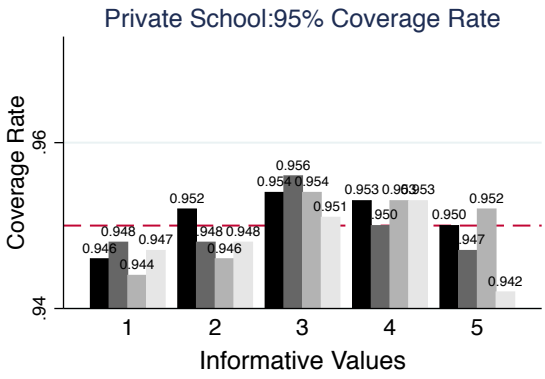
Figure 1.3: 95 % Coverage Rate











TABLES

Table 1.1: Variable descriptive statistics

PISA 2012	Name	Mean	SD	MIN	Max
Math score	PV1MATH	0.00	1.00	-3.44	3.82
Female	ST04Q01	0.49	0.50	0	1
Economic social and cultural status (ESCS)	ESCS	0.19	0.97	-3.8	3.12
Father work full-time	ST19Q01	0.73	0.45	0	1
Class size	CSIZE	0.09	0.28	0	1
Private	SC01Q01	26.23	5.28	13	43
School ESCS	ESCS (Average by school)	0.18	0.53	-1.81	1.25
Student final weights	W_FSTUWT	710.88	293.01	134.50	2597.88
School final weights	W_FSCHWT	145.25	204.10	21.97	1942.55
Number of students	4978				
Number of schools	162				
ECLS-K 2011	Name	Mean	SD	MIN	Max
Reading gain score	X2RTHETK2-X1RTHETK2	1.03	0.53	-3.36	3.38
Class size	A2DENROL	20.05	4.27	9	29
Female	X_CHSEX_R	0.49	0.50	0	1
Race (Other minority as base)	X_RACETH_R				
White	-	0.49	0.50	0	1
Black	-	0.14	0.35	0	1
Asian	-	0.07	0.25	0	1
Hispanic	-	0.24	0.43	0	1
SES	X12SESL	-0.06	0.77	-2.33	2.44
School location (Rural as base)	X2LOCALE				
City	-	0.33	0.47	0	1
Suburban	-	0.34	0.47	0	1
Town	-	0.09	0.29	0	1
Private	X2PUBPRI	0.16	0.37	0	1
School enrollment	X2KENRLK	81.96	51.43	14	250
Free lunch	2RFLCH2_I	44.98	31.90	0	100
Reduced lunch	X2RLCH2_I	7.77	9.07	0	100
Student final weights (A)	W12AC0	283.79	149.10	11.46	940.02
School final weights (B)	W2SCH0	79.04	46.87	4.45	372.03
Student level specific weights (C)	(A)/(B)	4.83	3.88	0.57	43.59
Number of students	10349				
Number of schools	678				

Table 1.2: Unweighted estimates across five software programs in PISA U.S. 2012

	STATA 14				Mplus 7				SAS 9.4				LISREL 9.30				HLM 7	
Model	Single		Multi		Single		Multi		Single		Multi		Single		Multi		Multi	
N=4978	Est		Est		Est		Est		Est		Est		Est		Est		Est	
Female	-0.089	*	-0.117	*	-0.089	*	-0.118	*	-0.089	*	-0.117	*	0.0573	*	-0.117	*	-0.117	*
SE	(0.028)		(0.024)		(0.028)		(0.024)		(0.028)		(0.024)		(0.012)		(0.024)		(0.024)	
ESCS	0.248	*	0.250	*	0.248	*	0.250	*	0.248	*	0.250	*	0.229	*	0.250	*	0.250	*
SE	(0.019)		(0.015)		(0.019)		(0.015)		(0.019)		(0.015)		(0.015)		(0.015)		(0.015)	
Father work full-time	0.209	*	0.183	*	0.209	*	0.181	*	0.209	*	0.183	*	0.128	*	0.183	*	0.182	*
SE	(0.031)		(0.028)		(0.031)		(0.028)		(0.031)		(0.028)		(0.013)		(0.028)		(0.028)	
Private	-0.434	*	-0.396	*	-0.434	*	-0.396	*	-0.434	*	-0.396	*	-0.414	*	-0.396	*	-0.395	*
SE	(0.114)		(0.101)		(0.114)		(0.101)		(0.114)		(0.101)		(0.017)		(0.100)		(0.102)	
Class size	-0.001		-0.003		-0.001		-0.003		-0.001		-0.003		0.146		-0.003		-0.003	
SE	(0.006)		(0.005)		(0.006)		(0.005)		(0.006)		(0.005)		(0.013)		(0.005)		(0.005)	
School ESCS	0.491	*	0.495	*	0.491	*	0.495	*	0.491	*	0.495	*	0.859	*	0.495	*	0.495	*
SE	(0.054)		(0.054)		(0.054)		(0.054)		(0.054)		(0.054)		(0.032)		(0.054)		(0.055)	

Note: * p<0.05;

Table 1.3: Weighted estimates across five software programs in PISA U.S. 2012

Model	STATA 14				Mplus 7				SAS 9.4				LISREL9.30	HLM 7
	Single	Multi-level			Single	Multi-level			Single	Multi-level (user to scale)			Multi-level	Multi-level
		Unscale	Size	Effective		Unscale	Size	Effective		Unscale	Size	Effective	Size	Size
N=4978	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est
Female	-0.071 *	-0.097 *	-0.094 *	-0.095 *	-0.071 *	-0.097 *	-0.094 *	-0.095 *	-0.071 *	-0.097 *	-0.094 *	-0.095 *	-0.096 *	-0.096 *
SE	(0.027)	(0.024)	(0.030)	(0.030)	(0.027)	(0.024)	(0.030)	(0.030)	(0.027)	(0.024)	(0.030)	(0.030)	(0.030)	(0.030)
ESCS	0.248 *	0.250 *	0.253 *	0.253 *	0.248 *	0.250 *	0.253 *	0.253 *	0.248 *	0.250 *	0.253 *	0.253 *	0.253 *	0.253 *
SE	(0.021)	(0.021)	(0.026)	(0.026)	(0.021)	(0.021)	(0.026)	(0.026)	(0.021)	(0.021)	(0.026)	(0.026)	(0.026)	(0.026)
Father work full-time	0.183 *	0.155 *	0.128	0.128	0.183 *	0.155 *	0.128	0.128	0.183 *	0.155 *	0.128	0.128	0.125	0.125
SE	(0.040)	(0.035)	(0.066)	(0.066)	(0.040)	(0.035)	(0.066)	(0.066)	(0.040)	(0.035)	(0.066)	(0.066)	(0.066)	(0.066)
Private	-0.420 *	-0.392 *	-0.400 *	-0.400 *	-0.420 *	-0.392 *	-0.400 *	-0.400 *	-0.420 *	-0.392 *	-0.400 *	-0.400 *	-0.397 *	-0.397 *
SE	(0.095)	(0.105)	(0.107)	(0.107)	(0.095)	(0.105)	(0.106)	(0.107)	(0.095)	(0.105)	(0.106)	(0.106)	(0.106)	(0.106)
Class size	-0.002	-0.006	-0.006	-0.006	-0.002	-0.006	-0.006	-0.006	-0.002	-0.006	-0.006	-0.006	-0.006	-0.006
SE	(0.007)	(0.008)	(0.008)	(0.008)	(0.007)	(0.008)	(0.008)	(0.008)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
School ESCS	0.521 *	0.705 *	0.713 *	0.713 *	0.521 *	0.705 *	0.713 *	0.713 *	0.521 *	0.705 *	0.713 *	0.713 *	0.718 *	0.718 *
SE	(0.060)	(0.100)	(0.103)	(0.103)	(0.060)	(0.101)	(0.105)	(0.103)	(0.060)	(0.101)	(0.105)	(0.105)	(0.102)	(0.102)

Note: * p<0.05

Table 1.4: Subgroup analyses by gender in PISA U.S. 2012

Unweighted	STATA 14		**Weighted**	STATA 14				Mplus 7				SAS 9.4				
Female	Single	Multi	Single	Multi				Single	Multi			Single	Multi (user to scale)			Multi
				Unscale	Size	Effective			Unscale	Size	Effective		Unscale	Size	Effective	Size
N=2,453	Est	Est	Est	Est	Est	Est		Est	Est	Est	Est	Est	Est	Est	Est	Est
ESCS	0.267 *	0.269 *	0.263 *	0.267 *	0.252 *	0.252 *		0.263 *	0.267 *	0.252 *	0.252 *	0.263 *	0.267 *	0.253 *	0.253 *	0.252 *
SE	(0.024)	(0.021)	(0.025)	(0.026)	(0.025)	(0.025)		(0.025)	(0.026)	(0.025)	(0.025)	(0.025)	(0.026)	(0.025)	(0.025)	(0.025)
Father full-time	0.175 *	0.161 *	0.141 *	0.121 *	0.099	0.099		0.141 *	0.121 *	0.099	0.099	0.141 *	0.121 *	0.098	0.098	0.097
SE	(0.041)	(0.039)	(0.053)	(0.051)	(0.125)	(0.125)		(0.053)	(0.050)	(0.124)	(0.124)	(0.053)	(0.051)	(0.125)	(0.125)	(0.124)
Private	-0.327 *	-0.289 *	-0.339 *	-0.290 *	-0.307 *	-0.307 *		-0.339 *	-0.290 *	-0.306 *	-0.306 *	-0.339 *	-0.290 *	-0.306 *	-0.306 *	-0.305 *
SE	(0.071)	(0.107)	(0.069)	(0.101)	(0.101)	(0.101)		(0.069)	(0.100)	(0.100)	(0.100)	(0.069)	(0.100)	(0.101)	(0.101)	(0.101)
Class size	-0.002	-0.003	-0.003	-0.003	-0.003	-0.003		-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003
SE	(0.006)	(0.005)	(0.007)	(0.007)	(0.007)	(0.007)		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
School ESCS	0.476 *	0.462 *	0.491 *	0.608 *	0.639 *	0.638 *		0.491 *	0.608 *	0.639 *	0.638 *	0.491 *	0.608 *	0.636 *	0.635 *	0.641 *
SE	(0.060)	(0.057)	(0.059)	(0.081)	(0.090)	(0.090)		(0.059)	(0.083)	(0.092)	(0.092)	(0.059)	(0.083)	(0.092)	(0.092)	(0.090)
	STATA 14		STATA 14					Mplus 7				SAS 9.4				
Male	Single	Multi-level	Single	Multi				Single	Multi			Single	Multi (user to scale)			Multi
				Unscale	Size	Effective			Unscale	Size	Effective		Unscale	Size	Effective	Size
N=2,525	Est	Est	Est	Est	Est	Est		Est	Est	Est	Est	Est	Est	Est	Est	Est
ESCS	0.231 *	0.232 *	0.234 *	0.237 *	0.260 *	0.260 *		0.234 *	0.237 *	0.260 *	0.260 *	0.234 *	0.237 *	0.256 *	0.256 *	0.260 *
SE	(0.026)	(0.022)	(0.029)	(0.028)	(0.041)	(0.041)		(0.029)	(0.028)	(0.041)	(0.041)	(0.029)	(0.028)	(0.041)	(0.041)	(0.041)
Father full-time	0.239 *	0.211 *	0.223 *	0.183 *	0.173 *	0.173 *		0.223 *	0.183 *	0.173 *	0.173 *	0.223 *	0.183 *	0.172 *	0.172 *	0.168 *
SE	(0.043)	(0.041)	(0.050)	(0.045)	(0.059)	(0.059)		(0.050)	(0.045)	(0.059)	(0.059)	(0.050)	(0.045)	(0.059)	(0.059)	(0.058)
Private	-0.533 *	-0.434 *	-0.498 *	-0.430 *	-0.444 *	-0.444 *		-0.498 *	-0.430 *	-0.444 *	-0.444 *	-0.498 *	-0.430 *	-0.445 *	-0.445 *	-0.439 *
SE	(0.181)	(0.118)	(0.153)	(0.130)	(0.134)	(0.134)		(0.153)	(0.130)	(0.134)	(0.134)	(0.153)	(0.130)	(0.133)	(0.133)	(0.133)
Class size	0.001	-0.002	-0.001	-0.008	-0.007	-0.007		-0.001	-0.008	-0.007	-0.007	-0.001	-0.008	-0.007	-0.007	-0.008
SE	(0.006)	(0.006)	(0.008)	(0.009)	(0.009)	(0.009)		(0.008)	(0.009)	(0.009)	(0.009)	(0.008)	(0.009)	(0.010)	(0.010)	(0.009)
School ESCS	0.498 *	0.507 *	0.542 *	0.720 *	0.707 *	0.707 *		0.542 *	0.720 *	0.707 *	0.707 *	0.542 *	0.720 *	0.713 *	0.713 *	0.710 *
SE	(0.062)	(0.063)	(0.075)	(0.115)	(0.126)	(0.126)		(0.075)	(0.115)	(0.126)	(0.126)	(0.075)	(0.116)	(0.128)	(0.128)	(0.125)

Note: * p<0.05; HLM does not support weighted subgroup analysis

Table 1.5: Subgroup analyses by school sector in PISA U.S. 2012

	Unweighted	STATA 14		STATA 14				Mplus 7				SAS 9.4				LISREL 9.30
Public	Level	Single	Multi	Single	Multi			Single	Multi			Single	Multi (user to scale)			Multi
					Unscale	Size	Effective		Unscale	Size	Effective		Unscale	Size	Effective	Size
N=4,551		Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est
Female		-0.108 *	-0.122 *	-0.081 *	-0.099 *	-0.104 *	-0.104 *	-0.081 *	-0.099 *	-0.104 *	-0.105 *	-0.081 *	-0.099 *	-0.104 *	-0.104 *	-0.105 *
SE		(0.025)	(0.025)	(0.027)	(0.025)	(0.031)	(0.031)	(0.027)	(0.025)	(0.031)	(0.031)	(0.027)	(0.025)	(0.031)	(0.031)	(0.031)
ESCS		0.252 *	0.254 *	0.249 *	0.251 *	0.249 *	0.248 *	0.249 *	0.251 *	0.249 *	0.248 *	0.249 *	0.251 *	0.249 *	0.249 *	0.249 *
SE		(0.020)	(0.015)	(0.022)	(0.022)	(0.027)	(0.027)	(0.022)	(0.022)	(0.027)	(0.027)	(0.022)	(0.022)	(0.027)	(0.027)	(0.027)
Father work full-time		0.208 *	0.181 *	0.181 *	0.153 *	0.109	0.109	0.181 *	0.153 *	0.109	0.109	0.181 *	0.153 *	0.109	0.109	0.107
SE		(0.032)	(0.029)	(0.042)	(0.037)	(0.072)	(0.072)	(0.042)	(0.037)	(0.071)	(0.071)	(0.042)	(0.037)	(0.072)	(0.072)	(0.071)
Class size		-0.002	-0.004	-0.002	-0.003	-0.003	-0.003	-0.002	-0.003	-0.003	-0.003	-0.002	-0.003	-0.003	-0.003	-0.003
SE		(0.006)	(0.005)	(0.007)	(0.009)	(0.009)	(0.009)	(0.007)	(0.009)	(0.009)	(0.009)	(0.007)	(0.009)	(0.009)	(0.009)	(0.009)
School ESCS		0.474 *	0.479 *	0.512 *	0.695 *	0.711 *	0.711 *	0.512 *	0.695 *	0.711 *	0.711 *	0.512 *	0.695 *	0.711 *	0.711 *	0.715 *
SE		(0.055)	(0.055)	(0.062)	(0.111)	(0.116)	(0.116)	(0.062)	(0.112)	(0.118)	(0.118)	(0.062)	(0.113)	(0.118)	(0.118)	(0.115)
	Unweighted	STATA 14		STATA 14				Mplus 7				SAS 9.4				
Private	Level	Single	Multi	Single	Multi			Single	Multi			Single	Multi (user to scale)			Multi
				-level	Unscale	Size	Effective	-level	Unscale	Size	Effective	-level	Unscale	Size	Effective	Size
N=427		Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est	Est
Female		0.063	-0.065	0.027	-0.047	-0.032	-0.032	0.027	-0.047	-0.032	-0.032	0.027	-0.047	-0.032	-0.032	-0.035
SE		(0.179)	(0.085)	(0.155)	(0.084)	(0.094)	(0.094)	(0.151)	(0.081)	(0.090)	(0.090)	(0.156)	(0.084)	(0.094)	(0.093)	(0.091)
ESCS		0.186 *	0.178 *	0.229 *	0.225 *	0.280 *	0.280 *	0.229 *	0.225 *	0.280 *	0.280 *	0.229 *	0.225 *	0.280 *	0.280 *	0.280 *
SE		(0.074)	(0.057)	(0.077)	(0.079)	(0.075)	(0.074)	(0.074)	(0.076)	(0.072)	(0.072)	(0.077)	(0.079)	(0.075)	(0.075)	(0.072)
Father work full-time		0.185 *	0.206 *	0.189 *	0.200 *	0.244 *	0.244 *	0.189 *	0.200 *	0.244 *	0.244 *	0.189 *	0.200 *	0.244 *	0.244 *	0.244 *
SE		(0.075)	(0.100)	(0.087)	(0.087)	(0.117)	(0.117)	(0.085)	(0.084)	(0.113)	(0.113)	(0.088)	(0.087)	(0.117)	(0.117)	(0.113)
Class size		0.016	0.000	0.005	-0.024 *	-0.024 *	-0.024 *	0.005	-0.024 *	-0.024 *	-0.024 *	0.005	-0.024 *	-0.024 *	-0.024 *	-0.024 *
SE		(0.018)	(0.016)	(0.015)	(0.012)	(0.012)	(0.012)	(0.014)	(0.012)	(0.011)	(0.011)	(0.015)	(0.012)	(0.012)	(0.012)	(0.011)
School ESCS		0.891 *	0.871 *	0.764 *	0.904 *	0.844 *	0.844 *	0.764 *	0.904 *	0.844 *	0.844 *	0.764 *	0.904 *	0.844 *	0.844 *	0.843 *
SE		(0.197)	(0.301)	(0.174)	(0.169)	(0.173)	(0.173)	(0.169)	(0.167)	(0.167)	(0.167)	(0.175)	(0.172)	(0.173)	(0.173)	(0.166)

Note: * p<0.05

Table 1.6: Empirical estimates in ECLS-K 2011 (STATA 14)

N=10349	Unweighted				Weighted											
	Single-level (a)		Multi-level (b)		Single-level (c)				Multi-level (d)(Size scaling)					Correlation		
	SRS	(1)				(2)		RDE(A)		(3)			RDE (B)	Student	Student level	School level
Covariate X	Est	SE (Robust)	Est	SE	Est	SE	I (A)	(2)/(1)	Est	SE	I (C)	I (B)	(3)/(1)	weights (A)	Weights (C)	(B)
Female	0.019	0.010	-0.001	0.002	0.016	0.012	0.43	1.13	0.017	0.010	0.50	0.09	0.971	-0.01	-0.01	0.00
White	0.020	0.022	0.016	0.009	0.009	0.030	2.71	1.41	0.020	0.026	6.95	10.48	1.191	0.05	-0.08	0.17
Black	-0.082 *	0.025	0.008	0.021	-0.061	0.035	2.15	1.43	-0.021	0.030	1.57	0.65	1.237	0.04	0.02	0.01
Asian	-0.022	0.029	-0.030	0.025	-0.023	0.038	11.56	1.31	0.021	0.037	8.80	5.05	1.254	-0.22	-0.11	-0.08
Hispanic	0.022	0.024	0.013	0.028	0.051	0.034	2.76	1.46	0.071 *	0.028	14.14	10.40	1.204	0.05	0.17	-0.17
SES	-0.065 *	0.008	0.037	0.023	-0.060 *	0.011	3.11	1.43	-0.047 *	0.009	9.01	6.92	1.175	-0.06	-0.11	0.11
Class size	-0.001	0.001	-0.049 *	0.008	-0.001	0.003	6.30	2.23	-0.001	0.003	22.31	20.49	1.923	0.12	0.27	-0.34
City	-0.054 *	0.015	-0.075 *	0.030	-0.062	0.034	2.17	2.26	-0.095 *	0.032	9.19	9.59	2.168	0.04	0.11	-0.16
Suburban	-0.042 *	0.014	-0.064 *	0.029	-0.048	0.032	5.67	2.25	-0.102 *	0.032	1.52	1.53	2.234	-0.11	-0.02	0.03
Town	0.000	0.020	-0.007	0.042	-0.007	0.047	7.54	2.31	-0.024	0.043	1.09	0.34	2.103	0.14	0.01	-0.01
Private	-0.026	0.017	-0.035	0.037	-0.061	0.040	9.12	2.31	-0.069	0.038	25.08	25.22	2.221	-0.17	-0.31	0.42
School enrollment	0.000	0.000	0.000	0.000	0.000	0.000	20.48	3.00	0.000	0.000	67.50	37.18	3.000	0.38	0.83	-0.62
Free lunch	0.001 *	0.000	0.001 *	0.000	0.001	0.001	5.68	3.00	0.000	0.001	14.93	14.07	2.500	0.11	0.18	-0.23
Reduce lunch	0.000	0.001	0.000	0.001	0.000	0.001	0.36	2.00	0.000	0.001	0.99	0.49	1.833	0.01	-0.01	0.01

Note: * p<0.05; RDE=Root Design Effect; I=informative index

Table 1.7: Simulation standard deviations of point estimators

Informativeness->	I=1 (High)		I=2		I=3		I=4		I=5 (Low)	
	SD	SE	SD	SE	SD	SE	SD	SE	SD	SE
Female										
NWS	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009
NWM	0.008	0.007	0.008	0.007	0.008	0.007	0.008	0.007	0.008	0.007
WS	0.010	0.010	0.011	0.010	0.010	0.009	0.009	0.009	0.009	0.009
WM	0.009	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.008
SES										
NWS	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
NWM	0.005	0.004	0.005	0.004	0.005	0.004	0.005	0.004	0.004	0.004
WS	0.005	0.005	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005
WM	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
Class size										
NWS	0.003	0.002	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002
NWM	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
WS	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
WM	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002
Private school										
NWS	0.027	0.027	0.028	0.029	0.025	0.026	0.026	0.025	0.025	0.025
NWM	0.023	0.024	0.023	0.024	0.023	0.024	0.025	0.024	0.025	0.024
WS	0.025	0.025	0.025	0.026	0.024	0.025	0.026	0.025	0.025	0.025
WM	0.027	0.028	0.029	0.030	0.025	0.026	0.025	0.025	0.025	0.024

Note: NWS=Unweighted single level; NWM=Unweighted Multi-level; WS=Weighted single level; NWM=Weighted Multi-level.
SD= the standard deviation of the 1000 point estimates; SE=the mean of the estimated standard errors.

REFERENCES

REFERENCES

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods, 35*(3), 439-460.

Asparouhov, T., & Muthen, B. (2006). *Multilevel modeling of complex survey data.* Paper presented at the Proceedings of the joint statistical meeting in seattle.

Bertolet, M. (2008). *To weight or not to weight? Incorporating sampling designs into model-based analyses.* (Ph.D.), Carnegie Mellon University, Ann Arbor.

Binder, D., & Roberts, G. (2006). *Approaches for analyzing survey data: a discussion.* Paper presented at the Survey Research Methods Section, American Statistical Association

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*(3), 279-292.

Binder, D. A., Kovacevic, M. S., & Roberts, G. (2005). *How important is the informativeness of the sample design.* Paper presented at the Proceedings of the Survey Methods Section.

Binder, D. A., & Roberts, G. (2009). Design-and model-based inference for model parameters. *Handbook of Statistics, 29*, 33-54.

Binder, D. A., & Roberts, G. R. (2001). Can Informative Designs be Ignorable? *Newsletter of the Survey Research Methods Section*(12), 1-3.

Binder, D. A., & Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of survey data* (pp. 29-48): John Wiley & Sons, Ltd.

Cai, T. (2013). Investigation of Ways to Handle Sampling Weights for Multilevel Model Analyses. *Sociological Methodology, 43*(1), 178-219.

Carle, A. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology, 9*(1), 1-13.

Chantala, K., Blanchette, D., & Suchindran, C. M. (2006). Software to compute sampling weights for multilevel analysis. *Carolina Population Center, UNC at Chapel Hill, Last Update*.

Chantala, K., & Suchindran, C. (2006). Adjusting for unequal selection probability in multilevel models: A comparison of software packages. *Proceedings of the American Statistical Association, Seattle, WA: American Statistical Association*, 2815-2824.

Cohen, S. B., Burt, V. L., & Jones, G. K. (1986). Efficiencies in variance estimation for complex

survey data. *The American Statistician, 40*(2), 157-164.

Eideh, A., & Nathan, G. (2009). Two-stage informative cluster sampling with application in small area estimation. *Journal of statistical planning and inference, 139*, 3088-3101.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*(1), 43-56.

Graubard, B. I., & Korn, E. L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical methods in medical research, 5*(3), 263-281.

Grilli, L., & Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology, 30*(1), 93-103.

Jenkins, F. (2008). *Multilevel analysis with informative weights.* Paper presented at the Proc. the Joint Statistical Meeting, ASA section on Survey Research Methods.

Jia, Y., Stokes, L., Harris, I., & Wang, Y. (2011). Performance of random effects model estimators under complex sampling designs. *Journal of educational and behavioral statistics, 36*(1), 6-32.

Kish, L. (1967). *Survey sampling* (2nd ed. ed.). New York: Wiley.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics, 8*(2), 183-200.

Korn, E. L., & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician, 49*(3), 291-295.

Korn, E. L., & Graubard, B. I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(1), 175-190.

Kovačević, M. S., & Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communications in Statistics-Theory and Methods, 32*(1), 103-121.

Koziol, N. A., Bovaird, J. A., & Suarez, S. (2017). A Comparison of Population-Averaged and Cluster-Specific Approaches in the Context of Unequal Probabilities of
Selection. *Multivariate Behavioral Research, 52*(3), 325-349.

Krieger, A. M., & Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology, 18*(2), 225-239.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American statistical Association, 99*(466), 546-556.

Lohr, S. L., & Liu, J. (1994). A comparison of weighted and unweighted analyses in the National Crime Victimization Survey. *Journal of Quantitative Criminology, 10*(4), 343-360.

Long, N. T. (1995). *Model-based methods for analysis of data from 1990 NAEP trial state assessment*. Washington, DC.

Mulligan, G. M., Hastedt, S., & McCarroll, J. C. (2012). *First-Time Kindergarteners in 2010-11: First Findings From the Kindergarten Rounds of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECS-K:2011) (NCES 2012- 049)*. Washington, DC: National Center for Education Statistics.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's Guide: Statistical Analysis with Latent Variables: User'ss Guide*: Muthén & Muthén.

OECD. (2012). PISA 2012 technical report. *Paris: Organisation for Economic Co-operation and Development*.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique, 61*(2), 317-337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical methods in medical research, 5*(3), 239-261.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it. *Survey Methodology, 37*(2), 115-136.

Pfeffermann, D., Krieger, A. M., & Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica, 8*(4), 1087-1114.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B, Statistical methodology, 60*(1), 23-40.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal, 2*(1), 1-21.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004a). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167-190.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004b). GLLAMM manual.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*(2), 301-323.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal*

of the Royal Statistical Society: Series A (Statistics in Society), 169(4), 805-827.

Randenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & Toit, M. D. (2011). *Hierarchical Linear and Nonlinear Modeling (HLM 7).* Lincolnwood, IL: Scientific Software International.

Rao, J. N. K., & Bellhouse, D. R. (1990). History and development of the theoretical foundations of survey based estimation and analysis. *Survey Methodology, 16*(1), 3-29.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1*(4), 381-397.

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. V. Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment: background, technical issues, and methods of data analysis* (pp. 117-153). London: Chapman Hall/CRC Press.

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical methods in medical research, 5*(3), 283-310.

SAS Institute Inc. (2013). *SAS/STAT® 13.1 User's Guide: The GLIMMIX procedure.*

Saw, G. K., & Schneider, B. (2015). Challenges and Opportunities for Estimating Effects with Large-Scale Education Data Sets. *Contemporary Educational Research Quarterly, 23*(4), 93-119.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating Causal Effects Using Experimental and Observational Designs (report from the Governing Board of the American Educational Research Association Grants Program).* Washington, DC: American Educational Research Association.

Scientific Software International. (2005-2012). "Multilevel Model" LISREL Documentation.

Skinner, C. (1994). *Sample models and weights.* Paper presented at the Proceedings of the Section on Survey Research Methods.

Skinner, C. J. (1989). Domain means, regression and multi-variate analysis. In: Skinner, C. J., Holt, D., & Smith, T. M. F., (eds). *Analysis of Complex Surveys*, 59-87.

Skrondal, A., & Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika, 68*(2), 267-287.

Smith, T. F. M. (1984). Present position and potential developments: Some personal views: Bayesian statistics. *Journal of the Royal Statistical Society. Series A (General), 147*(2), 245-259.

Smith, T. M. F. (1988). To weight or not to weight, that is the question. *Bayesian statistics, 3*, 437-451.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9*(4), 475-502.

Stapleton, L. M. (2012). Evaluation of conditional weight approximations for two-level models. *Communications in Statistics: Simulation & Computation, 41*(2), 182-204.

Stapleton, L. M. (2013). Incorporating sampling weights into single-and multi-level models. In L. Rutkowski, M. v. Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment.* (pp. 353-388). London: Chapman Hall/CRC Press.

Stapleton, L. M., Harring, J. R., & Lee, D. (2016). Sampling weight considerations for multilevel modeling of panel data. . In J. R. Harring, Stapleton, L. M. & Beretvas, S. N. (Ed.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*. (pp. 63-95). Charlotte, NC.: Information Age Publishing, Inc.

Stapleton, L. M., & Kang, Y. (2016). Design effects of multilevel estimates from national probability samples. *Sociological methods & research*, 1-28.

StataCorp, L. (2013). Stata multilevel mixed-effects reference manual. *College Station, TX: StataCorp LP*.

Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71*, 495-506.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Hagedorn, M. C., Daly, P., & Najarian, M. (2015). Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011). User's Manual for the ECLS-K: 2011 Kindergarten Data File and Electronic Codebook, Public Version. NCES 2015-074. *National Center for Education Statistics*.

Vieira, M. D. T., & Skinner, C. J. (2008). Estimating models for panel survey data under complex sampling. *Journal of Official Statistics, 24*(3), 343-364.

West, B. T., & Galecki, A. T. (2011). An overview of current software procedures for fitting linear mixed models. *The American Statistician, 65*(4), 274-282.

Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological methods & research, 23*(2), 230-257.

Xia, Q., & Torian, L. V. (2013). To weight or not to weight in time-location sampling: why not do both? *AIDS and Behavior, 17*(9), 3120-3123.

Zhu, M. (2014). Analyzing Multilevel models with the GLIMMIX procedure. *SAS Institute Inc.*

SAS026-2014.

CHAPTER 2
COMPLEX SAMPLING DESIGN IN MULTILEVEL MODELING

2.1 Introduction

In the era of evidence-based research, large-scale survey data have been increasingly utilized to provide solid evidence. Due to budgetary constraint and practical convenience, these high-quality data have been collected via complex sampling designs with some special features such as clustering, unequal selection of probabilities, stratification and non-response. For example, concerning international surveys, the basic sample design used in PISA is a stratified two-stage sample design that samples school at the first stage and students within schools at the second stage. In other international surveys such as The Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) classrooms may be sampled within schools (Martin & Mullis, 2012; OECD, 2012). National surveys such as the National Assessment of Educational Progress (NAEP), the Educational Longitudinal Studies (ELS), and the National Educational Longitudinal Studies (NELS) in general also utilize a stratified two-stage sampling design (Stapleton & Kang, 2016). However, the adoption of complex sampling designs may pose some challenges in statistical modeling. For instance, in the education context, the sampling of units at different stages produces data with clustering structures (e.g., students nested in classrooms or schools, etc.). This nesting structure violates the basic model assumption in regression models that the errors are identically and independently distributed (i.e., iid). This implies that running student-level regression analysis without adjusting the standard errors of the estimates for potential clustering effects may result in spurious significance due to spurious (smaller) standard errors, narrower confidence intervals and smaller p-values (increased likelihood of Type I error).

Multilevel models have been widely used in education in the past three decades to account for clustering (i.e., adjusting standard errors of the estimates), partition the outcome variance into components at various levels and estimate teacher and school effects separately (Lee & Fish, 2010; Lubienski & Lubienski, 2006; Palardy, 2010; Snijder & Bosker, 2012). Multilevel models have been used on both cross-sectional and longitudinal data where in the latter measurement occasions are nested within individual subjects. The basic multilevel models include random intercepts models; however, random slopes models have also been utilized. There are two general estimation methods to compute regression estimates and variances: the maximum likelihood (ML) and the restricted maximum likelihood (REML) (McCulloch, Searle, & Neuhaus, 2008; Raudenbush & Bryk, 2002).

In addition to clustering, unequal probability of selection is another issue that needs to be considered as it leads to an informative design, that is, the response variable is associated with sample selection even after controlling for covariates in the model. A high informative design implies that estimates from sample data could differ from the population parameters. For instance, if schools are sampled with a probability that is proportional to school size, which means larger schools would have a higher probability of being selected, and students score higher in these larger schools, ignoring this proportional sample would lead to biased population estimates. Therefore, to incorporate an informative design, applying sampling weights is a convenient and simple approach that has been recommended. The present study deals with the two features of clustering and unequal probability of selection via incorporating sampling weights in multilevel models.

2.2 Literature background

Investigating how to conduct model analysis on complex survey data has been of great interest in the literature (Firth & Bennett, 1998; Holt, Smith, & Winter, 1980; Konijn, 1962; Magee, 1998; Pfeffermann & Holmes, 1985; Pfeffermann & LaVange, 1989; Pfeffermann & Smith, 1985; A. Scott & Smith, 1969; Vella, 1998; Wedel, ter Hofstede, & Steenkamp, 1998; Wu, 2007). Different estimation methods for complex sample designs have been explored such as least squares, conditional empirical likelihoods, and pseudo empirical likelihoods (Chaudhuri, Handcock, & Rendall, 2010; Chen & Sitter, 1999; Francisco & Fuller, 1991; Fuller, 1984; Lin, Steel, & Chambers, 2004; Rao & Wu, 2010; A. J. Scott & Holt, 1982).

Sampling weights is a focal point that has received considerable attention and triggered much discussion (Pfeffermann, 1993, 1996). Regarding the sampling weights, the first long-term debated question is "whether to weight or not to weight" (Bertolet, 2008; DuMouchel & Duncan, 1983; Kish, 1992; C. Skinner, 1994; Smith, 1988). Theoretically, there are two fundamentally opposite schools of thought on making inferences from survey data: the design-based approach and the model-based approach (D. Binder & Roberts, 2006; D. A. Binder & Roberts, 2009; D. A. Binder & Roberts, 2003; Godambe & Thompson, 1986; Hansen, Madow, & Tepping, 1983; Rao, 1997; Rao & Bellhouse, 1990; Rao et al., 1999; Särndal et al., 1978). The design-based (or randomization sampling) approach is traditionally adopted to produce design-unbiased estimates of population quantities such as the mean, the ratio, and the total. The assumption of this approach is that the estimates for the finite population are fixed, thus the uncertainty exclusively comes from sampling error. In contrast, the model-based approach is typically used to conduct analyses for statistical inference producing regression coefficients, and the corresponding standard errors for population relationships. The assumption of the model-based approach is

about stochastic processes in which the data follow random distributions, thus the model estimates could be projected to the population regardless of sampling designs and features if the model is specified correctly.

In practice, data collection and model estimation cannot be separated. When research interests go beyond knowing descriptive statistics, a model-based analytic approach becomes essential. However, to many educational researchers it is still unclear when and how to use sampling weights appropriately in statistical models, especially in multilevel modeling. Comparatively speaking, incorporating sampling weights in single-level models has been a well-established procedure. Conventional design-based methods include Taylor series linearization, balanced repeated replication (BRR), Jackknife repeated replication (JRR) and bootstrap to produce least-biased variance estimation (Cohen, Burt, & Jones, 1986; K. Rust, 1985, 2013; K. F. Rust & Rao, 1996). For example, PISA adopts BRR (OECD, 2014) and TIMSS & PIRLS use JRR (Foy, 2014). General software commands such as "svy" commands in STATA and "surveyreg" in SAS have been used for single-level population-average or marginal models using complex survey data. Specialized software programs also have been developed, among which SUDAAN and PC Carp are widely used (Cohen et al., 1986; LaVange, Steams, Lafata, Koch, & Shah, 1996; Rodgers-Farmer & Davis, 2001). In model-based approaches, the pseudo maximum likelihood (PML) estimation method has been used to produce consistent estimates ((D. A. Binder, 1983; Gourieroux, Monfort, & Trognon, 1984; Krieger & Pfeffermann, 1992; C. J. Skinner, 1989). The idea of PML is to replace the population log-likelihood by a sample weighted likelihood and then sum across clusters assuming clusters are independent. Although PML is not real likelihood, it resembles a census population likelihood. Since the estimator from a census population likelihood is consistent, the PML estimator is also assumed to be consistent.

In the framework of multilevel modeling, the dependency of individual units within clusters gives arise to estimation difficulties and therefore the cluster level random effect needs to be integrated out in order to apply the idea of PML in multilevel models. Concerning the widely-used approach of probability weighted estimation, three techniques have been proposed to cope with complex sampling design: the method of moments, the maximum likelihood, and the least squares. Graubard and Korn proposed the weighted ANOVA estimators for one-way random-effects models, which is a specific approach with very limited applicability and there is no command available to perform this approach in any statistical software (Graubard & Korn, 1996; Jia, Stokes, Harris, & Wang, 2011; Korn & Graubard, 2003). In addition, there are two general estimation methods that have been proposed. Pfeffermann et al. (1998) introduced the method of probability-weighted iterative generalized least squares (PWIGLS) based on the iterative generalized least square (IGLS) developed by Goldstein (1986) (Goldstein, 1986; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). Concurrently, Rabe-Hesketh & Skrondal (2006) (Rabe-Hesketh & Skrondal, 2006) and Asparouhov (2006) (T. Asparouhov, 2006) proposed the MPML estimation method. In general, MPML and PWIGLS are preferred to the weighted ANOVA approach because of wider applicability in software packages. Specifically, STATA, Mplus and SAS implement MPML, while LISREL, HLM and MLwiN implement PWIGLS (Chantala & Suchindran, 2006; West & Galecki, 2011). It should be noted all these studies used a two-level random intercept model without covariates in their simulation investigation.

There is no consensus upon which estimation method works the best across different settings. Some prior studies compared the approaches of single-level and multilevel models on fixed effects (Koziol, 2016; Stapleton, 2013). Within multilevel models, through a simulation

investigation, Cai (2013) found that PWIGLS provided better estimates than the MPML whereas Asparouhov and Muthen (2007) (Tihomir Asparouhov & Muthen, 2007) found MPML worked better than PWIGLS. The comparison of the performance between MPML and PWIGLS is not straightforward. First of all, the evaluation of the performance of the two estimation methods depends on specific simulation scenarios such as informative designs and models (e.g., latent variable model or nonlinear model). Second, raw sampling weights at lower levels cannot be used as is and scaling procedures need to be applied. Nevertheless, the implementation of scaling in software programs is inconsistent. For instance, STATA provides three scaling options (i.e., "size", "effective" and "gk") and Mplus has two (i.e., "cluster" and "eclsuter") while LISREL only has one scaling choice. Moreover, the scaling in STATA and Mplus applies only to the lower-level of sampling weights whereas in LISREL it applies at both levels. Third, the scaling procedure as some functions of sample size may interact with other factors (e.g., ICC) to influence the performance of the probability weighted estimation method, which may add additional complexity.

2.3 Conceptual framework of sample weighted approach

The diagram in Figure 2.1 depicts one conceptual framework to help decide whether and when to apply sampling weights in multilevel models. Firstly, the answer depends on the availability of level-specific weights to be applied in multilevel models. For instance, when only overall sampling weights have been provided in the data set, a weighted single-level modeling approach is recommended (T. Asparouhov, 2006). Because if sampling weights are missing at some levels, applying weights at one level but not all could produce more bias than unweighted estimates (Grilli & Pratesi, 2004). Secondly, it also depends on whether the sample design is

informative or non-informative. In non-informative designs, unweighted estimators would be preferred with the advantage of providing efficient, consistent and unbiased estimates. In informative designs, sampling weights should be incorporated to represent population quantities. Nevertheless, determining design informativeness and its degree in practice is not an easy task. Thirdly, the decision depends on the performance of the weighted estimation method used (Bertolet, 2008). Specifically, the performance of different estimation methods depends on several factors such as design informativeness, scaling methods, sample size, intra-class and correlation (ICC) (T. Asparouhov, 2006; Cai, 2013; Jia et al., 2011; Pfeffermann, Skinner, et al., 1998). Currently, how these factors affect the performance of MPML is inconclusive and simulation evidence is scarce. Moreover, there is very little guidance regarding incorporating sampling weights into multilevel models including software application.

The impact of the informativeness of the design on statistical inference has received considerable attention (D. A. Binder, Kovacevic, & Roberts, 2005; Sugden & Smith, 1984). Previous studies have found that whether the sampling design is informative and the stage at which the sampling design is informative has substantial impacts on the estimation (Cai, 2013; Pfeffermann, Skinner, et al., 1998). For instance, when level 2 is informative while level 1 is non-informative, the unweighted estimates of the intercept and the second level variance will be biased, but the individual level variance is unbiased. Ignoring an informative sampling design at the first stage will result in biased estimates on the intercept and variance of random effects whereas ignoring an informative sampling design at the second stage will lead to slightly underestimated fixed effects and residual variance at lower level. When both levels are non-informative, estimates of involved parameters are least biased compared with informative designs.

Scaling of individual-level sampling weights is a primary tool of bias reduction, in which the weights at the individual level will be normalized to reflect some functions of cluster sample size. Two scaling methods have been proposed to provide the least biased estimates: "size" scaling and "effective" scaling. It is noteworthy that different scaling methods may work differently with different estimation techniques ((Potthoff, Woodbury, & Manton, 1992). There is no conclusion on which scaling method works better. For example, Pfeffermann et al. (1998) tentatively recommend the "size" scaling rather than the "effective" scaling whereas Stapleton found that the "effective" scaling method provided unbiased estimates of key parameters and their sampling variance while the "size" scaling produced negatively biased variance estimates (Stapleton, 2002).

Sample size is another factor involved in data collection and analysis. Asparouhov (2006) found that the multilevel weighted estimation is approximately unbiased when the cluster sample size is large (i.e., 100), but when cluster sample size is small (i.e., 5), MPML may produce relatively small bias even if the level 1 weights are non-informative given 100 cluster units. Korn and Graubard (2003) mentioned that even when using rescaled weights, survey weighted estimators could be badly biased for estimating variance components especially with smaller cluster sample size.

The ICC value is an important index in multilevel modeling, which in a two-level model is computed as the ratio of the second level variance over the total sampling variance. A relative large ICC value ensures that the cluster level variance is of great importance to include in multilevel model analyses. In addition, the ICC value is also informative for planning group-randomized experiments in education (Hedges & Hedberg, 2007). Intuitively, small ICC values

mean that the between-school variation is less influential than the within-school variation which translates to a smaller clustering effect.

2.4 The present study

There are a couple of literature gaps regarding the multilevel pseudo-likelihood approach. Theoretically, although this approach has some advantages in terms of producing consistent estimates with computation convenience, its weakness is evident. Specifically, the main caveat is that it produces larger error variances than the unweighted analyses, which affects statistical inference. In addition, the variance estimates may be inaccurate because the distribution of the weighted point estimator is in general unknown.

Practically, first of all, evidence of the performance of the two estimation techniques (i.e., MPML and PWIGLS) is very limited. Moreover, as the multilevel model involves scaling which is inconsistent across software programs, the comparison between these two methods is unclear and inconclusive. Relatively speaking, MPML outperforms PWIGLS in terms of computational simplicity (Kovačević & Rai, 2003) and software applications (Leite et al., 2015). Second, previous simulation studies are informative, but the findings are restricted to particular designs which are likely to be irrelevant and inapplicable to large-scale data in education. For example, Cai (2013) planned one informative sampling design in which the first-stage is non-informative while the second stage is informative, which is very rare. On the contrary, large-scale educational data sets typically would adopt unequal probability of selection at the first stage and that could be the driving force for the design informativeness.

Another approach to incorporate the informative sampling design in the multilevel modeling is the sample distribution approach (hereafter SDA), which has been proposed in the

1990s. This approach has much less application in practice due to lack of software implementation. Currently there is no software available to implement it directly. Instead the numerical technique and calculation is required, which may be beyond the competence of many empirical researchers.

To address the aforementioned issues, the present research consists of two investigation components corresponding to PML and SDA. The first investigation attempts to conduct a Monte Carlo simulation to examine the performance of MPML estimation method to see how the sampling weights, design informativeness and scaling procedure might affect its behavior. The quality of the parameter estimators will be discussed in terms of their relative bias, root mean square errors and coverage rates. The second investigation will use the exponential model to approximate the conditional expectation of the sampling selection and then fit a parametric model using ECLS-K data in 2011 via Bayesian analysis framework with a model checking component. The results obtained from MPML and SDA will be compared.

This study has potential contributions as follows. Firstly, it will shed light on providing some guidelines regarding when and how to incorporate complex sampling weights in multilevel modeling in the context of large-scale educational data. One current issue is that survey providers recommend that sampling weights should be incorporated into analysis when using large-scale educational data, but the conventional design-based approach they introduced such as BRR and JRR is incompatible with multilevel modeling. There is a lack of guidelines in practice and analysts could be easily confused and make mistakes.

Secondly, it will advance the methodology on both MPML and SDA. Regarding the MPML, the first study will evaluate the performance of the MPML estimation method and assess the effects of weighting, scaling and informative design via Monte Carlo simulation in the

context of large-scale educational data sets which the evidence has not been provided yet in the literature. With regard to the SDA, limited by the software availability, this approach has not been much studied yet. Therefore, exploring how to take advantage of this approach and compare with the sampling weights approach would be informative.

2.5 Investigation component I: MPML

2.5.1 Statistical model

Following past key studies, a two-level random intercept model with covariates in which individual student i in school j can be written as

$$Y_{ij} = \beta_0 + \boldsymbol{COV}_{(i)j}\boldsymbol{B_1} + u_j + e_{ij} \qquad u \sim N(0, \sigma_a^2) \quad \varepsilon \sim N(0, \sigma_e^2) \tag{A2.1}$$

where Y_{ij} is the outcome, β_0 represents the overall population intercept, $\boldsymbol{COV}$ refers to a row vector of covariates at the student (i.e., female and SES) and the school levels (i.e., class size and private school), Greek letter $\boldsymbol{B}$ represents a column vector of covariate coefficients for the fixed-effect parameters, u is a second-level residual and ε is the first-level residual. Both u and e are random-effect parameters, which assume to follow normal distributions with zero means and variances σ_a^2 and $\boldsymbol{\sigma_e^2}$ respectively.

Suppose $\hat{\theta}$ represents all parameters to be estimated which includes B, σ_a^2 and σ_e^2.

The conditional normal likelihood for student i in school j can be expressed as:

$$L_{ij}(\theta|y_{ij}) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left[-\frac{(y_{ij}-\hat{\bar{y}}_j)^2}{2\sigma_e^2}\right] \tag{A2.2}$$

Where $\hat{\bar{y}}_j$ is the estimated group mean for the j^{th} school. Thus, the marginal likelihood for school j is

$$L_j(\theta) = \int_{-\infty}^{+\infty} \prod_{i=1}^{N_j} L_{ij}(\theta|y_{ij})\phi(u_j)du_j \tag{A2.3}$$

where N is the number of observations in jth cluster in the population and ϕ is the density function.

And the overall marginal likelihood is

$$L(\theta) = \prod_{j=1}^{M} L_j(\theta) \quad \text{(A2.4)}$$

where M represents the number of clusters in the population.

For computational convenience, the log-likelihood form (here denoted by "*l*") is typically used, and the unknown parameters $\hat{\theta}$ can be obtained via either the maximum likelihood estimation (MLE) or the restricted maximum likelihood (REML) estimation methods.

The log-likelihood expression can be written as

$$l(\theta) = \sum_{j=1}^{M} log \int_{-\infty}^{+\infty} \{\exp [\sum_{n=1}^{N_j} log L_{ij} (\theta|u_j)]\} \phi(u_j) d_{u_j} \quad \text{(A2.5)}$$

Equations (1) to (5) denote the population model and the population likelihood. In practice, only sample data will be available instead of population data. From a model perspective, the data are assumed to be random with independent distributions. However, large-scale data seldom use simple random sampling selection. Instead, in complex sampling designs samples are selected in a particular way such as an unequal probability of selection.

Suppose the data are collected using a two-stage sampling design and the probability of selection at the first stage is p_j and the conditional probability of selection at the second stage is $p_{i|j}$. The corresponding weights at the first and second stages are w_j and $w_{i|j}$ respectively. Then, the next step is to apply the level-specific sampling weights into the log-likelihood function to account for potential effects due to the unequal probability of selection. In a single-level model, PML can be expressed as

$$l(\theta) = \sum_{j=1}^{m} w_j log\, L_j(\theta). \quad \text{(A2.6)}$$

However, in a multilevel model, PML cannot be applied directly because the individuals within each cluster are dependent, so the sample weighted log-likelihood cannot be summed up directly across clusters. Instead, integration of cluster variation needs to be done. Specifically, the MPML can be written as

$$l(\theta) = \sum_{j=1}^{m} w_j log \int_{-\infty}^{+\infty} \{exp\,[\sum_{n=1}^{n_j} w_{i|j} log L_{ij}\,(\theta|u_j)]\}\phi(u_j)d_{u_j} \quad \text{(A2.7)}$$

It is interesting that equation (7) is consistent with the Horvitz-Thompson principle (Horvitz & Thompson, 1952), a design-based approach, which replaces each sum over the cluster-level population unit *j* with a sample weights of $\frac{1}{p_j}$ and each sum over the level 1 units *i* by the sample weights of $\frac{1}{p_{i|j}}$ as follows. This shows that the design-based and model-based approach are compatible as equations (1.7) and (1.8) show.

$$l(\theta) = \sum_{j=1}^{m} \frac{1}{p_j} log \int_{-\infty}^{+\infty} \{exp\,[\sum_{n=1}^{n_j} \frac{1}{p_{i|j}} log L_{ij}\,(\theta|u_j)]\}\phi(u_j)d_{u_j} \quad \text{(A2.8)}$$

Previous studies have suggested that some scaling procedure is necessary for the individual-level sampling weights, which is referred to the λ_1 in the equation below.

$$l(\theta) = \sum_{j=1}^{m} w_j \lambda_2 log \int_{-\infty}^{+\infty} \{exp\,[\sum_{n=1}^{n_j} w_{i|j} \lambda_1 log L_{ij}\,(\theta|u_j)]\}\phi(u_j)d_{u_j} \quad \text{(A2.9)}$$

Although there is no theoretical result to show the gold standard scaling method, two scaling methods have been proposed to provide the least biased estimates (T. Asparouhov, 2006; Pfeffermann, Skinner, et al., 1998; Potthoff et al., 1992; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2002). The first scaling method scales the level-1 weights to the actual cluster sample size (n_j) and the second method scales the level-1 weights to its effective cluster size. Here they are referred as "size" and "effective" scaling following the name in STATA.

$$\lambda_{1_{size}} = \frac{n_j}{\sum_{i=1}^{n_j} w_{i|j}} \quad \text{(A2.10)}$$

$$\lambda_{1effective} = \frac{\sum_{i=1}^{n_j} w_{i|j}}{\sum_{i=1}^{n_j} {w_{i|j}}^2}. \quad \text{(A2.11)}$$

It should be noted that when $w_{i|j}$ is a constant, these two scaling methods are equal. To illustrate this point, I provide a simple example. Suppose 100 observations are selected from a total population of 1000. The probability of selection is 1/10 and the weights, $w_{i|j,}$ are equal to 10. Plugging this into equation (1.10) and (1.11) leads to the same result.

$$\lambda_{1_{cluster}} = \frac{n_j}{\sum_{i=1}^{n_j} w_{i|j}} = \frac{100}{(10+10+\cdots+10)} = \frac{100}{100*10} = 0.10$$

$$\lambda_{1Ecluster} = \frac{\sum_{i=1}^{n_j} w_{i|j}}{\sum_{i=1}^{n_j} {w_{i|j}}^2} = \frac{10+10+\cdots+10}{10^2+\cdots+10^2} = \frac{100*10}{100*10^2} = 0.10$$

There is a third scaling method, which is a little different from the "size" and "effective" scaling methods. Specifically, in the latter two methods the scaling takes place at level 1 weights whereas in the third scaling method the scaling takes place at level 2 weights w_j as in equation (1.12) below. It is derived from the moments estimators of the weighted ANOVA approach and it is referred to as the "gk" scaling in STATA (Graubard & Korn, 1996; Korn & Graubard, 2003).

$$w_j^* = \sum_{i=1}^{n_j} w_{i|j}\, w_j;\ w_{i|j}^* = 1 \quad \text{(A2.12)}$$

With regard to the computation technique, Asparouhov (2006) mentioned that MPML as a general estimator can be obtained via any optimization algorithm such as the EM-algorithm, the accelerated EM algorithm, and the Quasi-Newton algorithm. Nevertheless, there is no closed form solution for the MPML estimator for a random intercept model when the cluster sample size is unbalanced. However, assuming balanced data, the unweighted maximum likelihood estimators (MLE) have been provided (see McCulloch et al., 2008).

When σ_a^2 is positive, the analytical expressions of the estimators are as follows.

$$\widehat{\beta_0} = \bar{y}_{..}, \widehat{{\sigma_e}^2} = \frac{\sum_j \sum_i (y_{ij}-\bar{y}_j)^2}{m(n-1)}, \widehat{\sigma_a^2} = \frac{\sum_j (\bar{y}_j-\bar{y}_{..})^2}{m} - \frac{\widehat{\sigma_e^2}}{n} \quad \text{(A2.13)}$$

where $\bar{y}_{..}$ is the grand mean, $\bar{y}_j$ is the cluster mean, m is number of clusters and n is number of units in each cluster, which is a constant in balanced data.

Using the Laplace approximation, a well-known method for approximating the marginal densities, Asparouhov (2006) derived a closed form solution for the parameters of a random intercept model without any covariates when cluster sample size is constant across all clusters (i.e., balanced design). Using consistent notation in this study, the analytical expressions for the unscaled case can be expressed as follows

$$\widehat{\beta_0} = \frac{\sum_j w_j \bar{y}_j}{\sum_j w_j}, \widehat{\sigma_e^2} = \frac{\sum_j w_j \sum_i w_{i|j}(y_{ij}-\bar{y}_j)^2}{\sum_j w_j(n-1)}, \widehat{\sigma_a^2} = \frac{\sum_j w_j(\bar{y}_j-\bar{y}_{..})^2}{\sum_j w_j(n-1)} - \frac{\widehat{\sigma_e^2}}{n}. \quad \text{(A2.14)}$$

The "size" scaled case as

$$\widehat{\beta_0} = \frac{\sum_j w_j \bar{y}_j}{\sum_j w_j}, \widehat{\sigma_e^2} = \frac{\sum_j n\, w_j \sum_i w_{i|j}(y_{ij}-\bar{y}_j)^2}{\sum_j w_j(n-1)\sum_i w_{i|j}}, \widehat{\sigma_a^2} = \frac{\sum_j w_j(\bar{y}_j-\bar{y}_{..})^2}{\sum_j w_j(n-1)} - \frac{\widehat{\sigma_e^2}}{n}, \quad \text{(A2.15)}$$

and the "effective" scaled case as

$$\widehat{\beta_0} = \frac{\sum_j w_j \bar{y}_j}{\sum_j w_j}, \widehat{\sigma_e^2} = \frac{\sum_j(\sum_i w_{i|j})\, w_j \sum_i w_{i|j}(y_{ij}-\bar{y}_j)^2}{\sum_j w_j(n-1)(\sum_i {w_{i|j}}^2)}, \widehat{\sigma_a^2} = \frac{\sum_j w_j(\bar{y}_j-\bar{y}_{..})^2}{\sum_j w_j(n-1)} - \frac{\widehat{\sigma_e^2}}{(\sum_i w_{i|j})^2/(\sum_i w_{i|j}^2)}. \quad \text{(A2.16)}$$

With regard to the asymptotic covariance matrix, Rabe-Hesketh and Skrondal (2006) used the following sandwich form: $cov\left(\hat{\theta}\right) = I^{-1}JI^{-1}$ where "I" represents the observed (pseudo) Fisher information at the pseudo-maximum likelihood estimates and asymptotically the J is $\left(\sum_j (\lambda_2 w_j)^2 l_j' l_j'^T\right)$ where l' refers to the first derivative of the log-likelihoods.

2.5.2 Simulation setting

The simulation is carried out in the context of large-scale educational data which has special features. First, large-scale data sets have relatively large sample size from approximately 4,000 to 20,000. For example, in PISA 2015, the average number of students per school is 33

across 70 countries/economies. Since the cluster sample size is greater than 20 on average, previous findings about estimation bias relating to small sample size such as cluster sample size less than 10 may not be an issue for large-scale data sets. Second, large-scale data sets in education typically adopt a two-stage sampling design. For example, at the first stage schools are selected with an unequal probability and then at the second stage classrooms or students are selected with either an unequal probability selection or using simple random sampling. Therefore, the possible informative designs would be school (informative) and classroom/student (informative/non-informative). The school level selection seems to be the driving force of the design informativeness compared with student level selection. There are also three-stage sampling designs. Stapleton and Kang found that ignoring the sampling design beyond the levels in the model would have minor effects when using five public large-scale data sets from NCES (Stapleton & Kang, 2016).

This simulation investigation differs from previous studies in several ways. First, the simulation setup uses an informative design at the second level with some variation, while at first level the informative design is fixed. I intentionally omit the case where the second sampling stage is non-informative which is possible but is not of interest because for non-informative designs the sampling weights are not recommended. In addition, I fixed the ICC value at 0.20, which is the typical value in U.S. data. Moreover, based on PISA data which on average selects 35 students per school, I decided to have a balanced design with the cluster sample size set at 35. Cai (2013) examined four informative designs: informative-informative, informative-noninformative, noninformative-informative, and noninformative-noninformative. He simplified the simulation by fixing the sample size of 100 with clusters of size 50 and choosing a moderate ICC value as 0.33. Comparing with Cai (2013), it is pronounced that my simulation setting is

more realistic. For instance, it is very unlikely that large-scale surveys would employ simple random sampling at the first stage. Second, this study focuses on a multilevel linear model with a continuous outcome, which makes it different from Rabe-Hesketh & Skrondal (2006) and Grilli & Pratesi (2004) that examined the MPML estimation method in the context of logistic model and probit model respectively. Moreover, this simulation is different from the studies of Pfeffermann et al. (1998) and Asparouhov (2006). Pfeffermann et al. (1998) examined the performance of PWIGLS and their two-level model did not include any covariates. Asparouhov (2006) did not include covariates in the model either.

The sampling design mimicked the PISA data, in which schools were selected with probabilities that are proportional to school size. Suppose the school population is M=6000 and m=120 schools will be sampled with selection probability of 0.02 that are proportional to school size, which is defined as an exponential function of normal distribution for u_j with a mean of 0 and a variance of σ_a^2 . This means that schools with large variance are large schools with high selection probability. The selection probability (p_j) is then computed as p_j=1/(1+exp(4-u_j/x)) where x=(1/3, ½, 1, 2) to reflect informativeness from high to low. The weights w_j are computed as the inverse of p_j. At the second sampling stage, suppose there are 140 students in a certain grade and 35 students (i.e., 1/4) were randomly selected with probability of $p_{i|j}$ as $p_{i|j}$ =1/(1.35+exp(1-e_{ij}/2)) and the weights are the inverse of $p_{i|j}$. The individual student residual e was generated from a normal distribution with a mean of 0 and a variance of σ_e^2. The ICC is set as 0.20 with σ_a^2 =0.0625 and σ_e^2=0.25. The finite population values Y_{ij} were generated from the following model. The appendix 1A provides the STATA code. The Y_{ij} is the difference score in reading scores for each kindergartener between the spring and the fall. Female and private school are dummy variables with Bernoulli distributions while SES and class size follow normal

distributions. The "unusual" negative coefficients for the variables SES and private school show that students from a high SES family background or students who attend private schools have a smaller rate of change given their potential higher score to begin with.

$$Y_{ij} = 1 + 0.019female - 0.065 * SES - 0.001 * class\ size - 0.026 * private\ school\ + u_j + \varepsilon_{ij} \quad \text{(A2.17)}$$

Once the population is generated then sample data are selected based on p_j and $p_{i|j}$. The process was repeated 1000 times. Following Eideh and Nathan (2009) and Cai (2013), the quality of estimates is evaluated using three criteria: the empirical relative bias, the empirical root mean square error (RMSE) and the coverage rate that the true parameter falls within the 95% confidence interval using t-test based standard errors.

The relative bias is defined as

$$RBias = \frac{1}{\theta}\left[\frac{1}{1000}\sum_{i=1}^{1000}\left(\widehat{\theta_i} - \theta\right)\right] \quad \text{(A2.18)}$$

and the RMSE is expressed as

$$RMSE\left(\widehat{\theta}\right) = \sqrt{\left[\frac{1}{1000}\sum_{i=1}^{1000}(\widehat{\theta_i} - \bar{\widehat{\theta}})^2\right]} \quad \text{(A2.19)}$$

where $\bar{\widehat{\theta}} = \frac{1}{1000}\sum_{i=1}^{1000}\widehat{\theta_i}$.

2.5.3 Results

Results for simulation means

Table 2.1 reports the point estimators produced from the simulation for four fixed effects concerning the variables female, SES, class size and private school. Overall the estimates cover the true values except for private school where the estimates seem to have some bias. Especially, the variables with normal distributions (i.e., SES and class size) have more robust mean values

than the binary variables (i.e., female and private school) across designs with different degree of informativeness. In addition, weighted estimates seem to be a little different from unweighted estimates by and large in the multilevel framework. Nevertheless, the two scaling methods have similar performance.

Table 2.2 presents the point estimators produced from the simulation for the intercept and the variance components. With regard to the intercept β_0, the overall unweighted estimates produced positive bias while applying the weights helped correct this bias. With regard to σ_a^2, the results show that weighted but unscaled estimates have positive bias compared with unweighted or scaled estimates. For scaled estimates, the values fluctuate slightly around the true value across different informative designs. Regarding the scaling, the effective scaling produced slightly smaller estimates than the size scaling. Regarding σ_e^2, all four estimators have a slight negative bias. Comparatively speaking, the scaled estimates are preferred; weighted but unscaled estimates are the worst and the unweighted estimates lie in between.

Quality of simulated point estimates

Figures 2.2 to 2.4 along with Table 1.1 demonstrate the quality of the estimation concerning the fixed effects of the four variables in terms of relative bias, RMSE and 95% coverage rate. With regard to the relative bias, results evidently are affected by variation on design informativeness. In general, there is no clear pattern about which estimation type among the four generates the least biased estimates across variables. Regarding RMSE, the unweighted estimates consistently produce the lowest values compared to the three weighted counterparts, which have similar values. In terms of the 95% coverage rate, by and large, the performance of

the estimation is good and consistent across different settings. All variables have the values are all pretty close to 95%.

Figures 2.5 to 2.7 along with Table 2.2 present results for the intercept and the variance component. With respect to the relative bias, for β_0, the unweighted estimates have substantial positive bias and applying sampling weights reduces this bias. Once weights are applied, scaling does not seem to have additional benefits for β_0. For σ_e^2, the weighted but unscaled estimator has the greatest amount of negative bias and the scaling procedure helps to reduce this bias. For σ_a^2, the weighted but unscaled estimator generates substantial positive bias whereas unweighted or scaled estimators would be preferred. It seems that size scaling slightly outperforms the effective scaling. Corresponding to the analytic expressions in equations (A2.13) to (A2.16), when the predicted σ_e^2 has substantial negative bias, the estimator of σ_a^2 would have much positive bias. In terms of the RMSE, unweighted estimates have the lowest values. With regard to the 95% coverage rate, for β_0, the unweighted estimates have low coverage rates (from 64% to 86%); for σ_e^2, the weighted unscaled estimates have the lowest coverage rate (from 76% to 81%) where scaled estimates have coverage close to 95%; for σ_a^2, the weighted unscaled estimates have a relatively low coverage rate around 90%, whereas the other three counterparts all have coverage rates close to 95%.

Simulated standard error

Following Pfeffermann et al. (1998), the standard errors of the estimates were examined by comparing the average value of the 1000 estimated standard errors with the standard deviation of the 1000 point estimates for seven parameters including five fixed effect parameters (i.e., the intercept and four regression coefficients) and two random effect parameters (i.e., variance at

level one and level two). The closeness of these two indices indicates the good quality of the standard error estimation because by definition the standard error of the point estimate is the standard deviation of the sampling distribution of the point estimator. Table 2.3 contains results which show that the estimation of the standard errors performed very well. It is evident that the unweighted multi-level model produced the smallest standard errors for the point estimates across these parameters and the four informative designs.

2.6 Investigation component II: SDA

2.6.1 Background

To overcome the shortcoming of inefficiency (i.e., producing larger variance) in sample weighting approach, the method of SDA has been proposed. The basic idea of SDA is to extract a multi-level model from the sample data as a function of the corresponding population model as well as a sample selection probability at each level. The application and evidence of this method is very scarce for at least two reasons. First, there are several models (e.g., exponential and logistic models) that have been proposed to approximate the conditional expectations of selection probability at each level (Cai, 2013; Eideh & Nathan, 2009; Pfeffermann, Fernando Antonio Da Silva, & Pedro Luis Do Nascimento, 2006; Pfeffermann, Krieger, & Rinott, 1998). For instance, (Cai, 2013) used the logistic model for the conditional expectation of sampling. Pfeffermann et al., (1998) used polynomial and exponential models. Previous studies have indicated that SDA may not be that sensitive to the modeling of the conditional expectation of the inclusion probability, and due to simplicity and convenience, exponential models have been adopted more frequently (Eideh & Nathan, 2009). The notable feature of the exponential method is that the

joint sample likelihood could be simplified as a multivariate normal distribution with a shifted intercept while maintaining the same variance as in the population model.

Second, it is difficult to approximate the joint sample likelihood function, which involves numerical integration. Thus, it can be computational demanding for complex statistical models. Currently, there is no statistical software program available to perform it directly and researchers have to use numerical methods. For instance, Eideh & Nathan (2009) mentioned the "nlminb" function in S-PLUS but coding details were not provided. Cai (2013) used the Newton algorithm to approximate the joint sample likelihood function. The numerical methods were used to obtain the point estimates and then the variance of the point estimates were estimated via conventional methods (e.g., delta method and bootstrapping). Pfeffermann et al. (2006) used a Bayesian method to fit the resulting sampling model, but they did not provide model checking details, which is a crucial component for Bayesian analyses because the parameters rely on the convergence of the Monte-Carlo Markov Chain (MCMC) to be valid.

To summarize, the SDA is appearing as an alternative approach to overcome the limitations in the weighted probability approach. However, the difficulties of the estimation technique prevent it from moving forward and providing comparable empirical evidence. The study attempts to fill this gap via developing a Bayesian method to show the distribution of the variance of the point estimates after incorporating the design informativeness. This method coupled with the Bayesian framework could be a promising approach to incorporate design features in complicated multilevel models.

2.6.2 Statistical method

The same statistical model is used and expressed in a two-level format as follows.

Level 1: $y_{ij} = \beta_{0j} + \boldsymbol{B_1 STU} + e_{ij}$ i=1, 2, …,n_j $e \sim N(0, \sigma_e^2)$

Level 2: $\beta_{0j} = \gamma_{00} + \boldsymbol{\Gamma_1 SCH} + u_j$ j=1, 2, …, m $u \sim N(0, \sigma_a^2)$, (B2.1)

where in level 1, y_{ij} represents the outcome of student *i* in school *j*, β_{0j} is the school-level mean, ***STU*** stands for student level covariates including student's gender and socioeconomic status, $\boldsymbol{B_1}$ contains the corresponding coefficients, e is the first-level residual; in level 2 γ_{00} is the overall mean, ***SCH*** stands for school-level covariates and $\boldsymbol{\Gamma_1}$ refers to corresponding coefficients, and *u* is the second-level residual. Both *u* and *e* are random-effect parameters assumed to follow normal distributions with means zero and variances σ_a^2 and σ_e^2 respectively. Suppose $\hat{\theta}$ represents all the population parameters to be estimated which include γ_{00}, $\boldsymbol{B}$, $\boldsymbol{\Gamma}$, σ_a^2 and σ_e^2.

The sample pdf of y_{ij} can be expressed as a population pdf with an indictor I_{ij}. Let I_{ij} =1 indicate that a level 1 unit (e.g., student) in the population is selected and included in the sample and I_{ij} =0 otherwise. Using Bayes theorem, the sample pdf at level 1 can be expressed in a general form as

$$f_s(y_{ij}|X_{ij}) = f_p(y_{ij}|X_{ij}, I_{ij} = 1) = \frac{Prob(I_{ij} = 1|y_{ij}, X_{ij})}{Prob(I_{ij}=1|X_{ij})} f_p(y_{ij}|X_{ij}) \quad \text{(B2.2)}$$

where *X* represents covariates and the subscripts *p* and *s* stand for sample and population respectively.

Similarly, level 2 sample pdf of y_j can be written as

$$f_s(y_j|X_j) = \frac{Prob(I_j = 1|y_j, X_j)}{Prob(I_j=1|X_j)} f_p(y_j|X_j). \quad \text{(B2.3)}$$

Equations (B2.2) and (B2.3) demonstrate that the sample distributions can be regarded as a function of the population model and the sampling design. If the design is informative which means that the response variable is associated with the sampling selection given all the covariates

in the model, the sample distribution would be different from the population distribution. Therefore, in general, $Prob\left(I_j = 1\middle|y_j, X_j\right)$ is not equal to $Prob\left(I_j = 1|X_j\right)$.

Following (Pfeffermann, Krieger, et al., 1998) ,

$$Prob\left(I_{ij} = 1\middle|y_{ij}, X_{ij}\right) = E_p(\pi_{ij}|y_{ij}, X_{ij}) \quad \text{(B2.4)}$$

$$Prob\left(I_{ij} = 1|X_{ij}\right)=E_p(\pi_{ij}|X_{ij}) \quad \text{(B2.5)}$$

where $\pi_{ij} = prob(\left(I_{ij} = 1\middle|y_{ij}, Z_{ij}\right)$ which is different from $Prob\left(I_{ij} = 1\middle|y_{ij}, X_{ij}\right)$ and Z_{ij} denotes other design variables.

Similarly,

$$Prob\left(I_j = 1|y_j, X_j\right) = E_p(\pi_j|y_j, X_j) \quad \text{(B2.6)}$$

$$Prob\left(I_j = 1|X_j\right) = E_p\left(\pi_j\middle|X_j\right). \quad \text{(B2.7)}$$

Plugging equations (B2.4) - (B2.7) into equations (B2.2) - (B2.3), the final sample pdfs can be written as

$$f_s\left(y_{ij}\middle|X_{ij}\right)=\frac{E(\pi_{ij}|y_{ij},X_{ij})}{E_p(\pi_{ij}|X_{ij})} f_p\left(y_{ij}|X_{ij}\right) \quad \text{(B2.8)}$$

$$f_s\left(y_j\middle|X_j\right) = \frac{E_p(\pi_j|y_j,X_j)}{E_p(\pi_j|X_j)} f_p\left(y_j|X_j\right) \quad \text{(B2.9)}$$

The exact form of the conditional expectation of the sampling probability is usually unknown, but it can be approximated by exponentials via the Taylor series approximation. The exponential model has been adopted by several studies (Eideh & Nathan, 2009; Pfeffermann, Krieger, et al., 1998; Pfeffermann, Moura, & Silva, 2006; Pfeffermann & Sverchkov, 2007). Specifically,

$$E_p(\pi_{ij}|y_{ij}, X_j)=\exp\left(b_0 y_{ij} + B_1 X_j\right) \quad \text{(B2.10)}$$

and

$$E_p(\pi_j|y_j, X_j)=\exp\left(d_0 y_j + D_1 X_j\right). \quad \text{(B2.11)}$$

The coefficients can be estimated using sample data via the following relationships (see Pfeffermann & Sverchkov, 1999):

$$E(w_{i|j}|y_{ij}, X_{ij}) = 1/\, E_p(\pi_{ij}|y_{ij}, X_{ij}) = \exp\left[-\left(b_0 y_{ij} + B_1 X_{ij}\right)\right] \quad \text{(B2.12)}$$

$$E(w_j|y_j, X_j) = 1/E_p(\pi_j|y_j, X_j) = \exp\left[-(d_0 y_j + D_1 X_j)\right]. \quad \text{(B2.13)}$$

The appealing feature of the exponential approximation model is that the sample and population models both belong to the normal distribution, so the joint-sample likelihood function is simplified to a multivariate normal distribution with the mean shifted by a constant while the variance remains the same. In the specific model used in this study, the exponential approximation model could be written as

$$E_s\left(y_{ij}\middle|STU\right) = y_j + B_1 STU + b_0\sigma_e^2 \; ; Var\left(y_{ij}\middle|STU\right) = \sigma_e^2 \quad \text{(B2.14)}$$

$$E_s\left(y_j\middle|SCH\right) = \gamma_{00} + D_1 SCH + d_0\sigma_a^2 \; ; Var\left(y_j\middle|SCH\right) = \sigma_a^2. \quad \text{(B2.15)}$$

Eideh and Nathan (2009) pointed out that when $b_0, d_0 = 0$, the sample and populaiton distributions are the same and the sampling mechanism is ignorable. Based on equation (B2.14) and (B2.15) the value b_0 can be estimated by regressing $\log(w_{i|j})$ on y_{ij} and *STU* and d_0 by regressing $\log(w_j)$ on y_j and *SCH* in which the y_j is unknown but can be estimated by $\bar{y}_j$, the group mean for each cluster. To perform frequantist analyses, I used the empirical data of ECLS-K 2011 kindergarten year. The outcome is the difference in reading achievement scores in kindergarten between the spring and the fall; female, SES, class size and private school are the covariates/predictors. Corresponding to analysis component I, four models were employed (unweighted, unscaled, size-scaled and effective-scaled) and analyses were conducted in STATA 14. Using the real data, for SDA, the approximted values were $\hat{b}_0$ = -0.0397 and $\hat{d}_0$ = 0.0281, which will be utilized in the Bayesian analysis with flat priors. Below I delineate the Baysian methods used.

The difference between frequentist analyses and Bayesian analyses is twofold. First, philosophically, frequentists treat population parameter as fixed values while Bayesian analysts treat them as random variables with distributions of possible values or with uncertainty around their true value. Practically, frequentists provide point estimates and corresponding standard errors as summary statistics while Bayesian analysts provide different central tendency (e.g., mean, median and mode) and dispersion summaries for the posterior distribution. Second, the prior distribution is a unique or distinctive component of the Bayesian analysis as it allows analysts to include their own knowledge as a quantity to the model analysis in a formal way. However, the prior distribution is also a controversial element as it is subjective and arbitrary because different people may use different priors and consequently have potentially varied conclusions (Gelman, J. B., Dunson, Vehtari, & Rubin, 2013). Nevertheless, using flat priors which would not add additional information, the posterior distribution would be dominated by the likelihood function. Thus the mode of the posterior distribution, which maximizes the likelihood function, should be equivalent to the maximum likelihood estimator in the frequentist inference framework assuming unimodal distribution.

The unique tool of Bayesian analyses is the implementation of Monte Carlo Markov Chain (MCMC). This technique frees analysts from doing extreme complex integration that sometimes is impossible to obtain and instead replaces it with iterative work conducted by computers (Gill, 2008). For instance, obtaining the marginal posterior distribution needs to integrate out all the conditional parameters. The Metropolis-Hastings algorithm is a term for a family of Markov chain simulation methods that has been used to sample from any univariate distribution. The Gibbs sampler can be viewed as a special case of the Metropolis-Hastings and it has been used widely for multivariate distributions when practical application involves many

multidimensional problems. The Gibbs sampler samples from the distribution of each parameter in turn conditioning on the data and the current values of all the other parameters. Provided with a starting value, sequential samples build up a Markov Chain, until the algorithm converges to its stationary (equilibrium) distribution. The posterior then becomes just the joint distribution of all the parameters.

In the Bayesian framework, the posterior distribution, which is proportional to the product of the prior distribution and the likelihood function, can be expressed as

$$p(\theta|y) \propto p(\theta)f(y|\theta), \tag{B2.16}$$

where theta represents a vector of estimated parameters and *y* represents data. The likelihood function is the same as in the frequentist framework. As priors are assumed to be independent, the joint distribution $p(\theta)$ can be written as the products of all the priors as follows

$$p(\theta) = p(\gamma_{00}, B, \Gamma, \sigma_a^2, \sigma_e^2) = p(\gamma_{00})p(B_1)p(\Gamma_1)p(\sigma_a^2)p(\sigma_e^2). \tag{B2.17}$$

The likelihood term $f(y|\theta)$ is a sampling distribution of the data given the parameters of θ and the likelihood term can also be regarded as a mathematical function that needs to be maximized given that particular data with the expression below

$$f(\theta|y) = f(\gamma_{00}, B_1, \Gamma_1, \sigma_a^2, \sigma_e^2|y). \tag{B2.18}$$

Plugging equations (B2.17) and (B2.18) into equation (B2.16), for each data point, the posterior distribution is

$$p(\theta|y) \approx p(\gamma_{00},)p(B_1)p(\Gamma_1)p(\sigma_a^2)p(\sigma_e^2)f(\gamma_{00}, B_1, \Gamma_1, \sigma_a^2, \sigma_e^2|y). \tag{B2.19}$$

Then, for the whole data sets the posterior distribution is the joint distribution across clusters with hyper-parameters, namely

$$P(\theta|y) = \int \prod_{j=1}^{m} [N(\theta_j|\gamma_{00}, \sigma_a^2]p(\gamma_{00}, \sigma_a^2)d(\gamma_{00}, \sigma_a^2). \tag{B2.20}$$

This is the joint posterior distribution containing all the parameters. For the inference of each specific parameter, its marginal distribution can be obtained by integrating out the effect of all the other auxiliary parameters. Take γ_{00} for example,

$$P(\gamma_{00}|y) = \int\int\int\int f(\gamma_{00}, B_1, \Gamma_1, \sigma_a^2, \sigma_e^2|y) d_{B_1} d_{\Gamma_1} d_{\sigma_a^2} d_{\sigma_e^2}. \quad \text{(B2.21)}$$

One specific merit of Bayesian analysis is that it takes into account the effects of non-focal parameters in making an inference, which may thus produce some extra variation than the frequentist analysis. However, this multiple integration is typically intractable and the solution relies on the MCMC. In this study the Gibbs sampler is utilized which is appropriate for multivariate distribution. With initial values for parameter vector θ^0 =($\beta_0^0, \beta_1^0, \beta_2^0, {\sigma^2}_{a,}^0 {\sigma^2}_e^0$) (where the superscript represents the number of iteration), the sampling procedure is,

Step1: sample β_0^1 from $f(\beta_0|\beta_1^0, \beta_2^0, {\sigma^2}_{a,}^0 {\sigma^2}_e^0, y)$

Step2: sample β_1^1 from $f(\beta_1|\beta_0^1, \beta_2^0, {\sigma^2}_{a,}^0 {\sigma^2}_e^0, y)$

Step3: sample β_2^1 from $f(\beta_2|\beta_0^1, \beta_1^1, {\sigma^2}_{a,}^0 {\sigma^2}_e^0, y)$

Step4: sample ${\sigma^2}_{a,}^1$ from $f(\beta_2|\beta_0^1, \beta_1^1, \beta_2^1, {\sigma^2}_e^0, y)$

Step5: sample ${\sigma^2}_{e,}^1$ from $f(\beta_2|\beta_0^1, \beta_1^1, \beta_2^1, {\sigma^2}_a^1, y)$.

Now the first iteration is completed and θ^1 =($\beta_0^1, \beta_1^1, \beta_2^1, {\sigma^2}_{a,}^1 {\sigma^2}_e^1$). Repeat step 1 to step 5 N number of times to obtain a chain $\theta^{(1,2,3,\ldots,N)}$. In the initial iteration, the posterior distribution of one parameter is correlated with that of the others as it obviously depends on previously generated sample value. However, with a large number of iteration, independence will be achieved after a certain number of "Burn-in", which is a special term refers to certain amount of initial iterations that will be discarded as parameters are highly correlated at that stage.

In this study, the purpose is to utilize MCMC iterations to replace integration and resemble the likelihood function in a frequentist framework. Specifically, the flat priors for the random effects (i.e., σ_a^2 and σ_e^2) are uniform distributions from 0 to100 and for the fixed effects flat normal distributions with a mean of zero and a variances of 0.001 or flat uniform distribution (0,1). A calculation based on the method in Raftery and Lewis's (1992) indicates that 10,000 iterations would be sufficient to estimates parameters in this study. Therefore, the total number of iteration is 105,000 with 5,000 as burn-in and thinning of 10 (i.e., instead of storing each chain, every 10th chain will be saved). The purpose of thinning is just to reduce computer storage by recording only every kth value when the chain is run normally. Thinning thus does not relate to improving the quality of estimates. The final MCMC of 10,000 iterations is used for the inference of the posterior distributions. I compared the estimates between the frequentist and Bayesian approaches using the empirical data. The results section illustrates the details.

2.6.3 Results

The quality of the MCMC concerning convergence is evaluated via three types of graphs: autocorrelation plots, density plots and trace plots illustrated in Figures 2.8, 2.9, and 2.10 respectively. The autocorrelation graphs show that for the intercept and class size, independence achieved after some initial iterations while the other parameters achieve independence more quick. The density plots show that all parameters have good normal posterior distributions given the large number of iterations. The trace plots echo what the autocorrelation plots both showing that the intercept and class size converged less well than the other parameters.

Table 4 reports results using the empirical data. The frequentist approach contains four models: unweighted, unscaled, size-weighted and effective-weighted. Comparing these results

with the simulation results in Table 2.2, it is evident that the unscaled estimate of the cluster level variance has positive bias. For the estimates of the first level variance and the intercept, the pattern between the simulation and the empirical analysis is not very consistent, which may indicate that the simulation data with simplified normal residual did not fully capture the real data variation which may not be normal distribution. However, the estimates of the standard errors follow similar patterns in both simulation and empirical analyses. In general, the unweighted standard error is the smallest, while the unscaled counterpart is the largest.

When comparing results between the frequentist and Bayesian approaches, the Bayesian summary statistics of the posterior distributions perfectly match the unweighted estimates in terms of the center point estimates, the variances as well as the 95% confident intervals or high posterior density region (HPD) intervals which has been used more often than credible interval in practice (Gill, 2008). Clearly, the range of the interval for unweighted or Bayesian analysis is narrower than the range in the weighted analysis. That is, the standard errors are smaller as previous simulation results showed.

2.7 Discussion

Large-scale educational data typically adopt complex sampling designs with special features. Clustering and unequal probability of selection are two features that present a challenge in statistical analysis. Clustering data structures violate the statistical assumption of independent errors in regression model. As a result, multilevel modeling approaches have been utilized to incorporate clustering effects in the analysis.

The unequal probability of selection may lead to an informative design at each sampling stage. If the informativeness is ignored at either level when fitting a model with sample data, the

estimates will be biased (Cai, 2013). In order to correct this bias, design informativenes should be incorporated in the model analysis. Specifically, the use of sampling weights has been recommended in model analysis using a design-based approach, but the difficulty of applying the weights in multilevel models has been underestimated. To many data analysts, it is unclear when and how complex sampling weights should be used in multilevel modeling. The evidence and practical guidance in this area is badly needed, especially when researchers are increasingly relying on the large-scale data to produce scientific evidence.

In the literature, two approaches have been proposed to incorporate informative sampling designs from a model-based perspective. The first approach is derived from PML in which the sampling weights are incorporated in the likelihood to produce unbiased and consistent estimates. This approach is used in single-level model. In multilevel model, two estimation techniques have been proposed (i.e., MPML and PWIGLS). These techniques have also been used in structural equation modeling (T. Asparouhov, 2005; Stapleton, 2006). Many statistical software programs employ these two techniques such as STATA, Mplus and LISREL. One limitation of this approach is that it produces asymptotic unbiased estimates, but the exact distribution of the weighted point estimators is generally unknown. Therefore, when the sample size is small, substantial bias could be observed (D. Pfeffermann et al., 2006). Moreover, the standard errors of estimates in weighted analyses are typically larger than those in unweighted analyses, which is a main disadvantage of incorporating weights in analyses (Kish, 1992). Further, it is not clear how to conduct weighted multilevel analysis including scaling when having an informative design for statistical models with more than two levels.

This study aims to shed new light on how to incorporate informative deign in multilevel modeling by working on both approaches. The results indicated that in both the simulation and

the empirical investigation, unscaled estimates produced substantial bias for the variance component and the two scaling methods (size and effective) seems to perform equally well concerning the estimation of cluster level variance. In general, some findings in this study are consistent with previous studies. For example, previous studies showed that for a two-level random intercept linear model, when both levels have an informative design, the unweighted estimation produces biased estimates for all parameters involved (Cai, 2013; Pfeffermann, Skinner, et al., 1998). This conclusion is consistent with findings from this simulation. The different is that I found bias varied across variables while Cai reported that fixed effects are nearly unbiased or slightly biased within 10 percent of the true value. Pfeffermann et al. (1998) found that when an informative design exists in level 2 only, the bias of unweighted σ_e^2 disappears but the unweighted estimators β_0 and σ_a^2 remain biased. I used an informative design at both levels, so the unweighted estimate of σ_e^2 is also biased. This finding is in congruence with previous findings that which sampling stage is informative matters (Cai, 2013). With regard to the SDA, empirical results may suggest that this method fails to take into account the informative selection either because of the small informative value in the data or because adding error term does not change model estimates. Therefore, the tentative conclusion is that sample weighted estimation approaches (i.e., MPML and PWIGLS) would still be preferred.

There are several limitations in this study. First, only a simple multilevel random-intercept linear model is considered in this study. It is possible that results may vary when random slopes are introduced or non-linear models are used. Second, there are some practical procedures that have been applied on the sampling weights such as weights trimming and non-response adjustment, which have not been considered in the simulation.

For future research could further evaluate the performance of MPML in other model settings. For example, it may be informative to extend the research to nonlinear model as discrete responses are often used as outcomes (Goldstein, 1991) and there is an increasing body of research focusing on binary and count data (Chaudhuri, Handcock, & Rendall, 2008; Natarajan, Lipsitz, Fitzmaurice, Moore, & Gonin, 2008; Nordberg, 1989; Rodriguez & Goldman, 1995, 2001). Extending the research to longitudinal designs would also be interesting and needed (Jenkins, 2008; C. J. Skinner & de Toledo Vieira, 2007; Stapleton, Harring, & Lee, 2016; Vieira & Skinner, 2008).

In addition to substantial informative design and scaling procedure, other factors such as sample size and ICC value could also be included in simulation setup for further research. For example, Pfeffermann et al. (1998) found that the sample number of level 1 units instead of the sample number of level 2 is the critical factor affecting the bias of unscaled estimators especially for small cluster sample size. Regarding the effect of ICC, literature showed that when ICC values increase, the bias for the unscaled estimate decreases (Asparouhov, 2006; Kovačević & Rai, 2003). Moreover, currently, the literature development and software applications regarding sampling weights and scaling options mostly deal with two-level models. For example, in Stata 14 there is no availability for scaling procedures in models with more than two-levels. Future research may move forward to three-level models.

APPENDICES

APPENDIX 2.1: STATA SIMULATION CODE

```
/*********************************************************************/
local info 0.67 0.5 1 2
forvalue i = 1/1000 {
display "iteration `i'"
foreach j of local info {
display "informative `j'"
* generate school level data
quietly: set seed 1`i'1
quietly: set obs 6000
quietly: gen uj=rnormal(0,0.25)
quietly: gen pj=1/(1+exp(4-uj/`j'))
quietly: gen wj=1/pj
quietly: gsample 120 [aw=pj],gen(index1)
quietly: gen school = _n
* school covariates
quietly: gen clsize=rnormal(20,4)
quietly: gen private=uniform()<=0.16
quietly: expand 140
quietly: sort school
* generate student data
quietly: gen eij=rnormal(0,0.5)
quietly: gen pi_j=1/(1.35+exp(1-eij/2))
quietly: gen wi_j=1/pi_j
quietly: gen pij=pi_j*pj
quietly: gen wij=1/pij
quietly: gen female=uniform()<=0.49
quietly: gen ses=rnormal(0.06,0.9)
* merge two level data
quietly: gen yij=1+0.019*female-0.065*ses-0.001*clsize-0.026*private+uj+eij
*select final sample
keep if index==1
quietly: gsample 4200 [aw=pi_j]
/*********************************************************************/
```

APPENDIX 2.2 A: EQUATION DETAILS

1) In Equation (B2.2)

$$f_p(y_{ij}|X_{ij}, I_{ij} = 1)$$
$$= \frac{f_p(y_{ij}, X_{ij}, I_{ij} = 1)}{f_p(X_{ij}, I_{ij} = 1)}$$
$$= \frac{f_p(I_{ij} = 1|y_{ij}, X_{ij}) f_p(y_{ij}, X_{ij})}{f_p(X_{ij}, I_{ij} = 1)}$$

$$= \frac{f_p(I_{ij} = 1|y_{ij}, X_{ij}) f_p(y_{ij}|X_{ij}) f_p(X_{ij})}{f_p(I_{ij} = 1|X_{ij}) f_p(X_{ij})}$$

$$= \frac{f_p(I_{ij} = 1|y_{ij}, X_{ij}) f_p(y_{ij}|X_{ij})}{f_p(I_{ij} = 1|X_{ij})}$$
$$= \frac{Prob(I_{ij} = 1|y_{ij}, X_{ij})}{Prob(I_{ij} = 1|X_{ij})} f_p(y_{ij}|X_{ij})$$

2) In Equation (B2.4)

$$Prob(I_{ij} = 1|y_{ij}, X_{ij})$$
$$= \int prob(I_{ij} = 1|y_{ij}, X_{ij}, \pi_{ij}) f_p(\pi_{ij}|y_{ij}, X_{ij}) d\pi_{ij}$$
$$= \int \pi_{ij} f_p(\pi_{ij}|y_{ij}, X_{ij}) d\pi_{ij}$$
$$= E_p(\pi_{ij}|y_{ij}, X_{ij})$$

3) In Equation (B2.5)

$$Prob(I_{ij} = 1|X_{ij})$$
$$= \int E_p(\pi_{ij}|y_{ij}, X_{ij}) f_p(y_{ij}|X_{ij}) dy_{ij}$$
$$= E_p(\pi_{ij}|X_{ij})$$

APPENDIX 2.2 B: BAYESIAN CODE

```
######################### R package: R2WinBugs  ################################
###### model syntax
   model;{
   # likelihood
   for(i in 1:N){
   y[i]~dnorm(mu[i],tau.e)
   mu[i]<-
grand.mean[id.school[i]]+beta.female*female[i]+beta.ses*ses[i]+beta.cs*cs[i]+beta.private*private[i]-
0.0397*var.e+0.0281*var.int
   }
   # prior
   for (j in 1:n.sch){
   grand.mean[j]~dnorm(mu.int,tau.int)
   }
   mu.int~dnorm(0,0.001)
   beta.female~dunif(0,1)
   beta.ses~dnorm(0,0.001)
   beta.cs~dnorm(0,0.001)
   beta.private~dunif(0,1)

   sigma.e~dunif(0,100)
   var.e<-sigma.e*sigma.e
   tau.e<-1/var.e

   sigma.int~dunif(0,100)
   var.int<-sigma.int*sigma.int
   tau.int<-1/var.int
   }

#############################################################################
```

FIGURES

Figure 2.1: A diagram for the conceptual framework

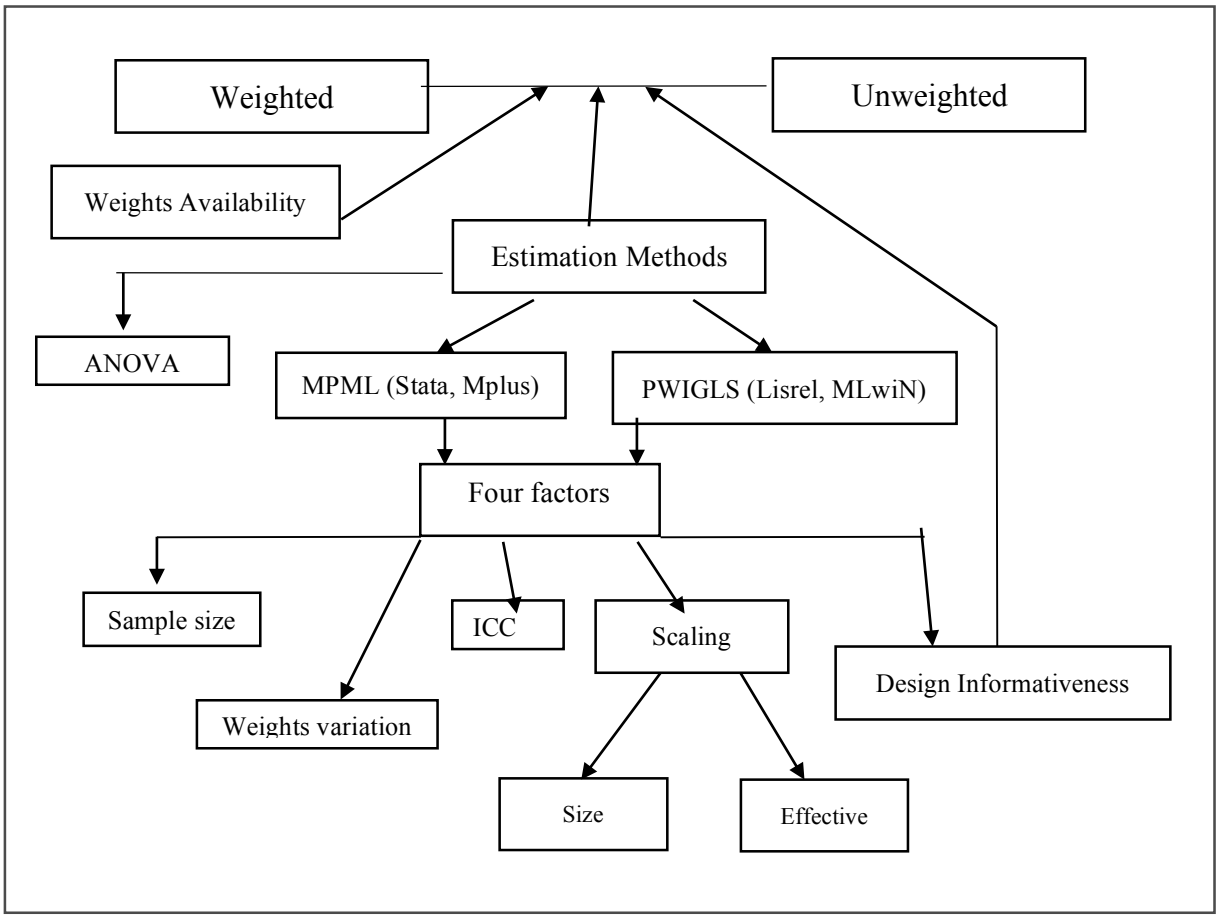


Figure 2.2: Relative bias for four covariates

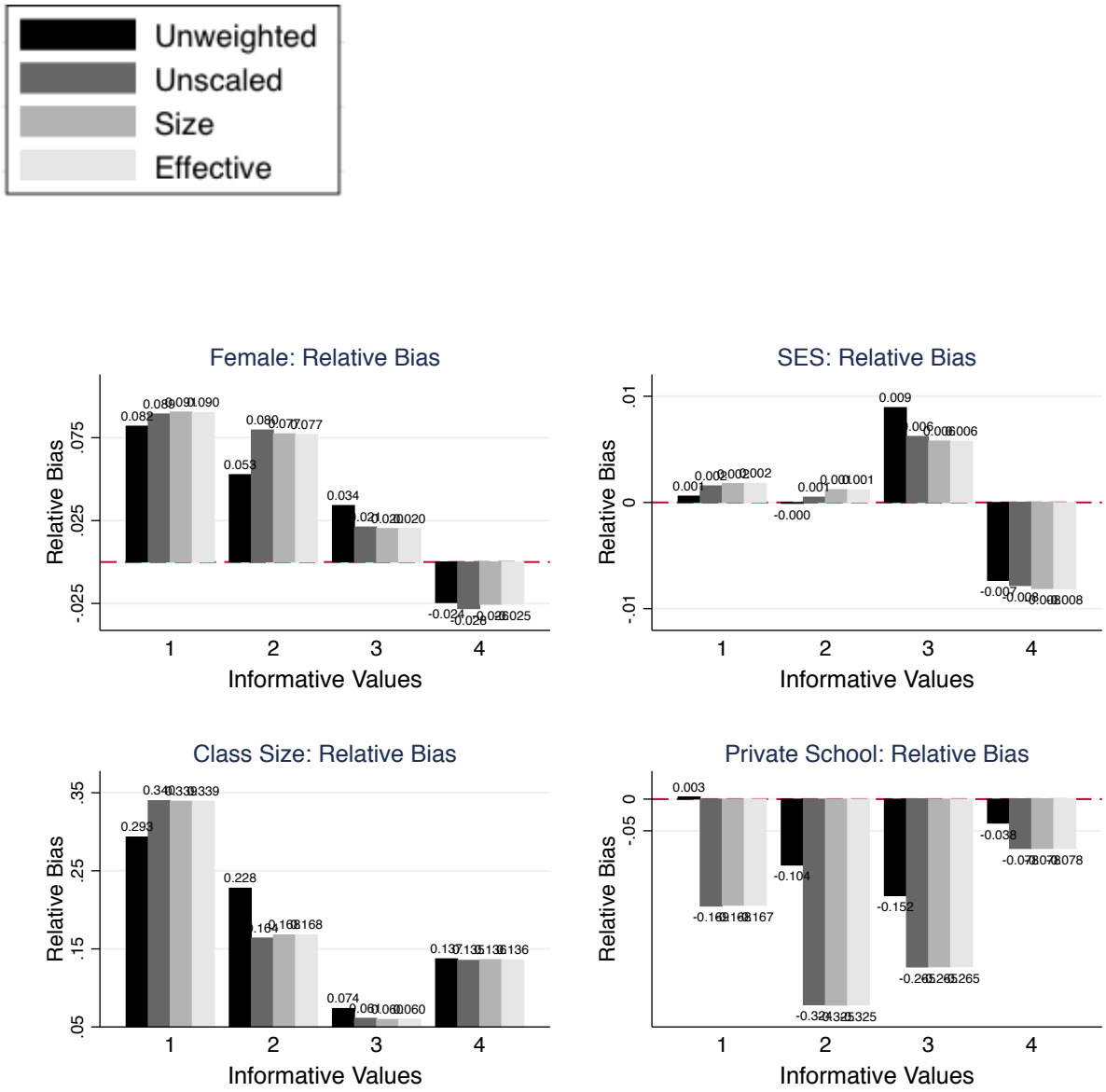


Figure 2.3: RMSE for four covariates

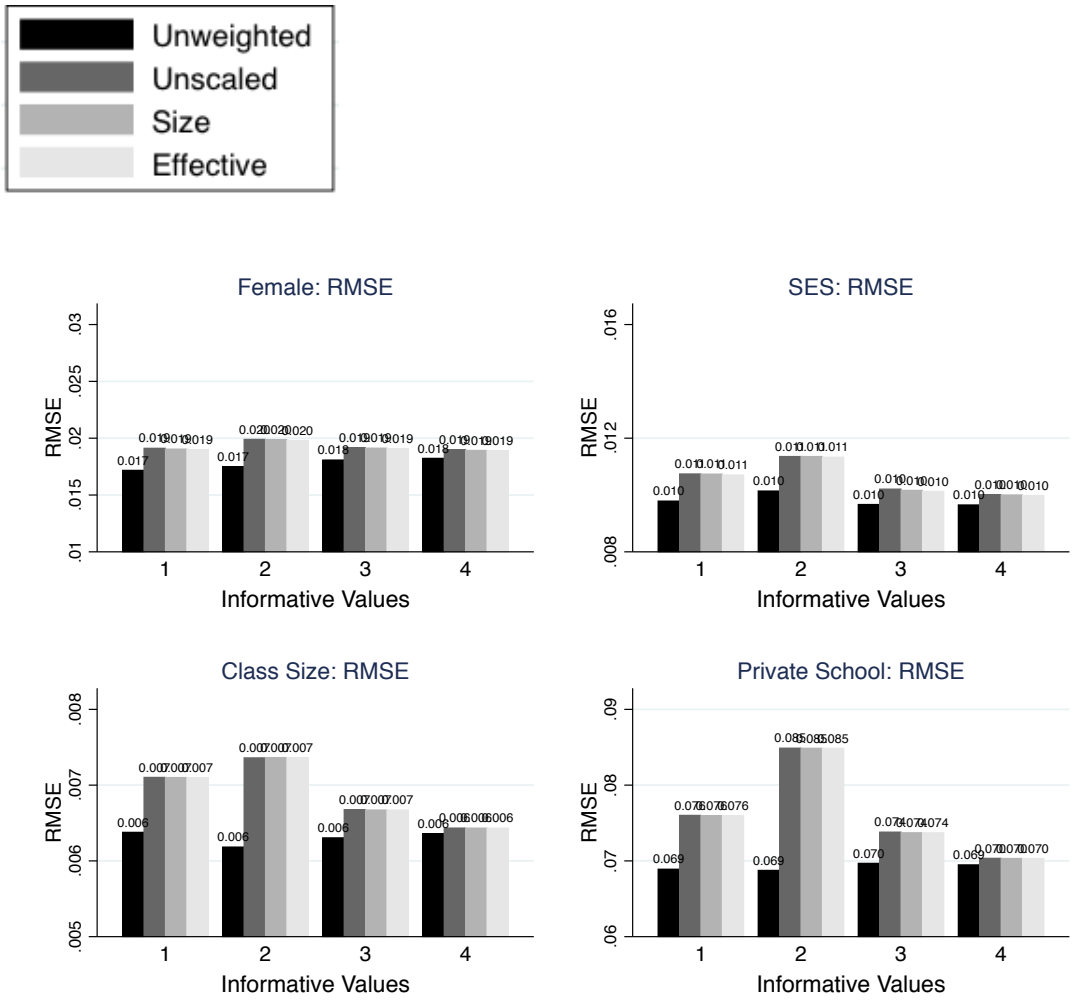


Figure 2.4: 95% coverage rate for four covariates

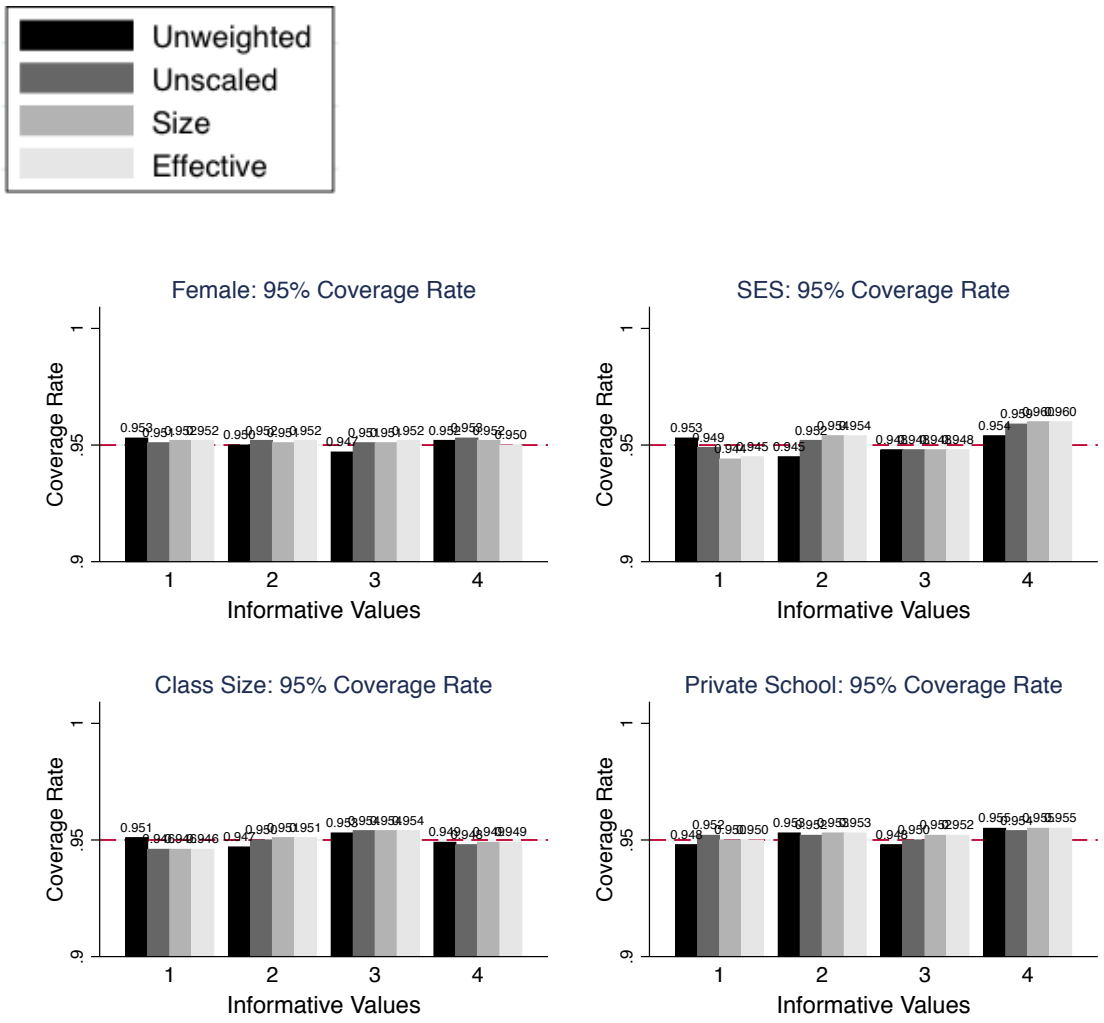


Figure 2.5: Relative bias for the intercept and variance component

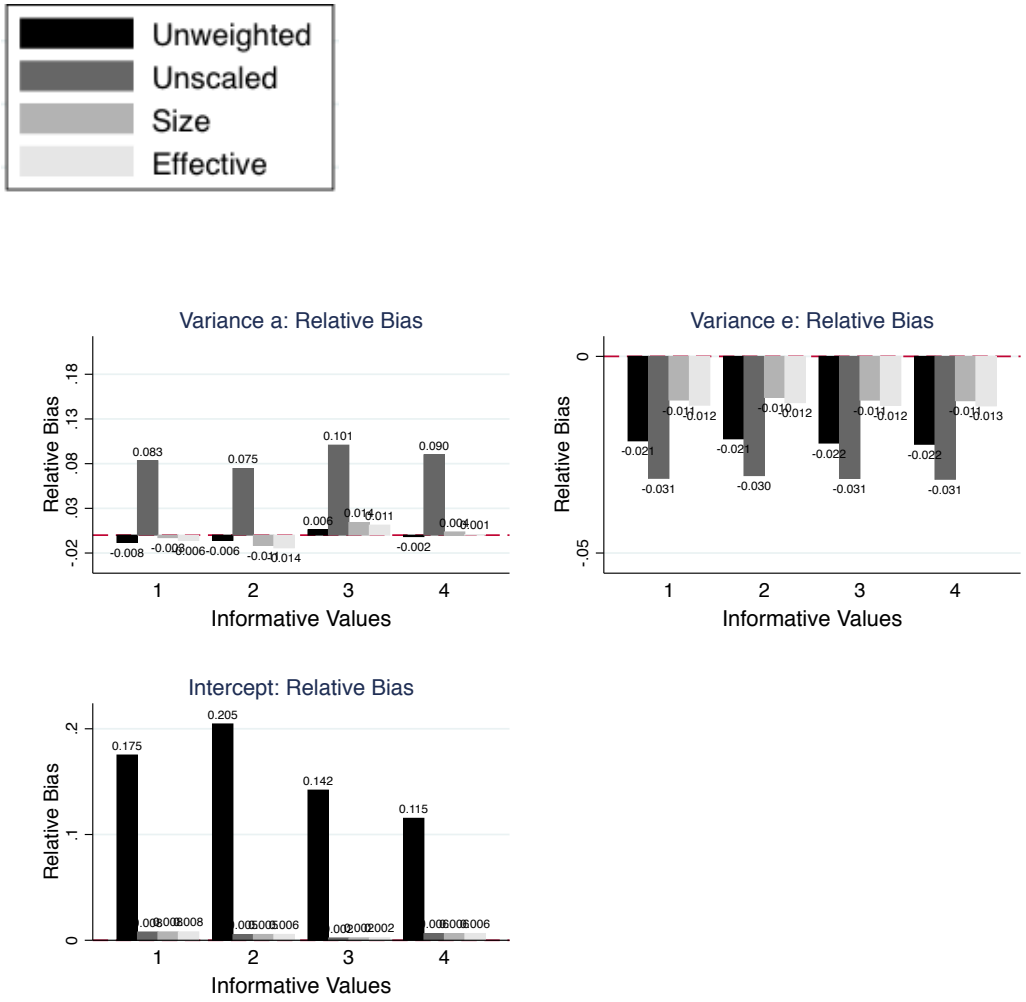


Figure 2.6: RMSE for the intercept and variance component

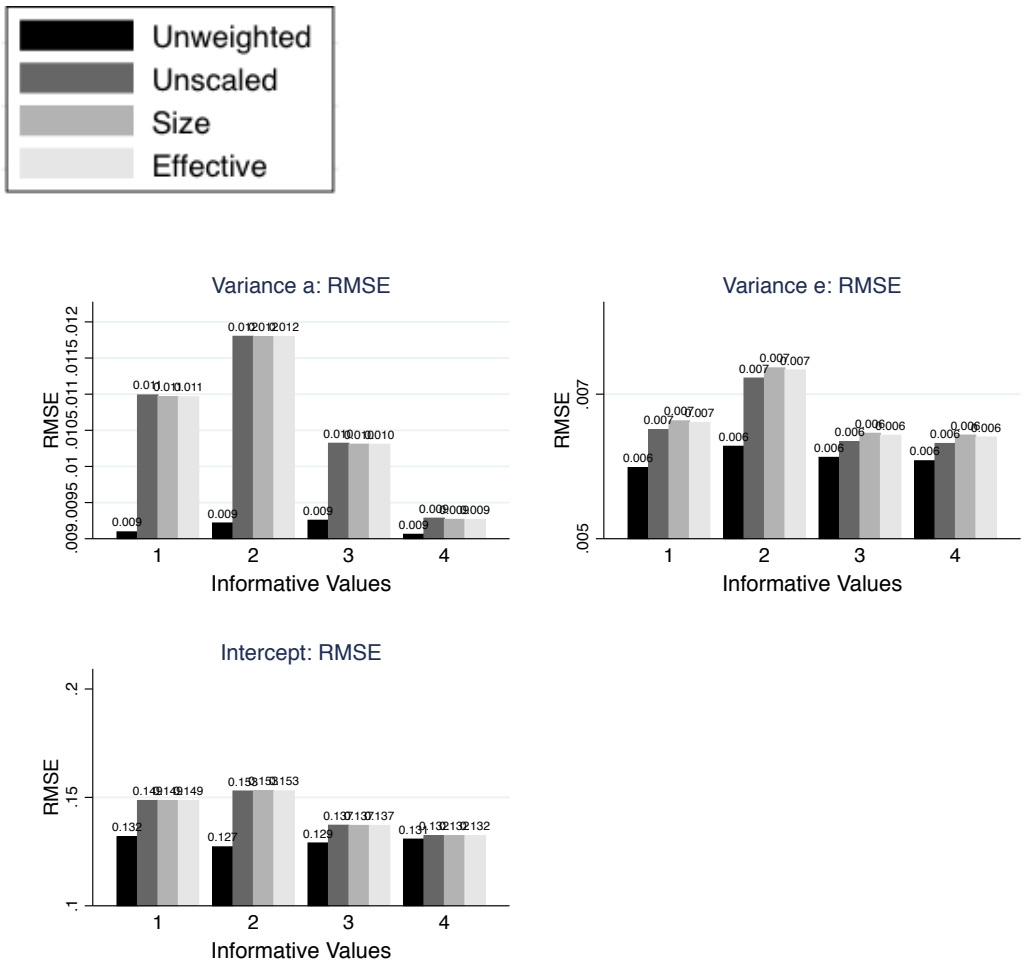


Figure 2.7: 95 % Coverage rate for the intercept and variance component

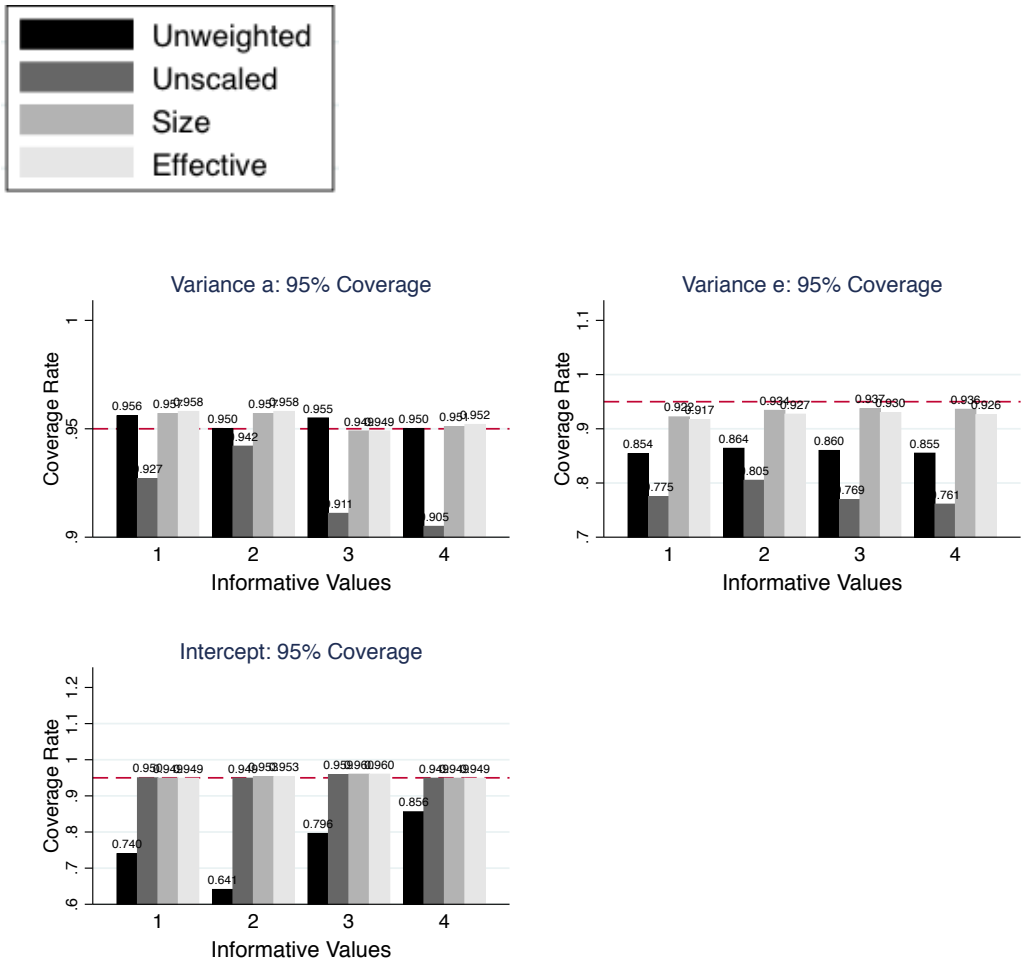


Figure 2.8: Autocorrelation plots

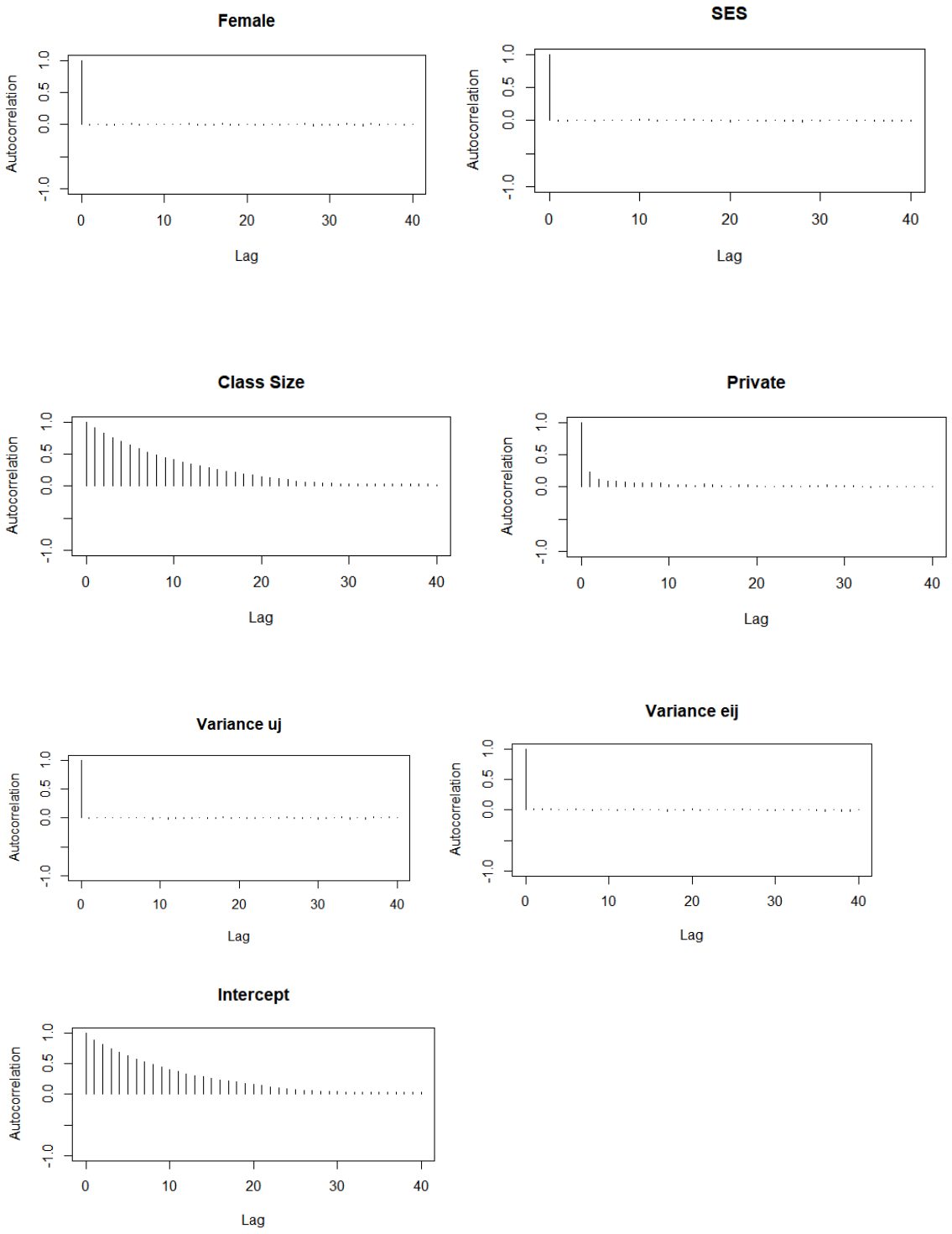


Figure 2.9: Density plots

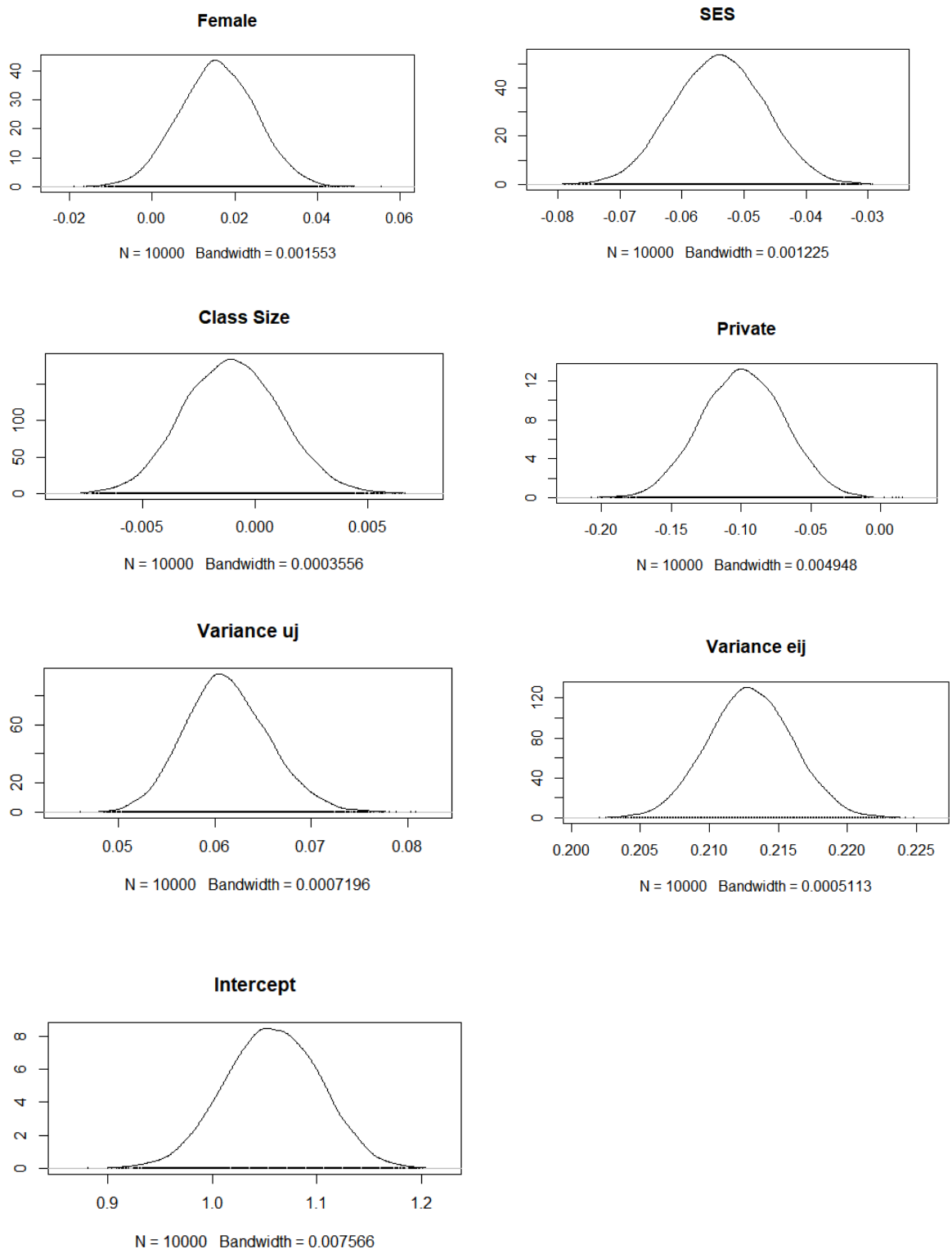
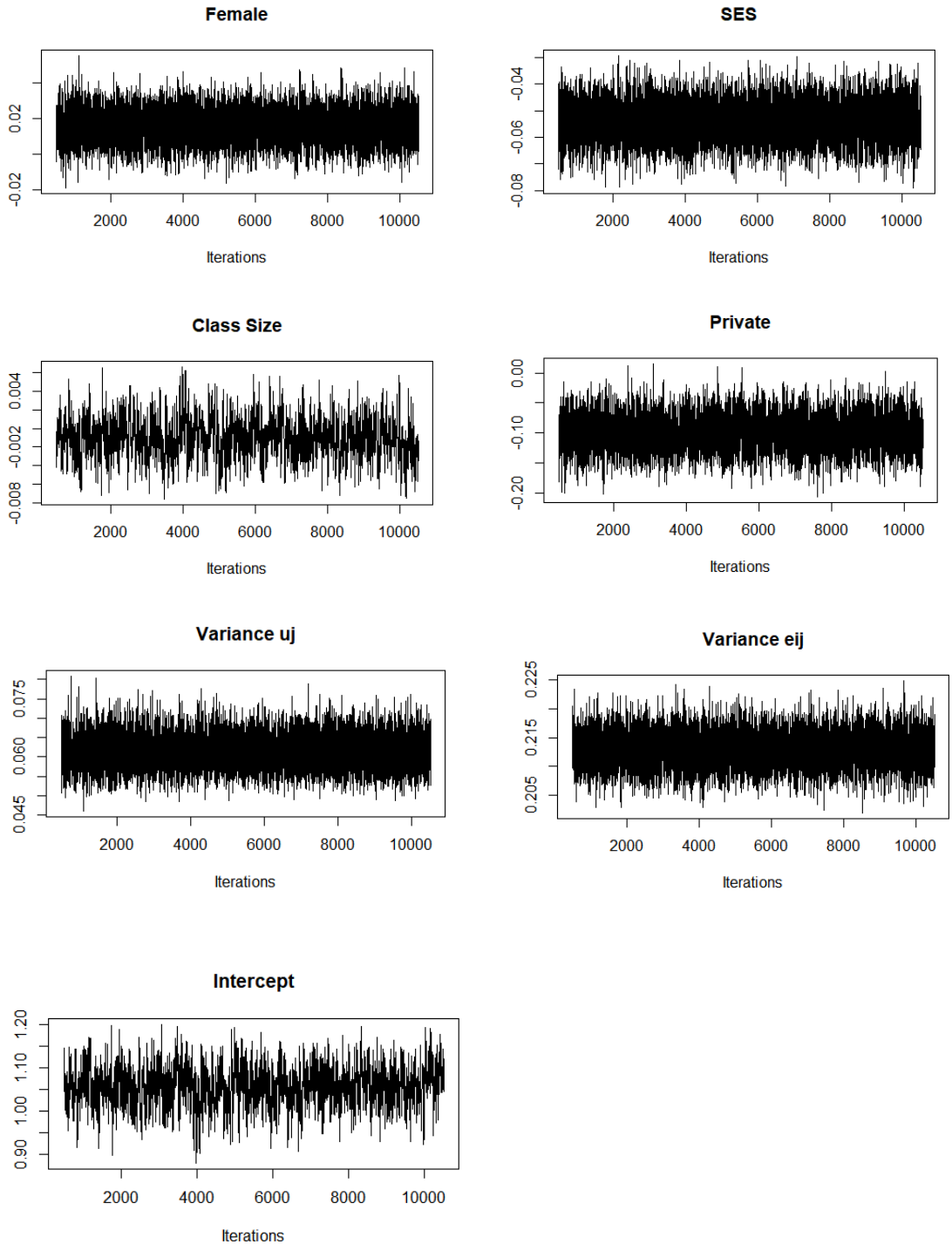


Figure 2.10: Trace plots



TABLES

Table 2.1: Results for covariates

	Info=1 (high)				Info=2				Info=3				Info=4 (low)			
Covariates	Mean	Rbias	RMSE	95%CR	Mean	Rbias	RMSE	95%CR	Mean	Rbias	RMSE	95%CR	Mean	Rbias	RMSE	95%CR
Female	**0.019**															
NW	0.021	0.082	0.017	0.953	0.020	0.053	0.017	0.950	0.020	0.034	0.018	0.947	0.019	-0.024	0.018	0.952
NS	0.021	0.089	0.019	0.951	0.021	0.080	0.020	0.952	0.019	0.021	0.019	0.951	0.018	-0.028	0.019	0.953
SIZE	0.021	0.091	0.019	0.952	0.020	0.077	0.020	0.951	0.019	0.020	0.019	0.951	0.019	-0.026	0.019	0.952
EFF	0.021	0.090	0.019	0.952	0.020	0.077	0.020	0.952	0.019	0.020	0.019	0.952	0.019	-0.025	0.019	0.950
SES	**-0.065**															
NW	-0.065	0.001	0.010	0.953	-0.065	0.000	0.010	0.945	-0.066	0.009	0.010	0.948	-0.065	-0.007	0.010	0.954
NS	-0.065	0.002	0.011	0.949	-0.065	0.001	0.011	0.952	-0.065	0.006	0.010	0.948	-0.064	-0.008	0.010	0.959
SIZE	-0.065	0.002	0.011	0.944	-0.065	0.001	0.011	0.954	-0.065	0.006	0.010	0.948	-0.064	-0.008	0.010	0.960
EFF	-0.065	0.002	0.011	0.945	-0.065	0.001	0.011	0.954	-0.065	0.006	0.010	0.948	-0.064	-0.008	0.010	0.960
ClassSize	**-0.001**															
NW	-0.001	0.293	0.006	0.951	-0.001	0.228	0.006	0.947	-0.001	0.074	0.006	0.953	-0.001	0.137	0.006	0.949
NS	-0.001	0.340	0.007	0.946	-0.001	0.164	0.007	0.950	-0.001	0.061	0.007	0.954	-0.001	0.135	0.006	0.948
SIZE	-0.001	0.339	0.007	0.946	-0.001	0.168	0.007	0.951	-0.001	0.060	0.007	0.954	-0.001	0.136	0.006	0.949
EFF	-0.001	0.339	0.007	0.946	-0.001	0.168	0.007	0.951	-0.001	0.060	0.007	0.954	-0.001	0.136	0.006	0.949
Private	**-0.026**															
NW	-0.026	0.003	0.069	0.948	-0.023	-0.104	0.069	0.953	-0.022	-0.152	0.070	0.948	-0.025	-0.038	0.069	0.955
NS	-0.022	-0.169	0.076	0.952	-0.018	-0.324	0.085	0.952	-0.019	-0.265	0.074	0.950	-0.024	-0.078	0.070	0.954
SIZE	-0.022	-0.168	0.076	0.950	-0.018	-0.325	0.085	0.953	-0.019	-0.265	0.074	0.952	-0.024	-0.078	0.070	0.955
EFF	-0.022	-0.167	0.076	0.950	-0.018	-0.325	0.085	0.953	-0.019	-0.265	0.074	0.952	-0.024	-0.078	0.070	0.955

Note: Rbias=relative bias; RMSE=root mean square error; CR=coverage rate; NW=unweighted; NS=unscaled; SIZE=size scaling; EFF=effective scaling

Table 2.2: Results for intercept and random effects

	Info=1 (high)				Info=2				Info=3				Info=4 (low)			
Covariates	Mean	Rbias	RMSE	95%CR	Mean	Rbias	RMSE	95%CR	Mean	Rbias	RMSE	95%CR	Mean	Rbias	RMSE	95%CR
Intercept	**1**															
NW	1.175	0.175	0.132	0.740	1.205	0.205	0.127	0.641	1.142	0.142	0.129	0.796	1.115	0.115	0.131	0.856
NS	1.008	0.008	0.149	0.950	1.005	0.005	0.153	0.949	1.002	0.002	0.137	0.959	1.006	0.006	0.132	0.949
SIZE	1.008	0.008	0.149	0.949	1.005	0.005	0.153	0.953	1.002	0.002	0.137	0.960	1.006	0.006	0.132	0.949
EFF	1.008	0.008	0.149	0.949	1.006	0.006	0.153	0.953	1.002	0.002	0.137	0.960	1.006	0.006	0.132	0.949
Variance_a	**0.0625**															
NW	0.0620	-0.0082	0.0091	0.9560	0.0621	-0.0058	0.0092	0.9500	0.0629	0.0060	0.0093	0.9550	0.0624	-0.0016	0.0091	0.9500
NS	0.0677	0.0832	0.0110	0.9270	0.0672	0.0746	0.0118	0.9420	0.0688	0.1007	0.0103	0.9110	0.0681	0.0899	0.0093	0.9050
SIZE	0.0623	-0.0028	0.0110	0.9570	0.0618	-0.0113	0.0118	0.9570	0.0634	0.0141	0.0103	0.9490	0.0627	0.0036	0.0093	0.9510
EFF	0.0621	-0.0057	0.0110	0.9580	0.0616	-0.0142	0.0118	0.9580	0.0632	0.0112	0.0103	0.9490	0.0625	0.0007	0.0093	0.9520
Variance_e	**0.250**															
NW	0.245	-0.021	0.006	0.854	0.245	-0.021	0.006	0.864	0.244	-0.022	0.006	0.860	0.244	-0.022	0.006	0.855
NS	0.242	-0.031	0.007	0.775	0.242	-0.030	0.007	0.805	0.242	-0.031	0.006	0.769	0.242	-0.031	0.006	0.761
SIZE	0.247	-0.011	0.007	0.922	0.247	-0.010	0.007	0.934	0.247	-0.011	0.006	0.937	0.247	-0.011	0.006	0.936
EFF	0.247	-0.012	0.007	0.917	0.247	-0.012	0.007	0.927	0.247	-0.012	0.006	0.930	0.247	-0.013	0.006	0.926

Note: Rbias=relative bias; RMSE=root mean square error; CR=coverage rate; NW=unweighted; NS=unscaled; SIZE=size scaling; EFF=effective scaling

Table 2.3: Simulation standard deviations of point estimators

Informativeness->	I=1 (High)		I=2		I=3		I=4 (Low)	
	SD	SE	SD	SE	SD	SE	SD	SE
Female								
NW	0.0172	0.0155	0.0175	0.0155	0.0181	0.0155	0.0182	0.0155
NS	0.0191	0.0189	0.0199	0.0201	0.0192	0.0185	0.0190	0.0181
SIZE	0.0191	0.0189	0.0199	0.0201	0.0191	0.0184	0.0189	0.0180
EFF	0.0190	0.0189	0.0198	0.0200	0.0191	0.0184	0.0189	0.0180
SES								
NW	0.0098	0.0086	0.0101	0.0086	0.0097	0.0086	0.0097	0.0086
NS	0.0108	0.0105	0.0114	0.0111	0.0102	0.0102	0.0100	0.0100
SIZE	0.0107	0.0105	0.0114	0.0111	0.0102	0.0102	0.0100	0.0100
EFF	0.0107	0.0105	0.0113	0.0110	0.0101	0.0102	0.0100	0.0100
Class size								
NW	0.0064	0.0061	0.0062	0.0061	0.0063	0.0062	0.0064	0.0061
NS	0.0071	0.0066	0.0074	0.0070	0.0067	0.0063	0.0064	0.0062
SIZE	0.0071	0.0066	0.0074	0.0070	0.0067	0.0063	0.0064	0.0062
EFF	0.0071	0.0066	0.0074	0.0070	0.0067	0.0063	0.0064	0.0062
Private								
NW	0.0689	0.0673	0.0688	0.0675	0.0697	0.0681	0.0695	0.0677
NS	0.0761	0.0709	0.0849	0.0747	0.0738	0.0691	0.0704	0.0667
SIZE	0.0760	0.0709	0.0849	0.0747	0.0738	0.0691	0.0704	0.0667
EFF	0.0760	0.0709	0.0849	0.0747	0.0738	0.0691	0.0704	0.0667
Intercept								
NW	0.1319	0.1257	0.1273	0.1259	0.1290	0.1262	0.1308	0.1260
NS	0.1487	0.1360	0.1530	0.1441	0.1372	0.1303	0.1324	0.1264
SIZE	0.1486	0.1360	0.1531	0.1441	0.1371	0.1302	0.1324	0.1264
EFF	0.1486	0.1360	0.1531	0.1441	0.1371	0.1302	0.1324	0.1264
Variance_a								
NW	0.0091	0.0090	0.0092	0.0090	0.0093	0.0091	0.0091	0.0091
NS	0.0110	0.0098	0.0118	0.0100	0.0103	0.0096	0.0093	0.0091
SIZE	0.0110	0.0098	0.0118	0.0100	0.0103	0.0096	0.0093	0.0091
EFF	0.0110	0.0098	0.0118	0.0100	0.0103	0.0096	0.0093	0.0090
Variance_e								
NW	0.0060	0.0054	0.0063	0.0054	0.0061	0.0054	0.0061	0.0054
NS	0.0065	0.0066	0.0072	0.0069	0.0063	0.0064	0.0063	0.0062
SIZE	0.0066	0.0067	0.0074	0.0071	0.0065	0.0065	0.0064	0.0064
EFF	0.0066	0.0067	0.0073	0.0070	0.0064	0.0065	0.0064	0.0063

Note: NW=unweighted; NS=unscaled; SIZE=size scaling; EFF=effective scaling

Table 2.4: Comparing Frequentist and Bayesian analysis using empirical data

		Frequentist				Bayesian		
Parameters		Unweighted	Unscaled	Size_Scale	Effective_Scale	Mode	Mean	
Intercept	Estimate	1.048 *	1.006 *	1.063 *	1.064 *	1.050	1.050	*
	Standard error	(0.043)	(0.060)	(0.049)	(0.049)	(0.045)		
	95% Interval	[0.964, 1.132]	[0.888, 1.125]	[0.967, 1.159]	[0.968, 1.160]	[0.970, 1.136]		
Female	Estimate	0.016	0.013	0.017	0.017	0.015	0.016	
	Standard error	(0.009)	(0.010)	(0.010)	(0.010)	(0.009)		
	95% Interval	[-0.002, 0.034]	[-0.008, 0.033]	[-0.002, 0.037]	[-0.002, 0.037]	[-0.003, 0.033]		
SES	Estimate	-0.054 *	-0.043 *	-0.051 *	-0.051 *	-0.054	-0.054	*
	Standard error	(0.007)	(0.010)	(0.009)	(0.009)	(0.007)		
	95% Interval	[-0.068,-0.040]	[-0.063, -0.023]	[-0.069, -0.034]	[-0.069, -0.033]	[-0.068, -0.039]		
Class size	Estimate	-0.001	0.001	-0.002	-0.002	-0.001	-0.001	
	Standard error	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)		
	95% Interval	[-0.005, 0.003]	[-0.005, 0.007]	[-0.006, 0.003]	[-0.006, 0.003]	[-0.005, 0.003]		
Private	Estimate	-0.098 *	-0.110 *	-0.120 *	-0.120 *	-0.100	-0.097	*
	Standard error	(0.029)	(0.035)	(0.032)	(0.032)	(0.029)		
	95% Interval	[-0.156,-0.041]	[-0.179, -0.041]	[-0.183, -0.057]	[-0.184, -0.057]	[-0.156, -0.041]		
Variance a	Estimate	0.061 *	0.074 *	0.060 *	0.060 *	0.060	0.061	*
	Standard error	(0.004)	(0.006)	(0.005)	(0.005)	(0.004)		
	95% Interval	[0.053, 0.069]	[0.062, 0.087]	[0.050, 0.071]	[0.050, 0.071]	[0.053, 0.070]		
Variance e	Estimate	0.213 *	0.203 *	0.197 *	0.197 *	0.213	0.213	*
	Standard error	(0.003)	(0.006)	(0.005)	(0.005)	(0.003)		
	95% Interval	[0.207, 0.219]	[0.191, 0.215]	[0.189, 0.206]	[0.187, 0.206]	[0.207, 0.219]		

Note: * $p<0.05$, frequentist is 95% confidence interval while Bayesian is 95% HPD interval.

REFERENCES

REFERENCES

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*(3), 411-434.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods, 35*(3), 439-460.

Asparouhov, T., & Muthen, B. (2006). *Multilevel modeling of complex survey data.* Paper presented at the Proceedings of the joint statistical meeting in seattle.

Asparouhov, T., & Muthen, B. (2007). *Testing for informative weights and weights trimming in multivariate modeling with survey data.* Paper presented at the Proceedings of the 2007 JSM meeting, Section on Survey Research Methods, Salt Lake City, Utah.

Bertolet, M. (2008). *To weight or not to weight? Incorporating sampling designs into model-based analyses.* (Ph.D.), Carnegie Mellon University, Ann Arbor.

Binder, D., & Roberts, G. (2006). *Approaches for analyzing survey data: a discussion.* Paper presented at the Survey Research Methods Section, American Statistical Association.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*(3), 279-292.

Binder, D. A., Kovacevic, M. S., & Roberts, G. (2005). *How important is the informativeness of the sample design.* Paper presented at the Proceedings of the Survey Methods Section.

Binder, D. A., & Roberts, G. (2009). Design-and model-based inference for model parameters. *Handbook of Statistics, 29*, 33-54.

Binder, D. A., & Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of survey data* (pp. 29-48): John Wiley & Sons, Ltd.

Cai, T. (2013). Investigation of Ways to Handle Sampling Weights for Multilevel Model Analyses. *Sociological Methodology, 43*(1), 178-219.

Chantala, K., & Suchindran, C. (2006). Adjusting for unequal selection probability in multilevel models: A comparison of software packages. *Proceedings of the American Statistical Association, Seattle, WA: American Statistical Association*, 2815-2824.

Chaudhuri, S., Handcock, M. S., & Rendall, M. S. (2008). Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(2), 311-328.

Chaudhuri, S., Handcock, M. S., & Rendall, M. S. (2010). *A conditional empirical likelihood approach to combine sampling design and population level information*. Retrieved from

Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica, 9*(2), 385-406.

Cohen, S. B., Burt, V. L., & Jones, G. K. (1986). Efficiencies in variance estimation for complex survey data. *The American Statistician, 40*(2), 157-164.

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American statistical Association, 91*(434), 883-904.

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American statistical Association, 78*(383), 535-543.

Eideh, A., & Nathan, G. (2009). Two-stage informative cluster sampling with application in small area estimation. *Journal of statistical planning and inference, 139*, 3088-3101.

Fang, F., Hong, Q., & Shao, J. (2009). A pseudo empirical likelihood approach for stratified samples with nonresponse. *The Annals of Statistics, 37*(1), 371-393.

Firth, D., & Bennett, K. E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society. Series B, Statistical methodology, 60*(1), 3-21.

Foy, P. (2014). Estimating Standard Errors for the TIMSS and PIRLS 2011 Achievement Scales. In Methods and Procedures in TIMSS and PIRLS 2011.

Francisco, C. A., & Fuller, W. A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics, 19*(1), 454-469.

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology, 10*(1), 99-118.

Gelman, A., Carlin, J. B., H. S. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. New York, 3rd edition: CPC Press: Taylor & Francis Group.

Gill, J. (2008). *Bayesian methods: a social and behavioral sciences approach.* Hoboken, NJ, 2rd edition: Chapman and Hall/CRC.

Godambe, V. P., & Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review/Revue Internationale de Statistique, 54*(2), 127-138.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*(1), 43-56.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika, 78*(1), 45-51.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Theory. *Econometrica, 52*(3), 681-700.

Graubard, B. I., & Korn, E. L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical methods in medical research, 5*(3), 263-281.

Graubard, B. I., & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science, 17*(1), 73-96.

Grilli, L., & Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology, 30*(1), 93-103.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American statistical Association, 78*(384), 776-793.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Holt, D., Smith, T. M. F., & Winter, P. D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society. Series A (General), 143*(4), 474-487.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association, 47*(260), 663-685.

Jenkins, F. (2008). *Multilevel analysis with informative weights.* Paper presented at the Proc. the Joint Statistical Meeting, ASA section on Survey Research Methods.

Jewell, N. P. (1985). Least Squares Regression with Data Arising from Stratified Samples of the Dependent Variable. *Biometrika, 72*(1), 11-21.

Jia, Y., Stokes, L., Harris, I., & Wang, Y. (2011). Performance of random effects model estimators under complex sampling designs. *Journal of educational and behavioral statistics, 36*(1), 6-32.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics, 8*(2), 183-200.

Konijn, H. S. (1962). Regression analysis in sample surveys. *Journal of the American statistical Association, 57*(299), 590-606.

Korn, E. L., & Graubard, B. I. (2003). Estimating variance components by using survey data.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1), 175-190.

Kovačević, M. S., & Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communications in Statistics-Theory and Methods, 32*(1), 103-121.

Koziol, N. A. (2016). *A comparison of population-averaged and cluster-specific approaches in the context of unequal probabilities of selection.* (AAI3689723 Ph.D.), The University of Nebraska - Lincoln.

Krieger, A. M., & Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology, 18*(2), 225-239.

LaVange, L. M., Steams, S. C., Lafata, J. E., Koch, G. G., & Shah, B. V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical methods in medical research, 5*(3), 311-329.

Lee, J., & Fish, R. M. (2010). International and interstate gaps in value-added math achievement: multilevel instrumental variable analysis of age effect and grade effect. *American Journal of Education, 117*(1), 109-137.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*(3), 265-284.

Lin, Y. X., Steel, D., & Chambers, R. L. (2004). Restricted quasi-score estimating functions for sample survey data. *Journal of Applied Probability, 41*, 119-130.

Lubienski, S. T., & Lubienski, C. (2006). School sector and academic achievement: A multilevel analysis of NAEP mathematics data. *American Educational Research Journal, 43*(4), 651-698.

Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society, 60*(1), 115-126.

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

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear and mixed models*. New York: Wiley.

Natarajan, S., Lipsitz, S. R., Fitzmaurice, G., Moore, C. G., & Gonin, R. (2008). Variance estimation in complex survey sampling for generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 57*(1), 75-87.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics, 5*(3), 223.

OECD. (2012). PISA 2012 technical report. *Paris: Organisation for Economic Co-operation and Development*. OECD. (2014). Survey weighting and the calculation of sampling variance. In PISA 2012 technical report.

Palardy, G. J. (2010). The multilevel crossed random effects growth model for estimating teacher and school effects: Issues and extensions. *Educational and Psychological Measurement, 70*(3), 401-419.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique, 61*(2), 317-337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical methods in medical research, 5*(3), 239-261.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it. *Survey Methodology, 37*(2), 115-136.

Pfeffermann, D., Fernando Antonio Da Silva, M., & Pedro Luis Do Nascimento, S. (2006). Multi-level modelling under informative sampling. *Biometrika, 93*(4), 943.

Pfeffermann, D., & Holmes, D. J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society. Series A (General), 148*(4), 268-278.

Pfeffermann, D., Krieger, A. M., & Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica, 8*(4), 1087-1114.

Pfeffermann, D., & LaVange, L. M. (1989). Regression models for stratified multi-stage cluster samples. In C. J. Skinner, D. Holt, & T. F. M. Smith (Eds.), *Analysis of Complex Surveys* (pp. 237-260). Chichester: Wiley.

Pfeffermann, D., Moura, F. A. D. S., & Silva, P. L. D. N. (2006). Multi-level modelling under informative sampling. *Biometrika, 93*(4), 943-959.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B, Statistical methodology, 60*(1), 23-40.

Pfeffermann, D., & Smith, T. M. F. (1985). Regression models for grouped populations in cross-section surveys, correspondent paper. *International Statistical Review, 53*(1), 37-59.

Pfeffermann, D., & Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series*

B, 61(1), 166-186.

Pfeffermann, D., & Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American statistical Association, 102*(480), 1427-1439.

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American statistical Association, 87*(418), 383-396.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 805-827.

Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs Sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 763-773). Oxford, U.K.: Oxford University Press.

Rao, J. N. K. (1997). Developments in sample survey theory: an appraisal. *Canadian Journal of Statistics, 25*(1), 1-21.

Rao, J. N. K., & Bellhouse, D. R. (1990). History and development of the theoretical foundations of survey based estimation and analysis. *Survey Methodology, 16*(1), 3-29.

Rao, J. N. K., Chaudhuri, A., Eltinge, J. L., Fay, R. E., Ghosh, J. K., Ghosh, M., . . . Pfeffermann, D. (1999). Some current trends in sample survey theory and methods (with discussion). *Sankhyā: The Indian Journal of Statistics, Series B, 61*(1), 1-57.

Rao, J. N. K., & Wu, C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 72*(4), 533-544.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models (2nd ed.)*. Tousand Oaks, CA: SAGE.

Rodgers-Farmer, A. Y., & Davis, D. (2001). Analyzing complex survey data. *Social Work Research, 25*(3), 185-192.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 73-89.

Rodriguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 164*(2), 339-355.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1*(4), 381-397.

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. V. Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment: background, technical issues, and methods of data analysis* (pp. 117-153). London: Chapman Hall/CRC Press.

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical methods in medical research, 5*(3), 283-310.

Särndal, C.-E., Thomsen, L. B., Hoem, J. M., Lindley, D. V., Barndorff-Nielsen, O., & Dalenius, T. (1978). Design-based and model-based inference in survey sampling [with discussion and reply]. *Scandinavian Journal of Statistics, 5*(1), 27-52.

Scott, A., & Smith, T. M. F. (1969). Estimation in Multi-Stage Surveys. *Journal of the American statistical Association, 64*(327), 830-840.

Scott, A. J., & Holt, D. (1982). The Effect of Two-Stage Sampling on Ordinary Least Squares Methods. *Journal of the American statistical Association, 77*(380), 848-854.

Skinner, C. (1994). *Sample models and weights.* Paper presented at the Proceedings of the Section on Survey Research Methods.

Skinner, C. J. (1989). Domain means, regression and multi-variate analysis. In: Skinner, C. J., Holt, D., & Smith, T. M. F., (eds). *Analysis of Complex Surveys*, 59-87.

Skinner, C. J., & de Toledo Vieira, M. (2007). Variance estimation in the analysis of clustered longitudinal survey data. *Survey Methodology, 33*(1), 3-12.

Smith, T. M. F. (1988). To weight or not to weight, that is the question. *Bayesian statistics, 3*, 437-451.

Snijder, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling, 2nd edition.* London: Sage Publication Ltd.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9*(4), 475-502.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling, 13*(1), 28-58.

Stapleton, L. M. (2013). Incorporating sampling weights into single-and multi-level models. In L. Rutkowski, M. v. Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment.* (pp. 353-388). London: Chapman Hall/CRC Press.

Stapleton, L. M., Harring, J. R., & Lee, D. (2016). Sampling weight considerations for multilevel modeling of panel data. In J. R. Harring, Stapleton, L. M. & Beretvas, S. N. (Ed.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*. Charlotte, NC: Information Age Publishing, Inc.

Stapleton, L. M., & Kang, Y. (2016). Design effects of multilevel estimates from national probability samples. *Sociological methods & research*, 1-28.

Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71*(3), 495-506.

Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources, 33*(1), 127-169.

Vieira, M. D. T., & Skinner, C. J. (2008). Estimating models for panel survey data under complex sampling. *Journal of Official Statistics, 24*(3), 343-364.

Wedel, M., ter Hofstede, F., & Steenkamp, J. (1998). Mixture model analysis of complex samples. *Journal of Classification, 15*(2), 225-244.

West, B. T., & Galecki, A. T. (2011). An overview of current software procedures for fitting linear mixed models. *The American Statistician, 65*(4), 274-282.

Wu, Y. Y. (2007). Estimation of regression coefficients with unequal probability samples.

CHAPTER 3
THE LONG-TERM AND CAUSAL EVIDENCE OF CLASS SIZE EFFECTS FROM ECLS-K

3.1 Introduction

Does class size matter? Does smaller class size make a difference to improve the quality of public education in the U.S.? From a historical perspective, choosing an appropriate class size for a grade-layered education system (i.e., primary, secondary and tertiary education) has been a practice perhaps as early as the establishment of public education. Increasing or decreasing one or two student in a classroom may not matter much, but doing it in a ten-unit would make a great difference pertaining to classroom composition and the way of organizing students (e.g., grouping or tracking). Specifically, it may change pupil-teacher communication, peer interaction, teaching staff pattern and instruction practice (Graue, Hatch, Rao, & Oen, 2007; Mosteller, Light, & Sachs, 1996). As classrooms of different size have various goals and structures, which will have an impact on student non-cognitive skills such as motivation and ultimately on academic learning and performance (Ames, 1992). Given many factors that class size is related to, it is not surprising that there has been considerable conversations and debates among parents, teachers, educational researchers and policymakers over the years about how important class size is in the U.S. education system (Douglass & Parkhurst, 1940; Ehrenberg, Brewer, Gamoran, & Willms, 2001; Finn, 2002; Hanushek, 1999, 2002; Hanushek, Mayer, & Peterson, 1999; Mishel & Rothstein, 2002).

On one hand, parents and teachers are raring to support any forms of class size reduction based on the common-sense knowledge that smaller class size would ameliorate teaching and learning environments. For example, previous studies showed that reducing class size helps

reduce teacher's work load and disciplinary control, allows teachers to devote more time to learning activities, facilitates individualized instruction and increases attention and support that each student would receive from his/her teacher (Blatchford, Bassett, & Brown, 2011; Blatchford, Russell, Bassett, Brown, & Martin, 2007; Konstantopoulos & Sun, 2014; Rice, 1999). On the other hand, although class size is a variable that can be manipulated easily without triggering significant changes in other components in education (e.g., curriculum), class size reduction involves hiring a great number of additional teachers, so it is an expensive initiative to implement from the financial perspective. Thus, opponents of small class-size policy argued that class size reduction is over-invested and it is a false promise (M. M. Chingos, 2011).

When pondering over the pros and cons of reducing class size, policymakers face a dilemma in decision making. The evidence on the effectiveness of class size reduction has been increasingly relied upon to make a wise decision. However, findings of class size effects have been mixed and inconclusive by and large (Akerhielm, 1995; M. M Chingos, 2013; Hattie, 2005; Milesi & Gamoran, 2006; Rice, 1999). There are at least several factors that attribute to this inconsistency as follows: (1) geographical locations of sample data (i.e., country, region, district and school); (2) research design (i.e., correlation, quasi-experiment and experiment); (3) grade level (i.e., pre-kindergarten, elementary, secondary, and postsecondary school); (4) subject area (e.g., reading, math, science, history and art); (5) student and school characteristics.

With regard to the geographical location, it is not uncommon that education system varies widely worldwide. The aspect of class size is no exception. For instance, the average class size in Europe and U.S. in elementary school is around 20 to 30 while in some east Asian countries (e.g., China and Japan) it would be around 40 to 50. Therefore, class size effects in east and west education are not comparable to some degree. Moreover, within a group of countries that have

similar average class size such as European countries, the class size effects are not consistent. Using large-scale data sets, some studies showed that there is no systematic pattern of class size effects, instead significant findings only observed for a particular country and in a particular year (Li & Konstantopoulos, 2016; Shen & Konstantopoulos, 2017; Wößmann, 2005; Wößmann & West, 2006). Therefore, it is very likely that class-size effect would be country and context specific.

In terms of the research design, it is well-acknowledged that findings from correlation studies including numerous production function studies may have substantial bias due to the issue of non-random placement. For instance, affluent parents are more likely to send their high-achieving children to schools that have resources to operate small sized classrooms. In this scenario, correlation of class size and students' achievement across schools would be spuriously negative showing smaller classroom is associated with better achievement. On the contrary, school administrators may place low-achieving students in a small class as a remedial treatment leading to an erroneous finding of positive correlation between class size and academic scores. Evidently, without carefully isolating confounding factors, non-causal study of class size will produce more mixed findings. Therefore, scientific evidence of class size requires experimental or quasi-experimental designs which the latter could be achieved via applying appropriate statistical methods to observational data.

Regarding grade level, past research has shown that the magnitude of class size effects would decrease as grade goes up. In particular, the most prominent evidence was observed in early grades (i.e., K-3). Moreover, class size effects vary among different subjects (e.g., reading, math and science). Finally, it is inconclusive whether class size would bring more benefits to particular types of schools and groups of students such as low achieving students and schools or

students with special education. Given all these aforementioned factors contributing to the inconsistent conclusion of class size effects, the literature review in the next section focuses exclusively on the experimental and quasi-experimental plus longitudinal studies of class size in early grades in the U.S., which the current research will build upon and contribute to.

3.2 Literature review

In the U.S., at an early stage, small-scale experimental and quasi-experimental designs have been adopted to investigate the effects of class size on student achievement. Rockoff (2009) summarized twenty-four small-scale field experiments that were conducted between 1920 and 1940 involving active assignment of students and teachers to different sized classrooms, among which two studies demonstrated an evidence of increased achievement in smaller classes (Rockoff, 2009). Slavin (1989) did a "best-evidence synthesis" on ten relatively high quality studies of the effects of small class in elementary grades (i.e., K-6) from 1968 to 1987 which used either random assignment or matching. It was found that substantial reductions in class size does have a positive effect on student achievement and the median effect size is 0.13 standard deviations across eight studies (Slavin, 1989). In addition, through a meta-analytic review approach, these early experiment studies revealed that small class size effects have been observed and the effects were greater in class with 10 to 20 students versus 30 to 40 students (G. V. Glass, Cahen, Smith, & Filby, 1982; Gene V Glass & Smith, 1979). Furthermore, there was evidence showing that minority, economically disadvantaged students and low-achieving students get more benefits for being placed in smaller classes in early grades (Gene V Glass & Smith, 1979; Slavin, 1989).

The Student-Teacher Achievement Ratio experiment (STAR) is the sole large-scale randomized experiment which has been design and executed well, which measured class size effects to a significant scale in early grades in education (Mosteller, 1995). The Project STAR was carried out in the state of Tennessee from 1985 to 1989 in grades of K-3. Starting from the entering kindergarten in 1985, about 11,571 students and their teachers were randomly assigned to three class conditions in each of the 79 schools: (1) small class with an average of 15 students; (2) regular class with an average of 22 students; (3) regular class with a teacher aid. Thus far, the project STAR has produced the most pronounced evidence of class size effects in early grades.

Taking advantage of this high quality data with high internal validity, research has revealed that students placed in smaller classes outperform those in larger classes on both cognitive achievement scores and non-cognitive outcomes (e.g., college attendance rate and earnings) and the effects are even long-lasting (Chetty et al., 2011; Finn, Gerber, & Boyd-Zaharias, 2005; Konstantopoulos & Chung, 2009; Krueger & Whitmore, 2001a; Nye, Hedges, & Konstantopoulos, 1999; B. Nye, L. V. Hedges, & S. Konstantopoulos, 2000; Nye, Hedges, & Konstantopoulos, 2001). In addition, attending small class size has larger and longer effects for disadvantaged students such as minority students (e.g., the black) or students receiving free lunch at schools (Finn et al., 2005; Krueger, 1999; Krueger & Whitmore, 2001b; Nye, Hedges, & Konstantopoulos, 2002, 2004). However, other research found that there is no evidence suggesting that small-class effects is larger for low SES students (B. A. Nye, L. V. Hedges, & S. Konstantopoulos, 2000), or for low achieving students (Konstantopoulos, 2008; Nye et al., 2002). Overall, the heterogeneous class size effects on particular subgroups in the Project STAR is inconclusive.

Because of the influential findings from Project STAR, many states (e.g., California, Florida, Wisconsin, Indiana, Minnesota) subsequently invested tremendously to implement class size reduction especially in early grades (Finn & Achilles, 1999). For example, in California, class size was reduced from 30 to 20 in Kindergarten to 3rd grade in 1990s, but the smaller class size benefits have been offset by recruiting a large body of inexperienced and uncertified teachers and by using multi-grade classes (Jepsen & Rivkin, 2009; David Sims, 2008; D. Sims, 2009; Stecher, Bohrnstedt, Kirst, McRobbie, & Williams, 2001). In Wisconsin, the Student Achievement Guarantee in Education (SAGE) program, a pilot project, was carried out in 1996-97 which involved reducing the teacher-pupil ratio to 15 in K-3 and results showed the effect size is about 0.2 standard deviations in first grade, which was consistent with STAR project (Molnar et al., 1999). In Florida, the class size cap is reduced in every core subject in every grade, but the results are not desirable (M. M. Chingos, 2012). In general, the state class size reform was not successful and further, larger benefits of small class for disadvantaged students were not found neither in California nor in Florida (M. M Chingos, 2013).

With regard to the quasi-experimental studies, the evidence has been mixed. Taking advantage of random variation of class size in the school-age population, Hoxby (2002) failed to detect class size effects in fourth and sixth grade in Connecticut schools even for schools with large proportions of disadvantaged or minority students. Utilizing Hoxby's method, however, it was found that reducing class size increases 3rd and 5th graders' reading and mathematics test scores in Minnesota (Cho, Glewwe, & Whitler, 2012). Applying student fixed effects and school-by-year fixed effects models on a panel data of about 200,000 students in over 3,000 public elementary grades plus 7th grade in Texas in 1990s, Rivkin, Hanushek, and Kain (2005) found positive effects of smaller classes on reading and mathematics in 4th and 5th grade, but

little or no effects in later grades (Rivkin, Hanushek, & Kain, 2005). Using differences-in-differences method on a panel data set of students in 127 elementary schools from grade 1 to grade 5 in the San Diego Unified School District, Babcock and Betts (2009) found there is exogenous class size effects and furthermore they found the small classes influence disadvantaged students via eliciting effort or engagement rather than teaching specific skills (Babcock & Betts, 2009).

The literature review above reveals that overall the evidences of class size effects and the heterogeneity on different students and schools are inconclusive across the available small number of high quality studies in the past. Specifically, there are at least three literature voids that need to be addressed.

First, class size studies that used data of specific state, district or school are informative, but relevant findings are not likely to be applied to other states, districts or schools due to the variation in conditions and contexts. For instance, Connecticut has relatively high-paid teachers who may be able to teach well regardless of class size condition. In this case, class size may be less important to this state than to other states (e.g., Florida). In order to provide a whole picture of class size effects at the national level, nationally representative samples would be preferred for having the advantage of external validity (i.e., generalizability). Although high-quality data collected by National Center for Education Statistics (NCES) are available to the public, they have not been fully utilized to present useful information of class size effects to researchers, practitioners and policymakers.

Second, most of the evidence of class size effects heavily focused on academic achievement outcomes. The effects of class size on non-cognitive skills have been overlooked in the literature. Considering the increasing recognition of the substantial impact that these non-

cognitive skills may play in individual's long-term academic attainment and life success, scientific evidence on non-cognitive domain is badly needed. (Dee & West, 2011) is one exceptional study which examined the return of class size reduction on non-cognitive outcomes using the 8th grade data of the National Educational Longitudinal Study of 1998 (NELS: 1988). However, evidence of class size effects on non-cognitive skills in early grades using national representative sample is still a missing piece.

Third, empirical evidence of high-quality quasi-experimental studies is still scarce. There is a lack of longitudinal class size study. Additional, the heterogeneous class size effects need to be further examined as it is still inconclusive whether small class effects would vary across different gender, ethnic groups, and SES backgrounds.

3.3 The present study

Given that the class size reduction has been prevalently appealed in the public while the effectiveness has been questioned and debated continuously yet still unclear, the object of this study is to shed more light on the evidence of class size. It will present the causal and long-term effects of class size on both cognitive and non-cognitive outcomes via applying appropriate statistical methods on the most recent national representative data of the Early Childhood Longitudinal Study Class of 2011 (ECLS-K 2011).

This research has several special features that will address the aforementioned literature gaps. First, the prominent evidence of class size from the data of project STAR is obsolete. As the current society is developing by leaps and bounds and the student population nowadays is also likely to be different, there is a natural call for an update of class size evidence using recent high quality data. The ECLS-K 2011 provides the latest large-scale data concerning children's

early education and development in U.S. To my knowledge, this data that have high external validity has not been utilized to generate long-term or causal class size effects in U.S. so far.

Second, appropriate statistical method is a key to generate less biased estimates of class size, which however has not been taken good advantage of. In the literature, pertaining to quasi-experimental methods in statistics, it is not uncommon that instrumental variable (IV) and regression discontinuity design (RDD) have been used (Akerhielm, 1995; Angrist & Lavy, 1999; Bonesrønning, 2003; Hoxby, 2000), other methods are of less usage. For example, one single study in the past evaluated propensity score matching on the data of Project STAR (Wilde & Hollister, 2007). This research will explore the application of the propensity scores (PS) methods in Kindergarten grade to facilitates causal inference based on a hypothesized binary treatment of small vs non-small sized class. In addition, the individual fixed effects (FE) models will be employed on longitudinal data from Kindergarten to 2nd grade to investigate how the change of class size would be associated with the change of outcome controlling some individual variation. The corresponding estimates of class size thus would be unaffected by time-invariant individual variables. Moreover, to deal with potential missing variable bias, I adopt Frank et al (2013)'s approach to quantify the percentage of bias that must be present to invalidate the significant findings.

Third, from the content perspective, evidence of class size effects on non-cognitive outcomes needs to be considered in decision making of class size policy. It is well-recognized that smaller classroom has been hypothesized to enhance early graders' non-cognitive skills such as learning behavior, motivation, interpersonal interaction and communication which consequently would help to increase their academic achievement. However, the examination of

class size effects on cognitive outcome is dominant in the literature while its effects on non-cognitive outcome is alarmingly rare. This study aims to fill in this literature gap.

Fourth, heterogeneity of class size has also been a focal debate. This study will conduct subgroup analysis in terms of student gender, ethnicity, parents' education, family economic status and school sector. For sensitivity analysis, full sample analysis will be conducted both with and without sampling weights to account for any potential design effects, but for subgroup analysis, only unweighted analysis would be appropriate.

3.4 Research methods

3.4.1 Data, sample and measures

The data of ECLS-K 2011 was collected by the National Center for Education Statistics (NCES). It is the most recent longitudinal study that follows a national probability sample of kindergarten students of diverse socioeconomic and ethnic backgrounds from kindergarten through early elementary grades in the U.S. ECLS-K provides information regarding children's early school experience. Data have been collected to study how students' cognitive, social and emotional development might be related to various family, classroom and school environments that students have been exposed to. Currently, kindergarten-second grade data file is available for public use. Although data were collected twice a year (i.e., fall and spring), in the fall of the first and second grades, only a third of schools were surveyed and the class size variable is not available for these two rounds. Therefore, four waves of data (i.e., kindergarten in fall and spring and spring in first and second grades) will be utilized in the fixed effects data analyses while for propensity score methods, data in kindergarten year will be used.

In terms of sampling design, the ECLS-K data has adopted a three-stage stratified sampling strategy in which 90 geographic regions served as the primary sampling units (PSUs) and then public and private schools with 5-year-old children were selected within sampled PSUs and finally students were selected within sampled schools (Tourangeau et al., 2015). Both the first and second sampling stages select sample with probabilities that are proportional to measures of population size. The base weights of school are the PSU weights multiplied by the weights of selecting school from the PSU and adjustments were made for public and private schools (Mulligan, Hastedt, & McCarroll, 2012). The base weights of student take into account the within-school student weight with non-response adjustment as well as nonresponse-adjusted school weights. In addition, student weights for Asian/Pacific Islander (API) students were calculated separately from non-API students due to oversampling.

The set of cognitive outcomes contains the direct assessment of each child's reading and math achievement in the format of Item Response Theory (IRT) scale scores, which can be compared with other children regardless of which specific items a child takes (Tourangeau et al., 2015). The scores tend to be normally distributed with a metric ranging from -6 to 6. IRT has several advantages comparing to raw score. For instance, it allows longitudinal measurement of gain in achievement, adjusts guessing probability for low-ability child, and keeps track of a consistent pattern of right and wrong answers regardless of omitted items which raw scoring would treat as have been answered incorrectly (Tourangeau et al., 2015).

In the literature, the non-cognitive outcome has been used as a catch-all term to refer to everything that is not captured by intelligence assessment or standardized achievement tests (West et al., 2016). The non-cognitive skill is regarded to be more malleable than IQ. It encompasses a broad range of competence, which consists of five general categories based on a

comprehensive framework developed by Farrington et al (2012) (Farrington et al., 2012). They include (1) academic behavior (i.e., going to class, doing homework, organizing materials, participating and studying); (2) academic perseverance (i.e., grit, tenacity, delayed gratification, self-discipline and self-control); (3) academic mindsets (e.g., beliefs and attitudes); (4) learning strategies (i.e., study skills, metacognitive strategies, self-regulated learning, time management and goal-setting); (5) social skills (i.e., interpersonal skills, empathy, cooperation, assertion, and responsibility).

In ECLS-K data, there are two sets of ratings on children's problem behaviors and social skills collected from teacher and parent. The teacher's rating will be used as it may be more professional than that of parents based on their education background in general. Their rating should also be more objective assuming equal evaluation for each student. In the fixed effect analysis, the set of non-cognitive outcomes that are available in all four rounds of data are included: (1) *self-control* (4 items) -- control temper, respect others' property, accept peers' ideas and handle peer pressure; (2) *interpersonal skills* (5 items) -- get along with others, forms and maintains friendships, help other children, show sensitivity to others' feelings and express feels, ideas and opinions in positive ways; (3) *externalizing problem behaviors* (5 items) -- argue, fights, get angry, act impulsively and disturb ongoing activities; (4) *internalizing problem behaviors* (4 items) -- exhibit anxiety, loneliness, low self-esteem and sadness; (5) *approaches to learning* (7 items) -- keep belongings organized; show eagerness to learn new things; work independently; easily adapt to changes in routine; persist in completing tasks; pays attention well; and follows classroom rules.

In each variable, higher value indicates that it is more likely to observe the skill or behavior. The score for each of these five non-cognitive measures in the data is the mean rating

on the items included in the scale so they are in a continuous scale roughly. The individual item details are not available in the user's manual which the information was obtained in a reference (Gottfried & Le, 2016). Comparing with the theoretical framework which traits seem to be more related to older children's learning, the non-cognitive outcomes in the ECLS-K data may emphasize more on social and emotional behaviors which are more evident given the characteristics of this younger age.

In propensity score analysis, only cognitive outcomes are used to generate causal inference of class size effects considering possibly substantial measurement error might exist in measuring these non-cognitive skills. For example, parents may tend to rate higher score to their own child and teachers to more attractive students. Their judgements have the issue of fake desirability, so the results could be misleading. To examine the consistency of non-cognitive rating, I calculated the correlation between teachers' and parents' rating on individual's self-control, interpersonal social skills and approaches to learning. The values turn out to be very small, which are 0.02, 0.04 and 0.06 respectively indicating teacher's and parent's rating on these three non-cognitive skills are not very correlated.

Besides the outcome variables, the main independent variable of interest is the teacher reported class size. Variable details are listed in Appendix Table.

3.4.2 Fixed effects models

The statistical method involves individual fixed-effect method, which is an econometric approach that has been used in analyzing the impact of time-varying variables (Wooldridge, 2016). Here the purpose is to control for all the time-invariant student characteristics both observed and non-observed, so that the class size estimates would be unbiased by these time-invariant factors such as family background which turns out to be highly correlated with student

learning outcomes. It should be noted that the successful application of this method relies on having enough variation in variables over time, which I think is valid for the class-size variable and outcomes as they are all continuous variables with some reasonable changes from year to year.

Although the individual fixed effect models purify the estimation of class size by getting rid of the impact of some important time-invariant variables, there still could be omitted variable bias. To deal with it, this study will quantify the percentage bias necessary to invalidate the inference for statistically significant results. Specifically, based on the method proposed by Frank et al (2013) (Frank, Maroulis, Duong, & Kelcey, 2013), the bias can be calculated using the formula: bias %= 1-(standard error X $t_{critical,df}$/coefficient) (where estimates of coefficient and standard error are generated by software while $t_{critical,df}$ would be close to 1.96 for a two-tailed t-test with 0.05 significance level when sample size is large as in this study). This bias quantification approach is superior to the statement about higher or lower statistical significance. According to Frank et al (2013), the median level of robustness is about 30% by rule of thumb for observational studies in education.

Based on data availability, the analysis will focus on four rounds of data (i.e., fall kindergarten, spring kindergarten, spring 1st grade and spring 2nd grade). For longitudinal data, attrition over time would be an issue. For example, it is not uncommon some students may switch to another school across different time points. If the new school is within the same school district, then these observations would still be available in the sample, otherwise, these observations become missing data. To make it simple, this study focuses on the observations (total N= 18,174) that are available in the first round of data cycle while ignoring these later added observations (delete 9.4% - 1,710). Further, only those individuals who remain in the same

school all the time were considered (delete 32.5% - 5,912), otherwise it would not be able to control the time varying class and school effects. Finally, only the first time kindergarteners were included, that is to further exclude kindergarten repeaters (delete 4.9% - 884). The final sample size is 9,668.

Specifically, the statistical model of fixed effect is as follows:

$$Y_{ijt} = \beta_0 + \beta_1 CS_{ijt} + v_i + LOC_j + Wave_t + e_{ijt} \quad (3.1)$$

where Y_{ijt} is the outcome for student *i* in school *j* at time *t*, β_0 is the intercept, *CS* represents the variable of interest class size and β_1 is its coefficient, v_i indicates individual fixed effect, LOC_j represents school location (i.e., city, stubborn, town and rural), $Wave_t$ indicates data cycles with t-1 values and the first wave as a reference, e_{ijt} is the error term. Cluster robust standard errors will be used to control for school clustering effect and residual heterogeneity.

This model analysis will be conducted for the full sample. To check the results sensitivity, subgroups analysis will also be performed by gender: female (49.27%, 4,763) vs male (50.54%, 4,886); by ethnicity: White (52.28%, 5,054), Black (10.15%, 981), Hispanic (24.19%, 2,339) and Asian (7.77%, 751) while other minority race or race combination consisting of small number of observations (5.49%, 531) are omitted; and by school sector: public (89.07%, 8,611) vs private (10.93%, 1,057). To consider the potential influence of informative sampling design, the full sample analyses with sampling weights will be conducted and used as the main statistical inference. The longitudinal sampling weights variable is W6C6P_6T0, which is the child base weights adjusted for nonresponse associated with child assessment, parent and teacher data for fall and spring kindergarten and spring for first and second grades.

3.4.3 Propensity score methods

Propensity score (PS) methods have been widely used in social sciences and education to facilitate causal influence on observational data. They contain a group of strategies that utilize PS to reduce selection bias or pre-determined differences on observed variables between/among treatment groups. The propensity scores are the predicted probabilities that each observation will be assigned to the treatment condition given a vector of observed covariates (X). The PS can be computed via logit or probit model. Possible variables may include: all measured baseline covariates, covariates that affect treatment assignment, covariates that affect outcome, and covariates that affect both treatment and outcome. Consensus has not been reached concerning which variables should be included in the PS model (Austin, 2011). Once PS has been applied, balance between treated and untreated groups could be achieved based on these covariates under the assumption that there are no unmeasured confounders. Under this condition, the observational data would then resemble a randomized experiment in which there is no systematic difference between treated and untreated groups not only on these observed variables, but also on other observed and unobserved variables.

In general, there are two frameworks to estimate the treatment effect. Rubin's the potential outcome framework (i.e., Rubin's causal model) defines the average treatment effect (ATE) as the difference between expected value of the outcomes for all the observations in the treatment group and that in the control group (Rubin, 1974). That is, $E[Y_i^t]$- $E[Y_i^c]$ where subscript *i* stands for each individual, *t* stands for treatment and c is for control. The other framework is the average treatment of the treated (ATT), $E[Y_i^t|T=1]$- $E[Y_i^c|T=1]$,which is the difference between the expected value of the observed outcome for the treated individuals and that of the potential outcome for those treated individuals. Both frameworks have a

requirement of assumptions on the strong ignorability of treatment assignment, adequate common support (i.e., the overlap of PS distribution) and stable unit treatment value assumption (SUTVA). In general, ATT has less strict assumption requirement than ATE (Leite et al., 2015).

There are four commonly used PS methods including matching, stratification and inverse probability of treatment weighting (IPTW) and PS as covariate adjustment (Austin, 2011; Guo & Fraser, 2015). PS matching has some subcategories: one-to-one pair matching or many-to-one matching, greedy or optimal matching, with or without replacement matching. The key element to distinguish among different PS methods is how coarse the weight is (Leite et al., 2015). For example, in the binary treatment case, one-to-one greedy would produce the coarsest weights in which (1) treated and (2) untreated but matched individuals receive weights of one and (3) untreated and unmatched individuals receive weights of zero. Once the particular matching procedure is done, a comparison between treated and untreated subjects can be made within the PS matched sample. The stratification method divides treated and untreated individuals into k strata and then there are kx2 different weights to be defined (assuming binary treatment), with which the treated and control groups will have similar distribution of X within each stratum after deleting unmatched observations. Thus, the weighted average of stratum-specific mean differences in the observed outcome across all k strata will be an unbiased estimate of the average causal effect (Rosenbaum, 1991; Rosenbaum & Rubin, 1984). The IPTW would generate as many weights as the number of observations and no deletion would be necessary. It will compare treated and untreated observations weighted by the inverse probability of treatment.

In the class size context, due to the sorting issue, children attending small-sized classrooms are likely to be different from their counterparts who are placed in relatively large-sized classrooms on many factors (e.g., family background and achievement level). Examining

the true impact of small class size would be challenging given these potential pre-existing differences between those individuals in different treatment groups. In this analysis part, PS method will be used to reduce the bias. The treatment of small class size is a binary variable in which the value would be one if the actual class size is at or below 20 students and zero otherwise. The rationale of using 20 as a cut-off point is based on the class size literature and convention in the early grades in the U.S. From the literature perspective, for instance, after review work of Glass and Smith (1979), Educational Research Service indicated that achievement benefits from small class do not become noticeable unless the class size is fewer than 20 students (Educational Research Service, 1980). In practice, for example, in California's class size reduction reform in k-3 in 1996, the class size cap is 20 students.

This study will mainly utilize the IPTW approach to yield the ATT estimate considering the sampling weights issue as well as comparing the relative strength and weakness of the other three approaches. First, previous studies suggested that the method of including PS directly as a model covariate is not highly recommended because it requires assumptions about relationship between the covariate and the outcome within each treatment group (Hong, 2010b, 2012b). Second, the PS matching is an enormous popular method, but there is a debate about whether this method fulfills the goal of preprocessing satisfactory data for causal inference. Some researchers regard PS matching as a good approaches (Austin, 2011; Rosenbaum & Rubin, 1985). However, King & Nielson (2016) argued that although there is nothing wrong with matching, PS based matching should not be used as it increases imbalance and bias (King & Nielsen, 2016).

Third, the PS stratification approach is a favorable choice often with 5 strata to successfully remove about 90% bias (Cochran, 1968). Nevertheless, this approach involves deletion of all unmatched observations which in a sense may lose many sample cases.

Considering the compatibility with sampling weights, the IPTW approach seems to be more satisfactory. Computationally, the IPTW can be treated as sampling weights, which create a balance between treated and untreated groups and it achieves high internal validity (i.e., causal inference). The application of sampling weights component enables the sample distribution resembles that in the population to achieve external validity (i.e., generalizability). It should be noted that other three approaches could also be implemented via software (Ho, Imai, King, & Stuart, 2011).

The analytical procedures are as follows. First, the outcomes of all cognitive and non-cognitive variables are computed as gain scores by subtracting the fall scores from the spring scores. The purpose is to reduce variation and create a value-added setting. Second, PS is computed based on selected covariates. The logit model will regress the binary small class variable on about forty covariates. The baseline of covariates in fall include child demographic information (i.e., gender, age, race and home language), parent and family background (e.g., parent's age, race, education and occupation, and the number of people in household), pre-kindergarten care, teacher quality (e.g., teacher's education and teaching years) and school characteristics (e.g., school sector and location). Covariates missing values are replaced with median values for continuous variables and zero for binary variables and missing dummy flags are created and included in the model to control for any missing data effect.

Third, based on the generated PS, weights will be computed using the formula below for the treated and untreated individuals (Hong, 2010a, 2012a; Leite et al., 2015).

$$W_i = 1 \; if \; T = 1; \; W_i = \frac{p(T_i=1|X)}{p(T_i=0|X)} \; if \; T = 0 \tag{3.2}$$

Fourth, the linear regression will regress each of the outcomes on the small class treatment variable in the following model using PS as weights.

$$Y_i = \beta_0 + \beta_1 CS_i + e_i \quad (3.3)$$

Robust standard errors will be used to control for heterogeneity. It needs to mention that since the PS weights reconstruct the distribution of the observations, so it does not make sense to control for the school clustering anymore, which in fact does not have an influence once the PS weights apply. Finally, covariates balance between two treatment groups will be checked before and after applying the PS weights for a diagnosis. For sensitivity analysis, results from the PS method will be compared with that of no PS weight or plus sampling weights. The used sample weight is W12AC0 which is a child base weight adjusting for nonresponse associated with spring kindergarten teacher-level questionnaire and the fall kindergarten child assessment. The main reason to choose this sampling weight variable is that the main variable of interest (i.e., class size) comes from spring teacher-level questionnaire also since gain score is the outcome, the children's assessment in fall kindergarten is also relevant.

In addition to full sample analysis, the same analysis steps will be repeated for subgroup samples concerning parents' education, family poverty level and school sector without sampling weights. The dummy variable of parents' education of Bachelor's degree is created based on two original variables about first and second parent's education. That is the value is one if at least one parent has a bachelor and above degree and zero otherwise. The family poverty level is also a binary variable, in which one stands for at or above 200 percent of poverty threshold and zero is for below 200 percent of poverty threshold. The school sector is private versus public schools. Possible bias will be quantified using the same method as in the fixed effect analysis.

3.5 Results

3.5.1 Fixed effects results

Table 3.1.1 reports unweighted descriptive statistics of the variables of interest in the fixed effects models. The mean of the IRT scores represents the average children's reading and mathematics ability. It shows that the first wave has lower mean score than the following three waves of scores that remain stable. The values of self-control, interpersonal skills, and approaches to learning are relatively high with an approximate mean at 3 in a scale from 1 to 4 and the sample has relatively lower values (i.e., <2) on externalizing and internalizing behavior problems. The average class size is round 20 in Kindergarten and about 21 in first- and second-grade. The school location is in city or suburban on average.

The results for the FE models with full sample are presented in Table 3.1.2 in which the % bias is reported below in the parentheses following the coefficients. It shows that the coefficients of class size are negative and significant for both reading and mathematics achievement at the level of 0.05 (the same significance level hereafter). This indicates that one unit decreases in the class size, each child's ability scores increase in reading by 0.005 (29.45%) and in mathematics by 0.004 (17.82%) for unweighted analyses. The weighted counterparts are a little bit larger and need more cases to invalidate the inference, that are 0.008 (63.22%) and 0.005 (41.83%) for reading and mathematics respectively. It is evidence that the weighted and unweighted estimates are consistent and the evidence is robust in terms of the threshold bar of 30% in social science. Obviously, small class size benefits on both cognitive measures are observed. Nonetheless, there is no evidence of class size effects on the six non-cognitive outcomes.

Table 3.1.3 to Table 3.1.5 report the results for sub-groups analyses by gender, race and school type. Table 3.1.3 represents the results by gender, which shows children in smaller

classrooms have better performance on reading achievement for both boys and girls, while regarding mathematics achievement, the benefit of having smaller class size is only observed for girls but not for boys. In addition, class size effects on non-cognitive outcomes are not significant.

Table 3.1.4 shows the results by race (i.e., white, black, Hispanic and Asian). Overall, the results for the White are consistent to that in the full sample analysis. With regard to the Hispanic children, increasing class size has negative effect on interpersonal skills with the coefficient of 0.007 (9.55%). This result may suggest that smaller classroom may help Hispanic children increase interpersonal communication skill.

Table 3.1.5 presents the results by school type (i.e., public and private school). In general, results show that small class size has significant positive effect on both reading (0.004) and mathematics (0.004) for public school, but only on reading (0.009) for private schools. There is no significant finding on non-cognitive outcomes.

3.5.2 Propensity score estimates

With regard to the sample size, there are more children in non-small classrooms (i.e.., 8,514 students) than in small classrooms (i.e., 5,766 students). Children in smaller classrooms have slightly higher reading and math scores and lower behavior in approaches to learning.

It needs to mention that here the class size is a dummy class size variable with one indicating small class size, which thus the results would have opposite coefficient sign to be congruent with the FE analyses of continuous class size variable. The estimates of the propensity score analysis are present in Table 3.2.1. For the full sample weighted analysis, the coefficients of class size treatment variable are 0.049 and 0.043 for reading and math score respectively both

of which are statistically significant at the 0.05 level. For subgroup analysis, results show that children whose parent has lower education are likely to get benefits from being placed in small-sized classroom on both reading (0.065) and math (0.053). In public school, reducing class size increases children's reading score (0.054). In general, the evidence in public school is more robust (19.13%).

Table 3.2.2 reports balance check using p-values of t-tests for full sample concerning covariates including both observed student, teacher and school variables and corresponding missing flags. When p-value is less than 0.05, it indicates that individuals in small class is significantly different from those in large class on that particular covariate. Before applying PS weights, there are considerable differences between treatment groups on many covariates for each of the three subgroup analyses while after applying the PS weights, balance has been improved. The analysis was repeated to check the balance for sampling weighted models. When the balance test failed for a few variables, they were included as covariates to be further controlled for in the final model when estimating treatment effect. Table 3.2.3 contains similar test results for subgroup analysis by family poverty level, parent education level and school sector. Propensity score matching successfully reduced systematic difference between treatment groups especially when family is in low poverty level, when parent has lower education level and in public schools. It is interesting to find that in private school, covariates difference between treatment groups was not observed, which homogeneous class size condition may be more homogeneous there.

3.6 Discussion

The idea of having small classes as a strategy to improve the quality of public education especially in early grades (e.g., K-3) is enormously popular in U.S. In response to the popular appeal, over the past four decades, billions of dollars have been spent to enact class size reduction (CSR) initiatives across the nation in more than twenty states (e.g., large-scale CRS programs in Californian and Florida) (M. M. Chingos, 2011). However, with the exception of the extensive evidence from Project STAR dated back to the 1980s, currently there is surprisingly few high-quality studies to provide updated rigorous evidence of class size effects in U.S.

This study sheds new light on the long-term as well as causal inference of class size effects on early children's academic outcomes and non-cognitive skills by employing individual fixed effects model and propensity scores methods. To account for possible missing variable issue, the quantification of potential bias needed to invalidate the inference is computed for significant results. Considering potential design effect, both sample weighted and unweighted analyses have been performed. In addition, subgroup analyses are also conducted.

The results indicate that PS methods generate larger estimates than the FE approach. It could imply that causal method plus gain score yields similar results as were observed in previous experiment studies where the evidence of class size effects is more pronounced. Another possible explanation is that class size benefit may be stronger in Kindergarten than in the following two years.

The results of long-term class size effects from FE methods indicate reducing class size is associated with an increase in children's reading and mathematics ability scores over time. It should be noted that the reading and math are standardized scale scores, so the coefficient estimates are in standard unit. The finding is stronger for reading (0.05-0.08 standard deviations

of the ability distribution / per 10 students reduction) compared with mathematics (0.04-0.05 /per 10 students). Also the evidence on mathematics ability scores differs by gender and by school sector. In terms of the race, small class benefits on cognitive outcomes are only observed for the White children. Nevertheless, decreasing class size may increase Hispanic children's interpersonal communication skills. The causal inference of class size effects from PS method revealed that small class increases children's performance on reading (0.49/per 10 students) and mathematics (0.43/per 10 students).

Overall, this research shows that class size does matter and it increases student academic achievement in early grades. The finding of cognitive outcomes is robust and in line with previous literature. For example, findings from FE analyses (i.e., magnitude of class size effects ranging from 0.04-0.08 in 10 units of students for mathematics and reading) are in congruence with quasi-experimental studies. For example, Cho et al (2012) found that a decrease of ten student in elementary grades in Minnesota would increase reading and mathematics test scores by 0.04-0.05 standard deviations (Cho et al., 2012). Rivkin et al (2005) found the estimated effects were between 0.08-0.11 standard deviations per 10-student decrease for 4^{th} and 5^{th} grade, 0.03 standard deviations for 5^{rd} grade reading and 0.04 standard deviations for 6^{th} grade mathematics. However, the magnitude of causal class size effects from PS analyses in Kindergarten in this study ranges from 0.43 to 0.49 for decreasing 10 students, which is larger than Krueger (1999)'s finding in the STAR experiment where dropping eight students in kindergarten classrooms would increase children's math and reading tests by 0.2 standard deviation on math and reading, which equals to 0.25 standard deviations every 10 students (Krueger, 1999). In addition, in this study, there is little evidence showing minority students get

more benefits for being placed in small-sized classes on academic achievement which is also different to the some previous finding in Project STAR (Finn & Achilles, 1990).

In terms of the class size effects on non-cognitive outcomes, FE models produced the evidence that assuming a linear relationship between class size and non-cognitive outcomes, reducing every ten students in kindergarten class may help to improve Hispanic children's interpersonal skills by 0.07 standard deviations but the evidence is weak. This positive finding is congruent with Dee and West (2011)'s findings in 8th grade where smaller classes have two-year persisting effects on school engagement with effect size ranging from 0.05 to 0.09 and the return of class size reduction in 8th grade on non-cognitive skills is about 0.05 overall and 0.08 in urban school.

In summary, large-scale CSR may fail cost-benefit test in the state level from the economic perspective (Yeh, 2009). Nevertheless, it should be emphasized that the cost of class-size reduction can be easily measured at a fixed time point, but the short and long-term small class benefits on non-cognitive skills are still unknown, which nonetheless needs to be taken into account when making a policy decision on class size issue. There was evidence showing that small class size benefits in early childhood education faded on the test scores, but re-emerged in adulthood and the suggestive explanation is because of non-cognitive skills such as efforts, initiatives, and lack of disruptive behavior which are highly correlated with earnings even controlling on test scores (Chetty et al., 2011). It is noteworthy that these non-cognitive skills can never be easily acquired by young children in large classes as they could be in small classes (Douglass & Parkhurst, 1940). Also early intervention matters to foster human capital and to change the current phenomenon in American society where the issue is that the very young children are under-invested while low-skilled old adults are over-invested (Heckman, 2000).

With the increasing recognition of the importance of non-cognitive skills in fulfilling academic attainment, excelling in the labor-market performance and achieving life success (Heckman, Stixrud, & Urzua, 2006; Kautz et al., 2014; Lindqvist & Vestman, 2011), there is a rush to embrace this missing piece in education. However, as the survey responses tend to be influenced by the social context, the findings from current available but flawed measures of non-cognitive skills could be misleading. In order to capitalize the potential values of these non-cognitive skills, knowing how to reliably measure these traits is the key. Therefore, future research on the effectiveness of class size reduction on cultivating non-cognitive skills calls for an improvement on measuring each specific non-cognitive category, which would then be the foundation for generating scientific evidence in this domain. For instance, in addition to parent and teacher rating, observational measures using expert rating on video-recording may be valuable and informative.

APPENDIX

APPENDIX 3:

Variable List	Scale	Kindergarten Fall	Kindergarten Spring	1st Grade Spring	2nd Grade Spring
Outcomes					
Reading IRT scale scores	Continuous	X1RTHETK2	X2RTHETK2	X4RTHETK2	X6RTHETK2
Mathematics IRT scale scores	Continuous	X1MTHETK2	X2MTHETK2	X4MTHETK2	X6MTHETK2
Self-control - teacher	Continuous	X1TCHCON	X2TCHCON	X4TCHCON	X6TCHCON
Interpersonal skill - teacher	Continuous	X1TCHPER	X2TCHPER	X4TCHPER	X6TCHPER
Externalizing problems - teacher	Continuous	X1TCHEXT	X2TCHEXT	X4TCHEXT	X6TCHEXT
Internalizing problems - teacher	Continuous	X1TCHINT	X2TCHINT	X4TCHINT	X6TCHINT
Approaches to learning - teacher	Continuous	X1TCHAPP	X2TCHAPP	X4TCHAPP	X6TCHAPP
Attention focus - teacher	Continuous	X1ATTNFS	X2ATTNFS	X4ATTNFS	NA
Inhibitory control - teacher	Continuous	X1INBCNT	X2NBCNT	X4INBCNT	NA
Self-control - parent	Continuous	X1PRNCON	X2PRNCON	X4PRNCON	NA
Social interaction - parent	Continuous	X1PRNSOC	X2PRNSOC	X4PRNSOC	NA
Impulsive/overactive - parent	Continuous	X1PRNIMP	X2PRNIMP	X4PRNIMP	NA
Sad/lonely - parent	Continuous	X1PRNSAD	X2PRNSAD	X4PRNSAD	NA
Approaches to learning - parent	Continuous	X1PRNAPP	X2PRNAPP	X4PRNAPP	NA
Covariates					
Class size (main predictor)	Continuous	A1DTOTAG	A2DENROL	A4DENROL	A6DENROL
Child gender (1=male)*	Binary	X_CHSEX_R			
Child assessment age in month *	Continuous	X1KAGE_R			
Child race (1=white) *	Binary	X_WHITE_R			
Child home language (1=English)*	Binary	X12LANGST			
Parent one age *	Continuous	X1PAR1AGE			
Parent one race (1=white) *	Binary	X1PAR1RAC			
Parent one education (1=bachelor and beyond) *	Binary	X12PAR1ED_I			
Parent one employment (1=more than 35 hours per week) *	Binary	X1PAR1EMP			
Parent one occupation prestige *	Binary	X1PAR1SCR_I			
Mother marriage status at birth (1=married) *	Binary	X12MOMAR			
Total number of people in household *	Continuous	X1HTOTAL			
Number of siblings in household	Continuous	X1NUMSIB			
Primary type of care (1=parental care) *	Binary	X1PRIMNW			
Hours spent in non-parental care now *	Continuous	X1HRSNOW			
Teacher highest education (1=master's degree and beyond)*	Binary	A1HGHSTD			
Number of years taught at this school	Continuous	A1YRSCH			
Number of years been school teacher*	Continuous	A1YRSTCH			
Taken exam for national board (1=take and passed) *	Binary	A1NATEXM			
School type (1=private school)	Binary	X1PUBPRI			
School location (1=City;2=Suburb;3=Town;4=Rural)	Categorical	X1LOCALE			
School district composite poverty level	Continuous	X_DISTPOV			
Sixteen aggregated school variables indicated with *	Continuous	As above			

TABLES

Table 3.1.1: Descriptive Statistics for fixed effects analysis

	Kindergarten Fall					Kindergarten Spring					Grade1 Spring					Grade2 Spring				
Main variables	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Reading IRT scores	9,379	-0.48	0.85	-3.09	2.98	9,478	0.51	0.76	-2.69	2.98	9,478	0.51	0.76	-2.69	2.98	9,478	0.51	0.76	-2.69	2.98
Mathematics IRT scores	9,350	-0.41	0.91	-5.86	5.32	9,455	0.51	0.74	-5.85	2.88	9,383	1.72	0.83	-1.87	4.64	9,326	2.52	0.78	-2.65	6.54
Self-control	8,219	3.12	0.61	1	4	8,962	3.23	0.62	1	4	8,418	3.24	0.61	1	4	8,418	3.24	0.61	1	4
Interpersonal skill	8,298	3.03	0.63	1	4	8,968	3.18	0.63	1	4	8,472	3.17	0.65	1	4	8,736	3.15	0.66	1	4
Externalizing problem behaviors	8,717	1.55	0.59	1	4	9,017	1.58	0.60	1	4	8,520	1.70	0.60	1	4	8,823	1.69	0.60	1	4
Internalizing problem behaviors	8,628	1.45	0.48	1	4	9,002	1.47	0.47	1	4	8,476	1.52	0.49	1	4	8,782	1.57	0.51	1	4
Approaches to learning	8,948	2.99	0.67	1	4	9,061	3.15	0.67	1	4	8,555	3.12	0.69	1	4	8,840	3.11	0.70	1	4
Class size	7,102	20.11	3.94	11	29	7,400	20.08	4.10	9	29	8,545	20.92	4.24	8	33	8,800	21.40	4.37	9	34
School location	9,517	2.29	1.15	1	4	9,517	2.29	1.15	1	4	9,311	2.27	1.14	1	4	9,311	2.27	1.14	1	4

Table 3.1.2: Fixed effect model for class size in full sample (cluster robust SE)

Outcome		Full Sample				
		Est	Sig	SE	N	%
Cognitive	Reading	-0.0047	*	0.0017	30,804	29.45
	(W)	-0.0075	*	0.0014	21,808	63.22
	Mathematics	-0.0036	*	0.0015	30,757	17.82
	(W)	-0.0047	*	0.0014	21788	41.83
Non-cognitive	Self-control	0.0011		0.0020	29,877	
	(W)	0.0010		0.0018	21,158	
	Interpersonal skill	-0.0013		0.0022	30,052	
	(W)	-0.0020		0.0019	21,297	
	Externalizing behavior	-0.0010		0.0016	30,485	
	(W)	-0.0010		0.0016	21,561	
	Internalizing behavior	-0.0013		0.0013	30,306	
	(W)	-0.0007		0.0016	21,471	
	Approaches to learning	0.0008		0.0021	30,711	
	(W)	-0.0005		0.0018	21,702	

Note: * p<0.05; W=sample weighted analysis; % indicates the percentage of bias to invalidate the results

Table 3.1.3: Fixed effect model for class size by gender (cluster robust SE)

Outcome		Male					Female				
		Est	sig	SE	N	%	Est	sig	SE	N	%
Cognitive	Reading	-0.0045	*	0.0020	15,515	12.7	-0.0050	*	0.0020	15,229	21.6
	Mathematics	-0.0031		0.0019	15,485		-0.0042	*	0.0017	15,212	20.66
Non-cognitive	Self-control	-0.0005		0.0027	15,054		0.0025		0.0021	14,769	
	Interpersonal skill	-0.0030		0.0029	15,082		0.0004		0.0023	14,916	
	Externalizing behavior	0.0000		0.0022	15,366		-0.0023		0.0017	15,062	
	Internalizing behavior	-0.0017		0.0018	15,250		-0.0010		0.0017	15,000	
	Approaches to learning	0.0006		0.0027	15,476		0.0011		0.0022	15,176	

Note: * p<0.05;W=sample weighted analysis; % indicates the percentage of bias to invalidate the results

Table 3.1.4: Fixed effect model for class size by race (cluster robust SE)

Outcome		White					Black					Hispanic					Asian				
		Est	sig	SE	N	%	Est	sig	SE	N	%	Est	sig	SE	N	%	Est	sig	SE	N	%
Cognitive	Reading	-0.0051	*	0.002	16,161	22.55	-0.0086		0.005	3,306		-0.0048		0.003	7,518		0.0038		0.007	2,110	
	Math	-0.0048	*	0.002	16,149	30.37	-0.0052		0.004	3,303		-0.0022		0.003	7,518		-0.0006		0.008	2,080	
Non-cognitive	Self-control	0.0032		0.003	15,759		0.0062		0.005	3,199		-0.0059		0.003	7,224		0.0064		0.011	2,037	
	Interpersonal skill	0.0002		0.003	15,849		0.0014		0.005	3,210		-0.0072	*	0.003	7,260	9.55	0.0077		0.010	2,053	
	Externalizing	0.0002		0.002	16,011		-0.0055		0.005	3,262		0.0007		0.002	7,438		-0.0059		0.009	2,086	
	Internalizing	-0.0020		0.002	15,977		-0.0043		0.003	3,223		0.0002		0.003	7,364		0.0051		0.005	2,057	
	Approaches to learning	0.0026		0.003	16,086		0.0031		0.005	3,292		-0.0053		0.003	7,511		0.0068		0.010	2,119	

Note: * p<0.05; W=sample weighted analysis; % indicates the percentage of bias to invalidate the results

Table 3.1.5: Fixed effect model for class size by school sector (cluster robust SE)

Outcome		Public					Private				
		Est	sig	SE	N	%	Est	sig	SE	N	%
Cognitive	Reading	-0.0038	*	0.0018	27,239	8	-0.0085	*	0.0035	3,565	19.32
	Mathematics	-0.0039	*	0.0017	27,194	14.52	-0.0026		0.0030	3,563	
Non-cognitive	Self-control	-0.0003		0.0023	26,454		0.0073		0.0047	3,419	
	Interpersonal skill	-0.0028		0.0024	26,631		0.0059		0.0055	3,417	
	Externalizing behavior	-0.0011		0.0018	27,027		-0.0007		0.0041	3,454	
	Internalizing behavior	-0.0012		0.0015	26,845		-0.0015		0.0030	3,457	
	Approaches to learning	-0.0008		0.0023	27,209		0.0078		0.0051	3,498	

Note: * p<0.05; W=sample weighted analysis; % indicates the percentage of bias to invalidate the results

Table 3.2 1: The estimates of class size using propensity score methods

	Reading					Math				
	Est	Sig	SE	N	%	Est	Sig	SE	N	%
Full sample										
PS weights	0.039	+	0.020	13986		0.033	+	0.019	13905	
PS weights + sampling weights	0.049	*	0.024	13116	5.36	0.043	*	0.020	13041	9.11
Subgroup sample										
(1) Family poverty level										
High										
PS weights	0.022		0.023	5528		0.021		0.022	5509	
Low										
PS weights	0.055	+	0.029	5192		0.043		0.028	5161	
(2) Parent education										
High										
PS weights	0.029		0.020	8494		0.028		0.020	8443	
Low										
PS weights	0.065	*	0.030	5492	9.23	0.053	*	0.027	5462	1.67
(3) School sector										
Private school										
PS weights	-0.020		0.049	2013		-0.079		0.041	2001	
Public school										
PS weights	0.054	*	0.022	12745	19.13	0.025		0.021	12713	

Note: * $p<0.05$ + $p<0.1$ (School cluster robust standard error was used for all models)

Table 3.2.2: Covariates balance check (full sample) in propensity score analysis

	Full sample		
	No Sampling Weights		+sampling weights
Covariates	Before match	After match	After match
Male child	0.18	0.35	0.55
Child age	0.01	0.03	0.02
Child race white	0.01	0.18	0.95
Speak English at home	0.00	0.61	0.42
Parent one age	0.40	0.02	0.04
Parent one race (white)	0.01	0.14	0.95
Parent one bachelor degree and beyond	0.59	0.00	0.00
Parent one work full-time	0.00	0.22	0.38
Parent one job prestige	0.06	0.00	0.00
Mother marriage at child birth	0.59	0.04	0.06
Total number of people at home	0.03	0.45	0.52
Number of siblings	0.82	0.70	0.58
parental primary care	0.79	0.84	0.89
Hours spent for nonparental care	0.01	0.27	0.37
Teacher degree in master and above	0.01	0.40	0.58
Teaching years in this school	0.13	0.20	0.88
Years of being a school teachers	0.18	0.17	0.67
Passed exam in national board	0.32	0.44	0.54
City	0.00	0.96	0.31
Suburban	0.26	0.31	0.22
Town	0.60	0.96	0.72
School district composite poverty level	0.05	0.82	0.78
Proportion of male students in school	0.32	0.47	0.26
Child average age in school	0.01	0.02	0.01
Proportion of white children in school	0.00	0.13	0.82
Proportion of native speakers in school	0.00	0.23	0.88
Parent one average age in school	0.29	0.00	0.00

Table 3.2.2. cont'd			
Proportion of white parents in school	0.01	0.11	0.89
Proportion of parents with bachelor degree in school	0.52	0.00	0.00
Proportion of parents work full-time in school	0.00	0.08	0.17
Parent one job prestige average in school	0.12	0.00	0.00
Proportion of mother who married at child birth in school	0.71	0.05	0.05
Average number of family in school	0.03	0.25	0.37
Proportion of parental care in school	0.68	1.00	0.55
Average number of hours for non-parental care in school	0.00	0.30	0.35
Proportion of teachers with master's degrees	0.00	0.38	0.71
Average year of being teachers in school	0.28	0.08	0.32
Proportion of people passed board exam in school	0.43	0.60	0.90
Child age - missing flag	0.00	0.01	0.34
Speak English at home - missing flag	0.03	0.94	0.97
Parent one race (white) - missing flag	0.04	0.22	0.61
Parent one work full-time - missing flag	0.03	0.19	0.67
Parent one job prestige - missing flag	0.00	0.07	0.62
Mother marriage at child birth - missing flag	0.03	0.90	0.88
Total number of people at home - missing flag	0.02	0.20	0.66
parental primary care - missing flag	0.02	0.16	0.62
Hours spent for nonparental care - missing flag	0.02	0.15	0.60
Teacher degree in master and above - missing flag	0.00	0.57	0.00
Passed exam in national board - missing flag	0.00	0.65	0.01
School location - missing flag	0.83	0.47	0.70

Table 3.2.3: Covariates balance check (subgroup sample)

	Family poverty level				Parent education				School sector			
	High		Low		High		Low		Private		Public	
Covariates	Before match	After match	Before match	After match	Before match	After match	Before match	After match	Before match	After match	Before match	After match
Male child	0.58	0.33	0.31	0.76	0.56	0.75	0.20	0.29	0.14	0.88	0.36	0.53
Child age	0.63	0.02	0.00	0.45	0.19	0.01	0.00	0.32	0.33	0.81	0.00	0.62
Child race white	0.02	0.57	0.00	0.82	0.17	0.10	0.00	0.84	0.58	0.79	0.06	1.00
Speak English at home	0.00	0.31	0.00	0.92	0.00	0.35	0.00	0.92	0.92	0.47	0.00	0.70
Parent one age	0.24	0.00	0.56	0.74	0.76	0.01	0.61	0.52	0.47	0.70	0.00	0.59
Parent one race (white)	0.07	0.35	0.00	0.71	0.18	0.02	0.00	0.79	0.37	0.62	0.13	0.69
Parent one bachelor+	0.13	0.00	0.17	0.03	0.69	0.00	NA	NA	0.77	0.99	0.00	0.71
Parent one work full-time	0.00	0.89	0.01	0.38	0.00	0.06	0.01	0.49	0.23	0.24	0.00	0.45
Parent one job prestige	0.01	0.01	0.07	0.11	0.03	0.00	0.03	0.31	0.17	0.84	0.43	0.74
Mother marriage at child birth	0.43	0.09	0.38	0.20	0.82	0.00	0.12	0.80	0.57	0.77	0.09	0.74
Total number of people at home	0.03	0.82	0.02	0.67	0.00	0.82	0.27	0.56	0.70	0.57	0.11	0.99
Number of siblings	0.20	0.65	0.62	0.74	0.36	0.75	0.95	0.71	0.20	0.34	1.00	0.80
parental primary care	0.99	0.31	0.19	0.54	0.81	0.26	0.09	0.66	0.69	0.43	0.92	0.56
Hours of nonparental care	0.00	0.78	0.92	0.54	0.01	0.87	0.29	0.27	0.36	0.99	0.00	0.90
Teacher degree in master+	0.00	0.15	0.04	0.82	0.02	0.18	0.01	0.76	0.97	0.79	0.03	0.94
Teaching years in this school	0.49	0.33	0.11	0.33	0.29	0.13	0.06	0.44	0.86	0.61	0.17	0.59
Years of being a school teachers	0.27	0.09	0.28	0.27	0.21	0.07	0.22	0.28	0.61	0.92	0.55	0.68
Passed exam in national board	0.93	0.74	0.14	0.39	0.77	0.60	0.10	0.48	0.33	0.98	0.08	0.79
City	0.04	0.57	0.00	1.00	0.00	0.92	0.00	0.72	0.36	0.70	0.00	0.86
Suburban	0.09	0.51	0.38	0.51	0.49	0.37	0.15	0.58	0.94	0.79	0.09	0.78
Town	0.43	0.89	0.52	0.87	0.34	0.92	0.81	0.98	0.14	0.82	0.79	0.81
School district composite poverty	0.00	0.96	0.62	0.90	0.05	0.99	0.23	0.60	0.12	0.86	0.05	0.60
Proportion of male students in school	0.12	0.63	0.91	0.54	0.20	0.50	0.75	0.52	0.29	0.74	0.56	0.46
Child average age in school	0.31	0.01	0.00	0.11	0.28	0.01	0.00	0.12	0.59	0.73	0.00	0.65
Proportion of white children in school	0.05	0.20	0.00	0.50	0.05	0.11	0.00	0.47	0.90	0.52	0.04	0.90
Proportion of native speakers in school	0.01	0.08	0.00	0.52	0.00	0.18	0.00	0.39	0.66	0.99	0.00	0.97
Parent one average age in school	0.18	0.00	0.29	0.20	0.76	0.00	0.08	0.27	0.35	0.63	0.00	0.56
Propotion of white parents in school	0.14	0.10	0.00	0.55	0.10	0.09	0.00	0.40	0.71	0.69	0.12	0.97
Proportion of parents with bachelor+in school	0.15	0.00	0.62	0.08	0.95	0.00	0.20	0.13	0.23	0.15	0.00	0.66
Proportion of parents work full-time in school	0.00	0.03	0.00	0.21	0.00	0.02	0.00	0.27	0.17	0.47	0.01	0.82
Parent one job prestige average in school	0.12	0.00	0.06	0.06	0.09	0.00	0.13	0.02	0.63	0.99	0.24	0.85
Proportion of mother marrital status in school	0.25	0.00	0.13	0.60	0.91	0.01	0.28	0.49	0.81	0.62	0.11	0.64
Average number of family in school	0.17	0.12	0.04	0.67	0.01	0.13	0.11	0.63	0.44	0.63	0.12	0.94
Proportion of parental care in school	0.95	0.60	0.39	0.52	0.98	0.78	0.41	0.92	0.28	0.97	0.68	0.62
Average number of non-parental care in school	0.00	0.33	0.12	0.50	0.00	0.35	0.01	0.49	0.88	0.91	0.00	0.72
Proportion of teachers with master's degrees	0.01	0.11	0.01	0.99	0.01	0.15	0.00	0.82	0.71	0.93	0.01	0.98
Average year of being teachers in school	0.26	0.01	0.37	0.18	0.34	0.04	0.27	0.16	0.57	0.96	0.80	0.51
Proportion of teacher passed exam in school	0.98	0.84	0.31	0.69	0.84	0.81	0.17	0.75	0.54	0.00	0.18	0.94
Child age - missing flag	0.00	0.02	0.03	0.54	0.00	0.02	0.01	0.21	0.91	0.83	0.00	0.00
Speak English at home - missing flag	0.91	0.42	0.21	0.85	0.12	0.28	0.35	0.96	0.70	0.59	0.07	0.91
Parent one race (white) - missing flag	0.59	0.12	0.16	0.75	0.20	0.04	0.12	0.51	0.51	0.65	0.19	0.28
Parent one work full-time - missing flag	0.31	0.10	0.10	0.95	0.12	0.03	0.15	0.49	0.39	0.44	0.17	0.28
Parent one job prestige - missing flag	0.00	0.45	0.03	0.31	0.00	0.19	0.02	0.15	1.00	0.89	0.03	0.35
Mother marital status - missing flag	0.97	0.70	0.16	0.92	0.14	0.21	0.65	0.61	0.93	0.82	0.09	0.79
Total number of people at home - missing flag	0.38	0.08	0.12	0.73	0.11	0.05	0.14	0.52	0.61	0.64	0.15	0.23
parental primary care - missing flag	0.22	0.08	0.06	0.78	0.10	0.04	0.17	0.46	0.55	0.59	0.16	0.28
Hours spent for nonparental care - missing flag	0.20	0.08	0.05	0.75	0.08	0.04	0.18	0.41	0.49	0.60	0.15	0.29
Teacher degree in master + - missing flag	0.00	0.95	0.00	0.42	0.00	0.53	0.00	0.83	0.12	0.67	0.00	0.92
Teacher passed exam - missing flag	0.00	0.98	0.00	0.98	0.00	0.96	0.00	0.99	0.35	0.91	0.00	0.96
School location - missing flag	0.64	0.50	0.73	0.51	0.60	0.56	0.62	0.41	0.52	0.47	0.58	0.79

REFERENCES

REFERENCES

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

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of educational psychology, 84*(3), 261-271.

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.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424.

Babcock, P., & Betts, J. R. (2009). Reduced-class distinctions: Effort, ability, and the education production function. *Journal of Urban Economics, 65*(3), 314-322.

Barrett, N., & Toma, E. F. (2013). Reward or punishment? Class size and teacher quality. *Economics of Education Review, 35*(4), 41-52.

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and Instruction, 21*(6), 715-730.

Blatchford, P., Russell, A., Bassett, P., Brown, P., & Martin, C. (2007). The effect of class size on the teaching of pupils aged 7–11 years. *School Effectiveness and School Improvement, 18*(2), 147-172.

Bonesrønning, H. (2003). Class size effects on student achievement in Norway: Patterns and explanations. *Southern Economic Journal, 96*(4), 952-965.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics, 126*(4), 1593-1660.

Chingos, M. M. (2011). *The False Promise of Class-Size Reduction: Introduction and Summary*. Retrieved from Center for American Progress.

Chingos, M. M. (2012). The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review, 31*(5), 543-562.

Chingos, M. M. (2013). Class size and student outcomes: Research and policy implications. *Journal of Policy Analysis and Management, 32*(2), 411-438.

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.

Cochran, W. G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics, 24*(2), 295-313.

Dee, T. S., & West, M. R. (2011). The non-cognitive returns to class size. *Educational Evaluation and Policy Analysis, 33*(1), 23-46.

Douglass, H. R., & Parkhurst, A. (1940). Size of class and teaching load. *Review of Educational research*, 216-221.

Educational Research Service. (1980). Class size research: a critique of recent meta-analyses. *Phi Delta Kappan, 62*, 239-241.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest*, 1-30.

Farrington, C. A., Beechum, N. O., Johnson, D. W., Keyes, T. S., Nagaoka, J., Allensworth, E., . . . Consortium on Chicago School, R. (2012). *Teaching Adolescents to Become Learners: The Role of NoncognitiveFactors in Shaping School Performance--a Critical Literature Review*.

Finn, J. D. (2002). Small classes in American schools: Research, practice, and politics. *Phi Delta Kappan, 83*(7), 551-560.

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

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21*(2), 97-109.

Finn, J. D., Gerber, S. B., & Boyd-Zaharias, J. (2005). Small classes in the early grades, academic achievement, and graduating from high school. *Journal of educational psychology, 97*(2), 214-223.

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis, 35*(4), 437-460.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). *School class size: research and policy*. Hills, CA: Sage.

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.

Gottfried, M. A., & Le, V.-N. (2016). Full-Versus Part-Day Kindergarten for Children With

Disabilities: Effects on Academic and Social-Emotional Outcomes. *American Educational Research Journal, 53*(3), 708-744.

Graue, E., Hatch, K., Rao, K., & Oen, D. (2007). The wisdom of class-size reduction. *American Educational Research Journal, 44*(3), 670-700.

Griswold, M. E. P., Localio, A. R. P., & Mulrow, C. M. D. M. (2010). Propensity Score Adjustment With Multilevel Data: Setting Your Sites on Decreasing Selection Bias. *Annals of Internal Medicine, 152*(6), 393.

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications (2nd ed.)*. Thousand Oaks: Sage.

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis, 21*(2), 143-163.

Hanushek, E. A. (2002). Evidence, politics, and the class size debate. *The class size debate*, 37-65.

Hanushek, E. A., Mayer, S. E., & Peterson, P. (1999). The evidence on class size. *Earning and learning: How schools matter*, 131-168.

Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research, 43*(6), 387-425.

Heckman, J. J. (2000). Policies to foster human capital. *Research in economics, 54*(1), 3-56.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411-482.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software,, 42*(8).

Hong, G. (2010a). Marginal Mean Weighting Through Stratification: Adjustment for Selection Bias in Multilevel Data. *Journal of educational and behavioral statistics, 35*(5), 499.

Hong, G. (2010b). Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of educational and behavioral statistics, 35*(5), 499-531.

Hong, G. (2012a). Marginal Mean Weighting Through Stratification: A Generalized Method for Evaluating Multivalued and Multiple Treatments With Nonexperimental Data. *Psychological methods, 17*(1), 44-60.

Hong, G. (2012b). Marginal mean weighting through stratification: a generalized method for

evaluating multivalued and multiple treatments with nonexperimental data. *Psychological methods, 17*(1), 44.

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

Jepsen, C., & Rivkin, S. (2009). Class Size Reduction and Student Achievement The Potential Tradeoff between Teacher Quality and Class Size. *Journal of Human Resources, 44*(1), 223-250.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., Borghans, L., & National Bureau of Economic, R. (2014). Fostering and measuring skills : improving cognitive and non-cognitive skills to promote lifetime success.

King, G., & Nielsen, R. (2016). Why Propensity Scores Should Not Be Used for Matching.

Konstantopoulos, S. (2008). Do small classes reduce the achievement gap between low and high achievers? Evidence from Project STAR. *The Elementary School Journal, 108*(4), 275-291.

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., & Sun, M. (2014). Are teacher effects larger in small classes? *School Effectiveness and School Improvement, 25*(3), 312-328.

Krueger, A. B. (1999). *Experimental estimates of education production functions*. Retrieved from The Quarterly Journal of Economics.

Krueger, A. B., & Whitmore, D. M. (2001a). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal, 111*(468), 1-28.

Krueger, A. B., & Whitmore, D. M. (2001b). *Would smaller classes help close the black-white achievement gap?* : Industrial Relations Section, Princeton University.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*(3), 265-284.

Li, W., & Konstantopoulos, S. (2016). Class Size Effects on Fourth-Grade Mathematics Achievement: Evidence From TIMSS 2011. *Journal of Research on Educational Effectiveness, 9*(4), 503-530.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive

ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*(1), 101-128.

Messick, S. (1979). Potential uses of noncognitive measurement in education. *Journal of educational psychology, 71*(3), 281-292.

Milesi, C., & Gamoran, A. (2006). Effects of class size and instruction on kindergarten achievement. *Educational Evaluation and Policy Analysis, 28*(4), 287-313.

Mishel, L., & Rothstein, R. (Eds.). (2002). *The Class Size Debate*. Economic Policy Institute.

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. *The future of children*, 113-127.

Mosteller, F., Light, R., & Sachs, J. (1996). Sustained inquiry in education: Lessons from skill grouping and class size. *Harvard educational review, 66*(4), 797-843.

Mulligan, G. M., Hastedt, S., & McCarroll, J. C. (2012). *First-Time Kindergartners in 2010-11: First Findings From the Kindergarten Rounds of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) (NCES 2012-049)*. Retrieved from Washington, DC: National Center for Education Statistics.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis, 21*(2), 127-142.

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. (2001). Are effects of small classes cumulative? Evidence from a Tennessee experiment. *The Journal of Educational Research, 94*(6), 336-345.

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.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2004). Do minorities experience larger lasting benefits from small classes? *The Journal of Educational Research, 98*(2), 94-100.

Nye, B. A., Hedges, L. V., & Konstantopoulos, S. (2000). Do the disadvantaged benefit more

from small classes? Evidence from the Tennessee class size experiment. *American Journal of Education*, 1-26.

Rice, J. K. (1999). The impact of class size on instructional strategies and the use of time in high school mathematics and science courses. *Educational Evaluation and Policy Analysis, 21*(2), 215-229.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. (2009). Field experiments in class size from the early twentieth century. *The Journal of Economic Perspectives*, 211-230.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)*, 597-610.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology, 66*(5), 688.

Shen, T., & Konstantopoulos, S. (2017). Class size effects on reading achievement in Europe: Evidence from PIRLS. *Studies in Educational Evaluation, 53*, 98-114.

Sims, D. (2008). A strategic response to class size reduction: Combination classes and student achievement in California. *Journal of Policy Analysis and Management, 27*(3), 457-478.

Sims, D. (2009). Crowding Peter to educate Paul: Lessons from a class size reduction externality. *Economics of Education Review, 28*(4), 465-473.

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

Stecher, B., Bohrnstedt, G., Kirst, M., McRobbie, J., & Williams, T. (2001). Class-Size Reduction in California: A Story of Hope, Promise, and Unintended Consequences. *Phi Delta Kappan, 82*(9), 670-674.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Hagedorn, M. C., Daly, P., & Najarian, M. (2015). Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011). User's Manual for the ECLS-K: 2011 Kindergarten Data File and Electronic Codebook, Public Version. NCES 2015-074. *National Center for Education Statistics*.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and Paradox: Measuring Students' Non-Cognitive Skills and the Impact of Schooling. *Educational Evaluation and Policy Analysis, 38*(1), 148-170.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis & Management, 26*(3), 455-477.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach (6th ed.)*. Boston, MA Cengage Learning.

Wößmann, L. (2005). Educational production in Europe. *Economic policy, 20*(43), 445-504.

Wößmann, 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.

Yeh, S. S. (2009). Class size reduction or rapid formative assessment?: A comparison of cost-effectiveness. *Educational Research Review, 4*(1), 7-15.

CHAPTER 4
THE CAUSAL CLASS-SIZE EFFECTS IN SECONDARY EDUCATION: EVIDENCE FROM TIMSS

4.1 Introduction

The class size effects have been of intense debate worldwide. The research interest on class size has never faded for a couple of reasons. First of all, small class size represents good education quality in school to some extent. It determines the optimal number of students in a classroom. Subsequently, increasing or decreasing class size could change member composition and unique characteristics of the classroom entity. The common-sense knowledge indicates that smaller class size would facilitate a much better educational environment. It is thus not surprising that the broad public are enthusiastically in favor of smaller-sized class especially in U.S. Second, class size is closely related to teacher labor force and resource allocation. In addition, it is the single factor that could be easily manipulated via administration procedure without disturbing the whole educational system. Therefore, education policymakers may particularly interest in knowing the return of class size reduction or the optimal class size for a particular level of schooling.

Previous correlational studies of class size have shown that overall class size is not significant, but there are significant positive or negative findings of class size effects. This inconsistency partially arises from non-random selection process involving student and teacher characteristics. For example, low-achieving students may be assigned to smaller classroom as a remedy to increase their academic achievement. Similarly, high effective teacher is likely to be assigned to teach larger classroom. In either case, there is spurious positive estimates of class size variable indicating larger class size produces better educational outcome. On the contrary,

parents may send high-achievement students to smaller classroom which thus generates spurious negative correlation between class size and achievement. Due to this sorting or selection possibly occurred in classroom, there could be substantial bias in correlational studies of class size effects along with corresponding meta-analysis studies. Therefore, scientific evidence of class size effects requires causal inference that may generate from either experimental design or quasi-experimental design. Nevertheless, it is not uncommon that randomized trial is not feasible to implement ethically and practically. More often, quasi-experimental approach via statistical methods such as instrumental variable (IV) and regression discontinuity design (RRD) have been adopted to produce causal evidence of class size effects. However, evidence of causal class size is sparse and the context of having small class size benefits is still unclear.

4.2 Literature

In general, educational system varies considerably across different regions and countries. The factor of class size is no exception (Biggs, 1998; Cheung & Chan, 2008). For instance, eastern and western countries have different average class size (i.e., 40 vs. 20 respectively in elementary schools overall) as well as class-size policy. For example, East Asia countries (e.g., Hong Kong and Singapore) with the exception of Japan have adopted small class teaching (SCT) while the U.S. and European countries implemented class size reduction (CSR)(Blatchford, Chan, Galton, Lai, & Lee, 2016).

In the U.S., the public has enthusiastic appealed for small sized class. Tremendous resources have been invested to implement class size reduction in many states. There has been relentless debate and research about class size effectiveness. The most pronounced evidence of class size effects comes from the data of “Student Teacher Achievement Ratio” (STAR). The

project STAR is a large-scale field experiment that was carried out in Tennessee U.S. in 1985-1986, in which students and teachers were randomly assigned to either smaller class (15 students on average) or regular class (23 students on average). Findings from project STAR have suggested that students in smaller classes have significantly higher achievement than students in larger classes and the effects are even long-lasting (Chetty et al., 2011; Finn, Gerber, & Boyd-Zaharias, 2005a; B. Nye, Hedges, & Konstantopoulos, 1999; Barbara Nye, Hedges, & Konstantopoulos, 2000, 2001).

Regarding other studies, applying IV method on the National Education Longitudinal Study of 1988 (NELS:88), Akerhielm found that investing in smaller class size contributes to higher achievement in eighth grade (Akerhielm, 1995). Using the same data set, Dee and West found that smaller classes are associate with more school engagement and the effect persisted two years later (Dee & West, 2011). Nevertheless, the evidence of class size effects in U.S. is inconsistent. For instance, Hoxby (2000) relied upon the random variation of class size due to natural variation in birth and found there is no class size effect in fourth and sixth grades (Hoxby, 2000). However, other researchers applied the same method to compute class size effects in Minnesota and positive effects of smaller classes on student achievement was detected (Cho, Glewwe, & Whitler, 2012).

With regard to international studies, one study is noteworthy. Angrist and Lavy's (1999) introduced an instrument variable of class size based on the maximum class size rule of 40 in Israel elementary school. They found a significant and positive effect of small classes on fifth grade reading and mathematics scores while in fourth grade the evidence was significant in reading, but not in mathematics (J. D. Angrist & Lavy, 1999). This IV method has been adopted by a few researchers around the world.

In Europe, Bonesrønning investigated class size effects using a maximum class-size rule of 30 students per classroom in Norway and results revealed significant but rather small benefit of class size effects in lower secondary schools (i.e., eighth through tenth grades) (Bonesrønning, 2003). Likewise, Browning and Heinesen (2007) used Danish administrative panel data and found small class slightly enhances years of education and upper secondary education completion for eighth graders (Browning & Heinesen, 2007). Using RD and controlling for lagged achievement and school fixed effects, Krassel & Heinesen found significant positive effects of reducing class size on achievement in secondary school in Denmark (Krassel & Heinesen, 2014). Taking advantage of a longitudinal data from Sweden in primary school, Fredriksson, Öckert and Oosterbeek found smaller class size has positive effects on completed education in age 10 to 13 and on wages and earnings at age 27 to 42 (Fredriksson, Öckert, & Oosterbeek, 2013).

In addition, there are some other causal methods that have been employed in class size study. Using a seasonal feature of school system between summer and school-year learning and difference-in-difference (DD) models, Lindahl (2005) found smaller classes result in higher test scores in Sweden (Lindahl, 2005). Utilizing within-school variation over time in the size of subject-specific classes, there was substantial positive effects of class size reduction on French examination in Denmark (Heinesen, 2010). Although the maximum rule does not work, using the variation of total grade enrollment, Gary-Bobo and Mahjoub (2013) found small significant negative effects of larger class size on grade promotion in French junior high school (Gary-Bobo & Mahjoub, 2013). Relying on random assignment of students to teaching classes, De Giorgi, Pellizzari and Woolston (2012) found that reducing class size results in positive outcome on academic and labor market performance for college students (De Giorgi, Pellizzari, & Woolston,

2012). One study eliminates unobserved family effects by using variation of class size between siblings. It revealed that reducing class size has statistically and economically significant effects on increasing mean length of education for about 8 days in post-compulsory schooling in Denmark (Bingley, Myrup Jensen, & Walker, 2005).

Outside Europe, in Japan, based on the maximum class size rule of 40, Akabayashi & Nakamura (2014) found reducing class size has significant positive impact on language test score in 6th grade via value-added model (Akabayashi & Nakamura, 2014). Urquiola (2006) studied third-grade students in Bolivia and found significant class size effects on test scores, with effect sizes as large as 0.30 standard deviations (Urquiola, 2006).

However, using multiple country data in large-scale international studies such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), findings have revealed that overall there is no systematic and significant class size effects on students' cognitive skills (Konstantopoulos & Shen, 2016; Konstantopoulos & Traynor, 2014; Li & Konstantopoulos, 2016; Shen & Konstantopoulos, 2017). For example, Wößmann and West explored class size effects in 15 European countries using data from TIMSS 1995 for eighth grade students (Wößmann & West, 2006). They found only two statistically significant relationships between class size and student achievement: a marginally significant effect in Norway and a highly significant effect in Iceland. Leuven et al (2005) found the class-size effect on student achievement at the end of lower-secondary school is equal to zero in Norway (Leuven, Oosterbeek, & Rønning, 2008). Also there is even perverse evidence of smaller class size in Bangladesh (Asadullah, 2005).

There are several literature voids that have been observed. First, even considering causal inference of class size effects, findings are still inconclusive. Clearly, there is a lack of

consistency regarding whether class size reduction has any significant effect on scholastic achievement. Possible explanation is that class size effects vary with specific context (Wößmann & West, 2006), which the implication is that class size research has to be conducted for separate countries or group of similar countries and also for separate education level (e.g., primary, secondary, or tertiary) (Leuven et al., 2008). Moreover, examining class size effects for separate subjects would be helpful because the content and cognitive domains among different disciplines such as science (e.g., chemistry and biology) become distinct when going beyond elementary level. The heterogeneity of class size effects across subjects is evident. For example, through random assignment of teachers to classes of different size, it was found that larger class has significant and sizable negative effect on student performance in mathematics but not in language skills involving freshmen in a public University in Italy (De Paola, Ponzo, & Scoppa, 2013). Second, evidences of class size have been focused exclusively on scholastic achievement whilst effects on non-cognitive skills (e.g., engagement and attentiveness) has been a missing piece in the literature although these non-cognitive skills have been proposed to be crucial to later success in school and in life. They are also likely to be more malleable to intervention than intelligence traits once developed beyond early childhood (Dee & West, 2011; West et al., 2016). Third, overall, many previous studies used data of relatively small-sized sample rather than nationally representative sample which sets a limitation for generalizing findings to a national level.

4.3 The present study

The purpose of this study is to shed more light on the causal evidence of class size effects on both cognitive and non-cognitive outcomes in eighth grade using national probability samples in four European countries that have participated in TIMSS 2003, 2007 and 2011. Research has

revealed that the cognitive and non-cognitive outcomes of learning a particular subject are strongly interrelated. For example, motivational beliefs (e.g., self-concept) and enjoyment of particular science subject have substantial positive predictive effects on students' cognitive learning, engagement and achievement of that subject (Areepattamannil, Freeman, & Klinger, 2011). The self-concept or self-belief refers to students' self-perceived judgement on a subject (Marsh, Trautwein, Ludtke, Koller, & Baumert, 2005). This study includes four self-belief variables ranking the degree of learning a subject well, learning it quickly, whether the subject is hard or whether it is a weakness.

This study has several specific features that are worth emphasizing. First, through employing appropriate statistical method on large-scale data in multiple countries across years, findings in this study have high internal and external validity. Specifically, the application of Instrumental variable (IV) method and regression discontinuity design (RDD) allows us to generate causal inference of class size effects across four countries, which fulfill the high internal validity. In addition, utilizing sampling weights allows the estimates form the sample to be projected to a particular national population which thus has merit in generalizability.

Second, the recognition of the predictive power of non-cognitive skills over future academic and life outcomes has escalated. Small classroom has been hypothesized to improve non-cognitive development such as learning motivation and school engagement. Thus it would be especially meaningful to empirically investigate the potential contributions of class size on developing malleable non-cognitive skills. Nevertheless evidence of the return of class size on non-cognitive outcome is overall extremely rare expect one study (Dee & West, 2011). Therefore, providing evidence of class size effects on non-cognitive outcomes is badly needed to contribute to the debate on the effectiveness of class size reduction.

Third, this study focused on eighth grade because for self-report non-cognitive outcomes, it is very likely that eighth graders would provide more accurate response than the 4th graders. More importantly, I choose to focus on math and specific sciences including physics, biology, chemistry and science in secondary education because being well-versed in science, technology, engineering and math (STEM) is an educational imperative in a technological society. Especially, in today's education, school faces the challenge of declined interest and enrollment in STEM major in high school and university especially in the U.S. Findings of this research would contribute to knowing whether class size reduction has positive effect on improving high-school students' non-cognitive skills and academic achievement on math and sciences. To my knowledge, the causal inference of class size effects on specific science subjects in secondary education pertaining to both cognitive and non-cognitive outcomes have not been examined before although previous studies used the same TIMSS data.

4.4 Methods

4.4.1 Data

TIMSS is the largest international study that has been used to monitor 4th and 8th grader's mathematics and science achievement over time across approximately fifty countries or economies around the world. High quality assessment data have been provided via IRT model that are equated across participating countries and also over time for each participated country. In addition, it provides rich survey questions about the students' family and school learning environment in the questionnaires for students, parents, teachers and principals.

Data were collected every four years since 1995. This study used data in 2003, 2007 and 2011 to investigate whether the findings would be consistent for a particular country across

time. The latest round data in 2015 was not considered because the variable of 8th enrollment variable is not available for public data file, which however is an essential ingredient to employ the statistical methods. TIMSS data before 2000 (i.e., 1995 and 1999) were not included either because the collected variables then are quite different from surveyed variables after year 2000. The TIMSS sampling design is a two-stage stratified cluster sampling where schools are sampled with probability that is proportional to school size and then one or more intact classroom are selected. In the data user's manual, the sampling weights include the overall student-level weights (i.e., TOTWGT) are recommended to be used to generate population estimates.

4.4.2 Country Selection

Because I targeted at causal inference of class size effects, selected country has to meet two criteria. First, the country should have maximum class size rules to be able to apply the methods of IV and RDD and also the rules should be followed relatively well in order to ensure that the estimates would be valid to some extent. Second, the country should participate in all three waves to examine whether the class size effects in a country would be consistent. Originally, there are seven European countries that data in all three years, but three were eliminated in which Norway and Sweden do not have national maximum class size rules (Eurydice, 2012) (page 153) while in Italy there is an official rule but it was not followed at all as the empirical data displayed.

I thus selected the following four countries that are listed in an alphabetical order with the maximum class size rules included in the parentheses: Hungary (30), Lithuania (30), Romania (30) and Slovenia (28) (Eurydice, 2005, 2009, 2012). It should be noted that although these four countries have maximum class size rules in elementary grade, in secondary education, rules are

available in 2011 only. To my knowledge, the rules of the maximum class size do not change much based on previous study (Shen & Konstantopoulos, 2017) and the conjecture is also supported by the empirical graphs presented in the result part.

In terms of education system, Hungary has a standardized education system governed by the Ministry of Education for providing free education in public schools (Hörner, Döbert, Kopp, & Mitter, 2007). The National Core Curriculum is responsible for prescribing unified content of public educational requirements up to tenth grade. The lower secondary education lasts from grade 5 to grade 8. Lithuania also has a centralized education system under the governance of the Ministry of Education and Science, but schools have some flexibility to make modification as needed. The lower secondary education lasts six years from grade 5 to grade 10. Romania's education is similar to Lithuania. In Slovenia, the lower secondary education is from grade 5 to grade 8, which is regulated by the Council of Experts for General. It is worth noting that these four countries in Central and Eastern Europe have gone through a reform of decentralization transition moving from communism-featured education towards education system similar to Western European countries. According to the transitional development, Hungary and Slovenia belong to the more advanced group whilst Lithuania and Romania are the less advanced group (Ammermüller, Heijke, & Wößmann, 2005).

4.4.3 Variables

The main cognitive variables are the achievement scores in mathematics, physics, biology, chemistry and earth science in the format of five plausible values (PV) derived from multiple imputation (MI) method to produce reliable and consistent estimates of test scores. It is because each individual student only responses to a proportion of the item pool, so the

uncertainty needs to be taken care of via the five PVs. The non-cognitive outcomes are based on questions that ask students to indicate their level of engagement and self-evaluation of learning five subjects. In each subject, five variables are selected which are available across three rounds of data: (1) learning the subject well; (2) learning the subject quick; (3) enjoy learning the subject; (4) the subject is hard; (5) the subject is one's weakness. Each of the five non-cognitive items has a 1-4 Likert-type format (i.e., 1=strongly agree; 2=agree; 3=disagree; and 4=strongly disagree). The first three items were reverse coded so in all five variables, large value would represent better learning outcomes on a particular subject. For convenience, these five non-cognitive variables will be referred to as "well", "quick", "enjoy", "hard" and "weak" hereafter.

The main independent variable is teacher reported 8th grade class size in the classroom of each particular subject (i.e., math, physics, biology, chemistry and earth science). The other covariates include: students' gender, age and a composite variable of home items, teachers' gender, education and experience, and 8th grade enrolment and percentage of economically disadvantaged students in schools. It needs to note that the teacher variables including class size variable come from teacher questionnaire corresponding to each of the five academic subjects. The Appendix provides the list of variables used in this study including coding details.

4.5 Statistical Analysis

I analyzed data of these four countries in TIMSS 2003, 2007, and 2011 via the IV method applied in full sample and RD sample.

4.5.1 IV

The IV method involves finding a valid instrumental variable and then applying the estimation of two stage least squares (TSLS) (Joshua D. Angrist & Imbens, 1995). In the

example of class study, at the first stage regression, class size is regressed on the instrumental variable of class size plus all the covariates and in the second stage regression, the fitted value from the first regression will be used as the predictor for the outcome variables controlling the same covariates that have been use in the first stage. The basic idea of IV method is that the fitted value from the first regression represents the pure class size variable that is uncorrelated from the deleted error term in the first regression that represents the unobserved confounding variables or sorting process. Therefore, the estimated coefficient of the fitted value of class size variable represents the causal effect.

The assumption of the IV method is that the instrumental variable should be valid in terms of meeting two requirements: (1) having high correlation with the actual class size variable which the rule of thumb is F-test of the instrument in the first-stage regression should be larger than 10 (Stock, Wright, & Yogo, 2002); (2) having little or no relationship with the outcome variable. In that way, the fitted value of class size represents an exogenous variable indicating that the change of the outcome is only due to change of the class size variable. However, if the instrumental variable is weak, the estimates could be worse than the OLS estimates.

The equations (4.1 to 4.3) below describe the computation details. First I followed Angrist and Lavy (1999) by creating the Instrument Variable, which is the average class size in 8th grade in each school j as

$$ACS8_j = \frac{Enrollment8_j}{\left\{INT\left[\frac{Enrollment8_j - 1}{rule}\right]\right\} + 1} \tag{4.1}$$

where *ACS8* is the average class size in 8th grade in the school j, *Enrollment8* is the student enrollment in 8th grade, *rule* is the maximum class size rule in each country in each year and *INT* stands for the function of generating the next smaller integer.

In the first stage regression, the teacher reported class size is regressed on the computed class size (i.e., the instrument) along with other covariates. Specifically, for student i in school j the model is

$$ClassSize_{ij} = \pi_0 + \pi_1 ACS8_j + \boldsymbol{STD_{ij}\Pi_2} + \boldsymbol{TCH_{ij}\Pi_3} + \boldsymbol{SCH_j\Pi_4} + \varepsilon_{ij} \quad (4.2)$$

where *ClassSize* is the teacher reported class size, *ACS8* is the average class size computed in each school (i.e., IV), ***STD*** includes student variables (i.e., gender, age, items at home), ***TCH*** represents teacher variables including teacher's gender, education degrees and experience in years, ***SCH*** includes the percent of economically disadvantaged students and grade enrollment with 3rd polynomial function (i.e., linear, quadratic and cubic terms) to capture any possible non-linear relationship. The regression estimates are captured by the $\boldsymbol{\Pi}$**s**.

Next, I computed the predicted or fitted values from equation (4.2) that is known. Meanwhile, I deleted the part of class size that is unknown (the error), which could incorporate selection bias. The fitted values represent now the new error-proof class size variable (represented as FV below) that incorporates known variables such as the computed average class size based on rules about maximum class size, and other measured student, teacher/classroom or school variables. The regression model at the second stage is

$$Y_{ij} = \beta_0 + \beta_1 FV_{ij} + \boldsymbol{STD_{ij}B_2} + \boldsymbol{TCH_{ij}B_3} + \boldsymbol{SCH_jB_4} + e_{ij} \quad (4.3)$$

where Y_{ij} represents the outcome, *FV* is the modified class size variable and β is the coefficient of interest (i.e., the association between class size and achievement). All other terms have been defined previously. The residual term e includes a student and a school component, namely *e* = (student, school) and the variance of the school component captures the clustering effect due to complicated sampling in TIMSS. To account for the missing data effect, covariates with missing values were imputed with median values for the continuous variables and zero for the binary

variables. Most covariates have missing rates less than 5%. For example, the average missing rate for the class size variable is 3.5% across all data sets. Given the low missing rate of the main predictor and also with the purpose to make the analysis comparable among different countries and outcomes and also between full sample and RD sample, missing flags will not be included in the model.

In the full sample analysis, there are 58 data sets instead of 60 (i.e., four countries x three years x five subjects) because Slovenia does not have data on earth science in 2003 and 2007. For each dataset, there are one cognitive outcomes in the format of five plausible values and five non-cognitive outcomes. The weighted estimates and standard errors (SEs) for each PV will be computed and then combined across five PVs using multiple imputation formulae (Schafer, 1999). The weighted analysis would produce the main results to report as it represents the inference in a national population. Nevertheless, unweighted estimates will also be computed serving the purposes of sensitivity checking and also to be compared with findings in RD which is unweighted. I use the command ivregress in STATA to conduct the analyses.

4.5.2 RD

In addition to creating the instrument variable of predicted class size, the maximum class size rules give rise to a (fuzzy) RD design with an up-and-down or discontinuity pattern as school enrollment increases. For example, in this study, except in Slovenia, the maximum class size rule is 30. Once the enrollment passes cut-off points (e.g., 30, 60 and 90) for example by one unit, one more class will be added according the class size cap, so the average class sizes on the right side of the cut-off points become much smaller as 15.5, 20.3 and 22.7 in the first three segments while the counterparts in the left side are 29, 29.5 and 29.7. The main idea of RDD is

that the data around the cut-off points of multiples of the rule would resemble a local treatment effect of a randomized experiment when certain assumptions are met.

In a RDD, typically there is a running variable (Z) that determines the treatment status (T). Based on the magnitude of compliance, there are two types of design in general: sharp RD and fuzzy RD (G. W. Imbens & Lemieux, 2008). If Z fully captures the variation of T, it is a sharp RDD. Under this scenario, the estimates of T would be unbiased given Z is also included in the regression model. When Z cannot fully determine the T, this is a fuzzy RDD and the estimate of T could have bias even controlling Z in the model. Nevertheless, the IV method that is similar to "intention-to-treat analysis" could be embedded in the framework of RDD to get rid of some bias (Joshua D Angrist, Imbens, & Rubin, 1996). For RDD analysis, choosing an appropriate bandwidth is of great importance to balance the trade-off between bias and precision (G. Imbens & Kalyanaraman, 2012). On one hand, the narrower the bandwidth is the more accurate the estimates could be but there may not be sufficient sample to perform powerful statistical analysis. On the other hand, having wider bandwidth would have sufficient sample size but it is less accurate as it may not resemble local treatment any more. In the literature, one approach (i.e., optimal bandwidth) as well as corresponding software program in R or STATA have been developed to address this issue (Calonico, Cattaneo, Farrell, & Titiunik, 2017; Calonico, Cattaneo, & Titiunik, 2014a, 2014b, 2015).

In order to ensure a valid RD, it would be necessary to check that there is no discontinuity pattern in the running variable plus no other school resources would exhibit the same discontinuous pattern (Lee & Lemieux, 2010). Moreover, predetermined covariates should be balanced across observations just above and below the thresholds, so any jump or change in

the outcome would be due to the discontinuity in the class size change, which consequently becomes a causal inference of class size.

In the specific class size setting, the running variable is grade enrollment and the binary treatment is small vs large class size. The RDD of class size has several special characteristics. First, as some schools may not follow the maximum rule, it is a fuzzy RDD in which IV method could be employed. Second, it has multiple cutoffs with disproportional number of observations located around each cut-off point. The first and second segments may have more observations than other segments, but it is also possible that observations spread evenly across different segments. It is not uncommon to normalize the running variable, combine observations and produce an overall treatment effect across all cut-off points. In this study, the enrollment variable will be centered around each cut-off point, but it will not be standardized in order to maintain original segment variation. Third, the treatment status is unknown on the surface but could be generated with appropriate choice of bandwidth across different segments. Because of the limitation on creating treatment variable, optimal bandwidth does not work.

The statistical models are similar to the equations (4.2) and (4.3) in full sample IV analysis with the same covariates including the linear, quadratic and cubic terms of enrollment. Comparing with the IV analysis of full sample, the only difference is using RD sample with a dummy instrument of class size representing small class treatment instead of using the continuous predicted average class size as IV. It should be noted that reverse coding the dummy instrument yields the same results although the correlation between the dummy instrument of class size and teacher reported class size would have opposite sign. Using the RD graph of 4^{th} polynomial, it was found that 4^{th} power may overcorrect non-linearity. Therefore, I used cubic function as in the IV full sample analysis, which should appropriately capture possible non-

linearity across segments in general and thus the segment dummy variable was not included. The RD data generating details are explained in the result part.

4.6 Results

1) Descriptive

The sample size of data per county, per year and per subject is presented in Table 4.1. Lithuania and Slovenia have the smallest and largest number of schools, which is 141 and 186 respectively in 2011. The number of class ranges from 155 in Hungary in 2003 to 266 in Romania in 2007. The number of teacher varied among five subjects in addition to the variation in country and year. By and large, the number of teachers in specific sciences (i.e., physics, biology, chemistry and earth science) is similar, but is different from that in math. The smallest and largest number of students are 3301 in Hungary in 2003 and 5523 in Romania in 2011 respectively. On average, there are 4277 students, 190 teachers, 231 classrooms and 151 schools across countries, years and subjects.

Table 4.2 provides descriptive statistics on the outcome variables and class size. To be concise, the minimum and maximum values are not reported but will be briefly mentioned. With regard to the cognitive outcomes of achievement scores, comparing among countries, in general, Hungary has the highest average scores in each subjects, followed by Lithuania and Slovenia while Romania the lowest. The mean of the non-cognitive outcomes ranges between 2 and 3 and have a variation of 0.7 to 1 overall within the scale range of 1 to 4. The average class size across subjects in 2003 is about 23 in Hungary, 26 in Lithuania, 24 in Romania and 21 in Slovenia. In 2007, the numbers of average class size are 24 in Hungary, 26 in Lithuania, 22 in Romania and Slovenia. In 2011, numbers are similar as 23, 25, 24 and 22 in Hungary, Lithuania, Romania and

Slovenia respectively. The smallest average class size is about 15 in math class in Slovenia in 2011 and the largest one is about 27 in Lithuania in 2003 for Chemistry and Earth Science classes.

2) IV results

The model estimates of class size variable are rounded in two decimal places. The significant results at the level of 0.05 are reported first by year and then by country. Table 4.3 presents estimates in weighted analysis and negative sign means as class size unit increases the outcome decreases. In 2003, in Romania, there are some statistically significant negative estimates in academic achievement: math (-5.60), physics (-5.41), chemistry (-5.09), and earth science (-8.75). In 2007, in Romania, the non-cognitive variable of math enjoy has the significant estimate (-0.05). In Lithuania, the significant positive estimates were found in the variables of "biology well" (0.02), "biology quick" (0.03) and "biology weak" (0.03) and one significant negative estimate in "earth science quick" (-0.03). In 2011, in Lithuania, there are three significant negative estimates in "biology enjoy" (-0.03), "chemistry well" (-0.03) and "chemistry enjoy" (-0.04).

The unweighted estimates are included in Table 4.4, in which the significant results are reported as follows. In 2003, the significant estimate is in math score (-7.35) and chemistry score (-8.11) in Romania only. In 2007, significant results were found just in Lithuania: "biology quick" (0.03); "biology weak" (0.03); and "earth science quick" (-0.03). In 2011, three countries have significant estimates. In Lithuania, they are in "physics hard" (0.04), "biology enjoy" (-0.04) and "chemistry enjoy (-0.03). In Romania, "chemistry well" is significant (-0.05). In Slovenia, there are two significant positive estimates in earth science: "enjoy" (0.03) and "hard" (0.03).

To examine the strength of the instrumental variable of class size, the F-test values in the first-stage IV analysis are reported in Table 4.5. It needs to mention that for each subject, the results are pretty similar between each of the five plausible values and each of the five non-cognitive variables, so Table 4.5 only provides the analysis using the first plausible value in each subject. Concerning the aforementioned significant estimates, the F-test values are all above 10 except the earth science in Romania in 2003 with a value of 9.6 which is also close to 10. Therefore, the IV does not have the problem of weak instrument and the findings in full sample are all reliable. In addition, correlations between teacher reported class size in each subject and the predicted average class size based on maximum class size rules are above 0.5 overall, which further confirms that the class size instrument is valid and results are reliable.

3) RD data and results

The RD data were constructed according to the cut-off points based on maximum class size rule as well as the number of observations around each cut-off points, which vary by country and year. As the sample size in Table 4.1 and descriptive statistics in Table 4.2 show that there is very small variation regarding teacher variables including the class size variable among the data of the five subjects. To save text space, I used math data to demonstration RD data generating process as graphs and figures are similar across subjects. Figure 4.1 visually displays the discontinuity pattern using math data in four countries across three years. Specifically, the dots represent reported average class size in each school and the straight line is the computed average class size using maximum class size rules. How the dots align with the line indicates the degree of compliance with the class size cap regulation. Comparatively speaking, Lithuania and Romania follow rules better than Hungary and Slovenia.

In Figure 4.2 the histograms show the distributions of 8th grade enrollment and demonstrate the number of observations around each cut-off point in each country per year. For example, in Hungary 2011, most observations locate in schools that have grade enrollment less than 100, so the corresponding RD data would have three segments. Overall, there are observations spread around each cut-off points in each subgraph instead of having observations clustered on the left side of the discontinuity points. This suggests that it is unlikely that schools or parents manipulated school grade enrollment and class size.

Table 4.6 lists the RD sample details concerning the number of segments, bandwidth, sample size and a dummy treatment of class size. For instance, in Hungary, based on the information in Figure 4.1 and 4.2, I chose three segments and computed the largest possible bandwidth is 6 which allows to create a dummy instrument of class size treatment without overlapping. The segments thus are [25, 36], [55,66] and [85,96] and the cut-off points are 30, 60 and 90. On the left side of the cut-off points, the predicted average class size would be [25, 30], [27.5,30] and [28.3, 30] while on the right side it would be [15.5, 18], [20.3, 22] and [22.7, 24]. Across three segment, the predicted class size on the left side is [25, 30] and on the right side is [22.7, 24]. Therefore, the treatment variable of class size, a dummy instrument, can be created which treat=1 if ACS8<24.5 and zero otherwise. The "ACS8' represents the continuous instrument of class size based on the maximum class size rule that has been used in full sample analysis. Similarly, other bandwidth of 5, 4 and 3 can be chosen. The same procedure employs in other three countries. Across all four countries, the largest bandwidth was selected as long as the treatment variable of class size can be defined without overlapping between the left and right side of the cut-off points. The lowest bandwidth is 3, which followed Angrist and Lavy (1999).

In Figure 4.3, the RD plot considered the 4th power of polynomial function using the widest bandwidth data per country and per year. The enrollment was centered around each cut-off point. Observations are combined across segments.

Table 4.7 reports the RD results in 2003. It shows that the estimates are sensitive to the choice of bandwidth. In Lithuania with bandwidth of 3, in physics, significant estimates are "physics score" (2.29), "physics well" (-0.12), "physics quick" (-0.07), "physics enjoy" (-0.13) and "physics hard" (-0.07); in biology, including "biology well" (0.08), "biology quick" (0.11), "biology enjoy" (0.15), "biology hard" (0.05) and "biology weak" (0.12); in Chemistry for chemistry score (10.09) and in earth score (10.85).

Table 4.8 reports the RD results in 2007. In Lithuania, across the bandwidth of 4 and 3, consistent significant estimates are in earth: **"earth well" (-0.08 and -0.14), "earth hard" (-0.07 and -0.09)** and **"earth weak" (-0.11 and -0.13**). Other significant results were found with bandwidth 3: "physics enjoy" (0.06), **"biology well" (-0.03), "biology quick" (-0.05), "earth quick" (-0.14)** and **"earth enjoy" (-0.15)**. In Romania, with the bandwidth 4, significant estimates are "chemistry weak" (0.06), "earth quick" (0.11) and "earth hard" (0.10); with bandwidth 3, "chemistry quick" (0.07) and "earth score" (7.92).

Table 4.9 reports the RD results in 2011. The significant estimates are as follows: in Hungary with bandwidth 5 and 4, "math enjoy" (-0.06); in Lithuania with bandwidth 3, "earth score" (8.45), "earth well" (0.05), "earth hard" (0.11) and "earth weak" (0.10); in Romania with bandwidth 3, "earth score" (12.53).

Like the full sample analysis, I checked the strength of the dummy instrument by examining the F-test value in IV first stage (>=10) and absolute value of correlation between real class size and the dummy instrument (>0.4) and the indexes were provided in Table 4.10. It was

found that for the aforementioned significant findings in Table 4.7 to 4.9, significant estimates from valid instrument are highlighted with bolded format.

In addition, as endogenous sorting around discontinuity points may violate statistical assumptions of RD design, I checked the balance of covariates in the RD sample in 2007 where valid results were detected. Table 4.11 reports t-test results for the average differences of covariates between the groups on the right- and the left-side of the cut-off points in 2007. It is clear that covariate balance was achieved as there is no significant test difference in Lithuania with both bandwidth 4 and 3 in 2007. Since randomized control trial supposed to be balanced not only on the observed covariates but also on unobserved covariates, it may be helpful to check the balance of missing dummy flags in the RD analysis. However, given the low missing rate in the full sample and a further reduction of sample size in RD sample, overall it is not feasible to do the check.

4) Result summary and bias quantification

Although valid IV assumes to provide unbiased estimates, there still could be uncontrolled missing variable bias. To deal with it, this study further quantifies the percentage bias necessary to invalidate the inference for these statistically significant results. The formula is bias %= 1-(standard error X $t_{critical,df}$/coefficient) (where estimates of coefficient and standard error are generated by software while $t_{critical,df}$ would be close to 1.96 for a two-tailed t-test with 0.05 significance level when sample size is large) based on the method proposed by Frank et al (2013) (Frank, Maroulis, Duong, & Kelcey, 2013). This statistical method is superior to the statement about higher or lower statistical significance. According to Frank et al (2013), the median level of robustness is about 30% by rule of thumb for observational education studies.

Table 4.12 further summarizes previous valid statistically significant estimates as well as the bias quantification. To make the results comparable between the test score and the non-cognitive outcomes which have different measurement scale, I computed effect size as the estimate over the standard deviation of the dependent variable indicating the change of the outcome in standard deviation unit corresponds to one unit change in class size. It should be emphasized that unweighted standard deviation of the dependent variable was used for the weighted full sample estimates. It is because in single-level model analysis, the overall sampling weights were used to adjust for the standard errors which would not bring much change to the coefficient estimation. This usage of sampling weights in the overall regression model is different from weighted descriptive statistics such as computing weighted mean value.

I intend to report only the findings from the weighted full sample analysis and the unweighted RD analysis while findings from the unweighted full-sample analysis would be a reference. The bias quantification is included in parentheses alongside effect size plus the sign of coefficient in which negative sign indicates that smaller class generates positive outcome and positive sign means larger class size is preferred.

The results from weighted full sample analysis indicate that there are significant class size evidences in Romania and Lithuania. In Romania in 2003, increasing class size decreases students' academic scores in math (-0.06, 25%), physics (-0.07, 12%), chemistry (-0.05, 2%)and earth science (-0.09, 17%). Meanwhile larger class also decreases student's math learning enjoyment (-0.05, 8%) in 2007. In Lithuania in 2007, findings diverge. In 2007, it was found that larger class size increases student's non-cognitive functions in biology: well (0.03, 15%), quick (0.04, 32%), weak (0.04, 29%), but in earth science, in smaller class students learn quick (-0.03,

16%). In 2011, students in smaller-sized classroom are more likely to enjoy learning biology (-0.03, 20%), chemistry leaning well (-0.03, 30%) and enjoyment (-0.03, 30%).

Previous study indicates that schools and households may manipulate class size due to liberalized education market and violate the validity of RD estimates (Urquiola & Verhoogen, 2009). I checked the validity of RD via both visual graphs and statistic tests. The findings from RD analysis seem to resemble randomized experiment and thus have high internal validity. Results showed that in Lithuania in 2007 reducing class size significantly leads to more positive non-cognitive outcomes in biology well (0.04, 5%) and quick (0.06, 7%) and in earth science well (0.19, 67%), quick (0.16, 65%), enjoy (0.16, 80%), hard (0.10, 44%) and weak (0.15, 72%). The evidence in Lithuania in 2007 concerning some non-cognitive outcomes however is contradictory to the findings in IV.

4.7 Discussion

Over the past three decades, there is a large body of research examining the effects of class size particularly on student's academic performance, which have undergone continued debate. The U.S. singles out in providing the most pronounced evidence of small class size benefits in the early grades (Finn, 2002; Finn & Achilles, 1990, 1999; Finn, Gerber, & Boyd-Zaharias, 2005b; Krueger, 1999), which provided impetus for turning public's passion and appeal for small class size into the implementation of class size reduction in more than twenty states (Dee & West, 2011).

However, in general, evidences of large-scale data have shown that decreasing class size has little or no effect on lifting students' academic achievement. Consequently, some researchers argued that class size is not important on improving student academic achievement

and class size reduction failed in the cost-benefits tests, so investment on reducing class size is not worthwhile (Chingos, 2013). What has been missing in evaluating the class size effectiveness is the class size effects on non-cognitive skills. Small class size has been proposed to increase achievement by providing a better educational environment where it is more likely to reduce class disciplinary management, increase interaction and individualized teacher support such as the amount of individual attention pupil received, and promote students' engagement in learning (Blatchford, Bassett, & Brown, 2011; Blatchford, Moriarty, Edmonds, & Martin, 2002; Galton & Pell, 2012; Hargreaves, Galton, & Pell, 1998). In fact, the attention has been extensively placed on students' cognitive achievement, the non-cognitive variables are treated as mediating factors rather than outcomes. Although non-cognitive abilities have played a vital role on individuals' long-term academic, economic and life success, it is more difficult to measure precisely and the direct potential substantial benefits on non-cognitive development arising from being placed in a small-sized class has been underappreciated to a great extent.

I address this important literature void by investigating the class size effects on the non-cognitive outcomes of five subjects (i.e., math, physics, biology, chemistry and earth science) alongside cognitive outcomes in eighth grade across four European countries (i.e., Hungary, Lithuania, Romania and Slovenia) using TIMSS data in 2003, 2007 and 2011. The causal inference was facilitated by employing valid IV method based on the maximum class size rules in each country across the years using full sample and RD sample.

Full sample analyses have revealed that in Romania small class size boosts the enjoyment of learning math in 8^{th} grade with effect size of 0.047 and raises academic achievement in math, physics, chemistry and earth science with effect sizes ranging from 0.054 to 0.093. The finding of significant class size effects in Romania is in line with previous studies

(Li & Konstantopoulos, 2016; Shen & Konstantopoulos, 2017). Studies in the literature indicated that reducing class size (or student-teacher ratio) is one of the most effect way to improve student academic performance in secondary education in Romania (Kallai & Maniu, 2004).

In Lithuania, findings are not very consistent between two years. In 2011, reducing class size helps enhance the learning of biology and chemistry. However, in 2007, results have been mixed. The weighted full sample analysis shows that in larger class, students are more likely to report learning biology well, quickly and biology is less likely to be a weakness while the RD sample analysis demonstrates that students in smaller class is more likely to report learning biology well and quickly. Nevertheless, regard the subject of earth science, the finding is consistent which the small class has positive effect in learning the subject quickly from both the weighted full sample (0.033) and RD sample (0.155). Additionally, small class also has positive effect on learning earth science well (0.192), enjoy it (0.156) and it is less likely to feel earth science is hard (0.097) and a weakness (0.150). It is unclear whether the discrepancy is due to the effect of sampling weights, the methods between IV and RDD or because of the measurement error involved in self-report responses.

One interesting finding is that the analysis in RDD yields much stronger evidence than that in the full sample, which may imply the RDD is more compelling as it is closer to randomized experiment compared with other quasi-experiment approach (e.g., IV and DD) (Lee & Lemieux, 2010). Another interesting finding is that the class size effects were found in Romania and Lithuania but not in Hungary and Slovenia in which the former two countries have slower educational development than the latter two. This is perhaps due to the fact that class size effects are more likely to be detected in resource-limited countries with lower quality teachers, which was suggested by some previous studies (Altinok & Kingdon, 2012; Shen &

Konstantopoulos, 2017). More research is needed to verify whether it would be the case that low-quality teachers rely more on smaller class to promote effective teaching while high-quality teacher may deliver effective instruction regardless of the size of class. Another piece of information is that holding teacher's quality constant, class size effective would be more pronounced for low achieving students in low achieving classes (Bressoux, Kramarz, & Prost, 2009) or countries such as Colombia (Breton, 2014).

This study has several limitations. One limitation is that the measurement of the non-cognitive outcomes is much less precise than that of the standardized test scores in TIMSS. It is acknowledged that self-report response may threaten the measurement validity (Donaldson & Grant-Vallone, 2002). For instance, people may tend to report a better image than they actually are due to social desirability (van de Mortel, 2008). It is also possible that students may over or under estimates their subject learning ability and self-concept. A second possible limitation is about when enrollment information was collected. Angrist and Lavy (1999) pointed out that the grade enrollment information at the beginning of the school year may accurately predict average class size, but TIMSS data collect the assessment as well as school enrollment approaching the end of a school year, thus might introduce some bias. Third, I acknowledge that there might be potential heterogeneous effects across different segments which has not been investigated in this study given the great amount of RD analysis that would be involved. However, it would recommend to address the estimates of RDD with multiple cutoffs appropriately if the heterogeneity of effect would be the research interest (Cattaneo, Keele, Titiunik, & Vazquez-Bare, 2016).

To facilitate wise decision making pertaining to the policy of class size reduction, there is an urgent need for more studies of class size effectiveness on specific non-cognitive skills on

more specific aspects because non-cognitive outcome is a catch-all term that encompass all the aspects that have not been covered by standardized tests for measuring intelligence ability. More importantly, it would be imperative to provide more reliable measures on non-cognitive skills. For example, in addition to self-report, objective ratings on these traits by teachers or other experts based on daily behavior log or classroom video-recording might be useful.

For future research, when measurement on non-cognitive skills are valid, it would be interesting to explore whether the effect would be more salient in secondary education than in elementary education or vice versa. Moreover, longitudinal studies of class size would be informative as previous studies have showed that class size effect may be accumulative. Furthermore, with more waves of data, it would be interesting to model the trends of class size effects.

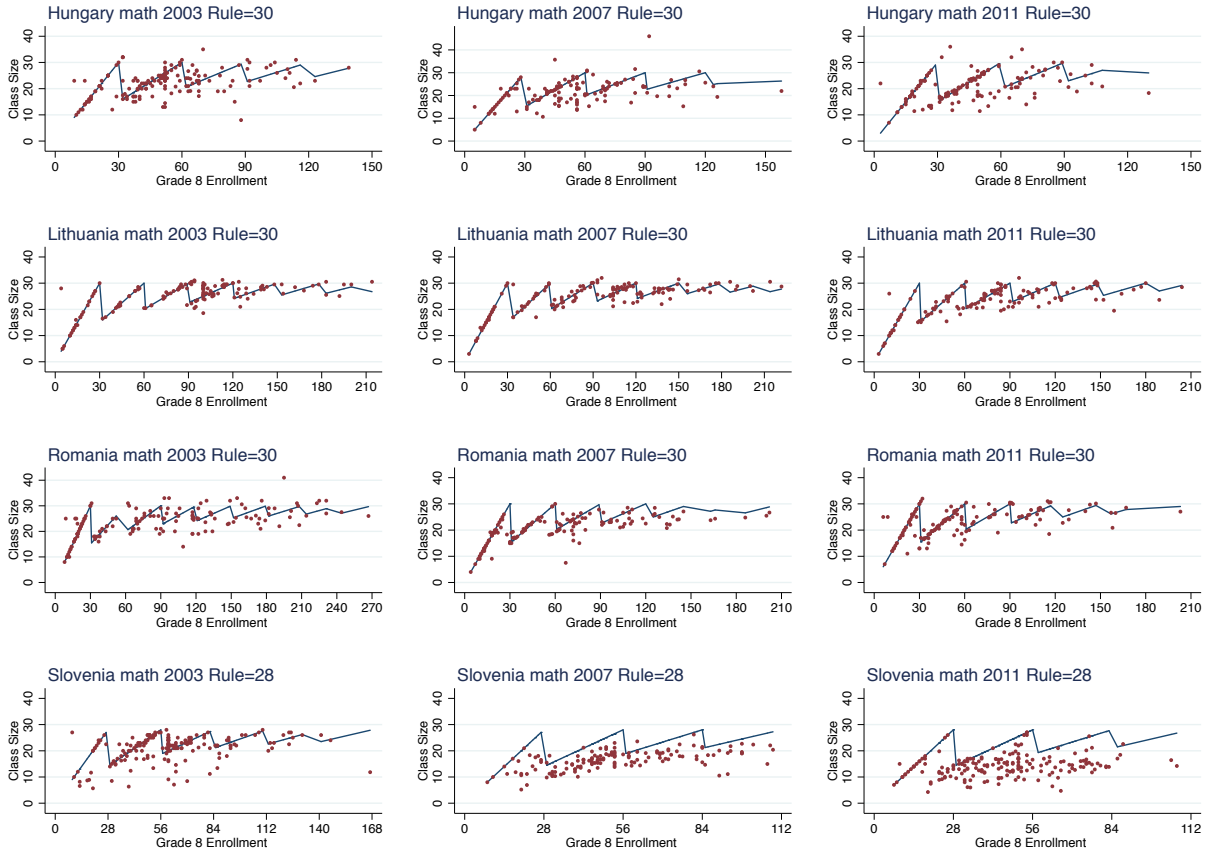
APPENDIX

APPENDIX 4: Variables list in TIMSS 8th Grade

TIMSS Variables	Descriptions	2003	2007	2011
Outcomes				
Math scores	A set of five overall math score plausible values	BSMMAT01-05	BSMMAT01-05	BSMMAT01-05
Do well in math	Usually do well in math (1 to 4 scale reverse coded)	BSBMTWEL	BS4MAWEL	BSBM16A
Learn quickly in math	Learn quickly in math (1 to 4 scale reverse coded)	BSBMTQKY	BS4MAQKY	BSBM16D
Enjoy learning math	I enjoy learning math (1 to 4 scale reverse coded)	BSBMTENJ	BS4MAENJ	BSBM14A
Math more difficult	Math is more difficult for you than for your classmates	BSBMTCLM	BS4MACLM	BSBM16B
Math is not a strength	Math is not one of your strength	BSBMTSTR	BS4MASTR	BSBM16C
Physics scores	A set of five overall physics score plausible values	BSSPHY01-05	BSSPHY01-05	BSSPHY01-05
Do well in Physics	Usually do well in Physics	BSBPTWEL	BS4PAWEL	BSBP32A
Learn quickly in Physics	Learn quickly in Physics	BSBPTQKY	BS4PAQKY	BSBP32D
Enjoy learning Physics	I enjoy learning Physics	BSBPTENJ	BS4PAENJ	BSBP30A
Physics more difficult	Physics is more difficult for you than for your classmates	BSBPTCLM	BS4PACLM	BSBP32B
Physics is not a strength	Physics is not one of your strength	BSBPTSTR	BS4PASTR	BSBP32C
Biology scores	A set of five overall Biology score plausible values	BSSLIS01-05	BSSBIO01-05	BSSBIO01-05
Do well in Biology	Usually do well in Biology	BSBBTWEL	BS4BAWEL	BSBB20A
Learn quickly in Biology	Learn quickly in Biology	BSBBTQKY	BS4BAQKY	BSBB20D
Enjoy learning Biology	I enjoy learning Biology	BSBBTENJ	BS4BAENJ	BSBB18A
Biology more difficult	Biology is more difficult for you than for your classmates	BSBBTCLM	BS4BACLM	BSBB20B
Biology is not a strength	Biology is not one of your strength	BSBBTSTR	BS4BASTR	BSBB20C
Chemistry scores	A set of five overall Chemistry score plausible values	BSSCHE01-05	BSSCHE01-05	BSSCHE01-05
Do well in Chemistry	Usually do well in Chemistry	BSBCTWEL	BS4CAWEL	BSBC28A
Learn quickly in Chemistry	Learn quickly in Chemistry	BSBCTQKY	BS4CAQKY	BSBC28D
Enjoy learning Chemistry	I enjoy learning Chemistry	BSBCTENJ	BS4CAENJ	BSBC26A
Chemistry more difficult	Chemistry is more difficult for you than for your classmates	BSBCTCLM	BS4CACLM	BSBC28B
Chemistry is not a strength	Chemistry is not one of your strength	BSBCTSTR	BS4CASTR	BSBC28C
Earth Science scores	A set of five overall Earth Science score plausible values	BSSEAS01-05	BSSEAR01-05	BSSEAR01-05
Do well in Earth Science	Usually do well in Earth Science	BSBETWEL	BS4EAWEL	BSBE24A
Learn quickly in Earth Science	Learn quickly in Earth Science	BSBETQKY	BS4EAQKY	BSBE24D
Enjoy learning Earth Science	I enjoy learning Earth Science	BSBETENJ	BS4EAENJ	BSBE22A
Earth Science more difficult	Earth Science is more difficult for you than for your classmates	BSBETCLM	BS4EACLM	BSBE24B
Earth Science is not a strength	Earth Science is not one of your strength	BSBETSTR	BS4EASTR	BSBE24C
Student Variables				
Female	Dummy (1=female student)	ITSEX	ITSEX	ITSEX
Age	Student age	BSDAGE	BSDAGE	BSDAGE
Home items	Sum of possession items at home	BSBGPS01-16	BS4GTH01-09	BSBG05A-K
Classroom/Teacher Variables				
Math class Size	Number of students in the math classroom	BTBMSTUD	BT4MSTUD	BTBG12
Science class Size	Number of students in each science classroom	BTBSSTUD	BT4SSTUD	BTBG12
Male teacher	Dummy variable for male teacher in classroom	BTBGSEX	BT4GSEX	BTBG02
Teacher education	Dummy variable of math teacher education (1= master degree and above)	BTBGFEDC	BT4GFEDC	BTBG04
Teacher experience	Math teachers teaching experience	BTBGTAUT	BT4GTAUT	BTBG01
School Characteristics				
Grade enrollment	8th grade enrollment	BCBGEENR	BC4GEENR	BCBG02
economically disadvantaged (ED percent)	The percentage of students coming from economically disadvantaged homes four categories	BCBGSBED	BC4GSBED	BCBG03A
Total Weight	Total weight at student level	TOTWGT	TOTWGT	TOTWGT

FIGURES

Figure 4.1: Class size by enrollment.



Note: Reported and computed average class size in mathematics by 8th grade enrollment in four countries across three years.

Note: The dots represent reported average class size in each school and the straight line is the computed average class size using maximum class size rules.

Figure 4.2: Histograms of 8th grade enrollment across four countries and three years.

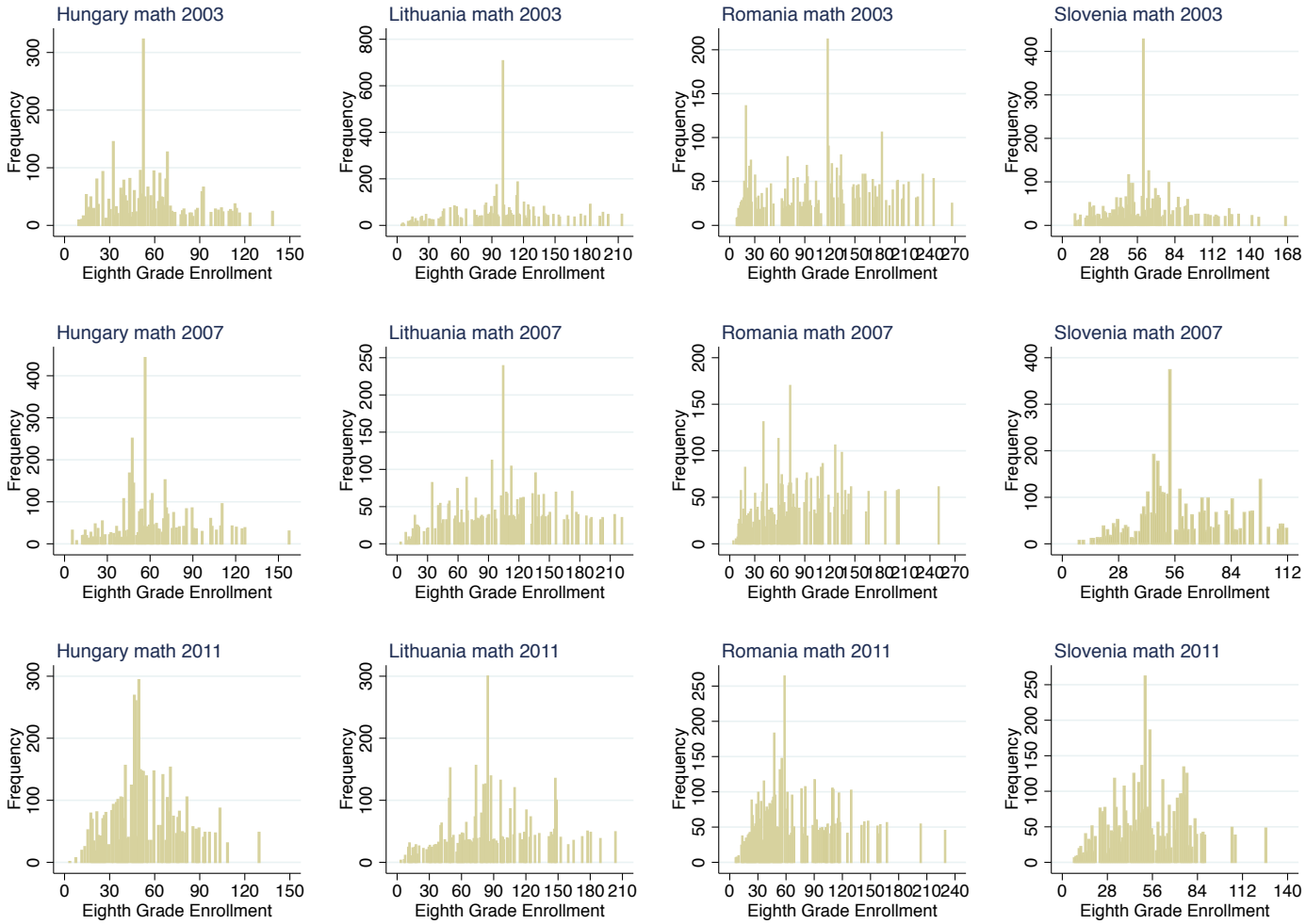
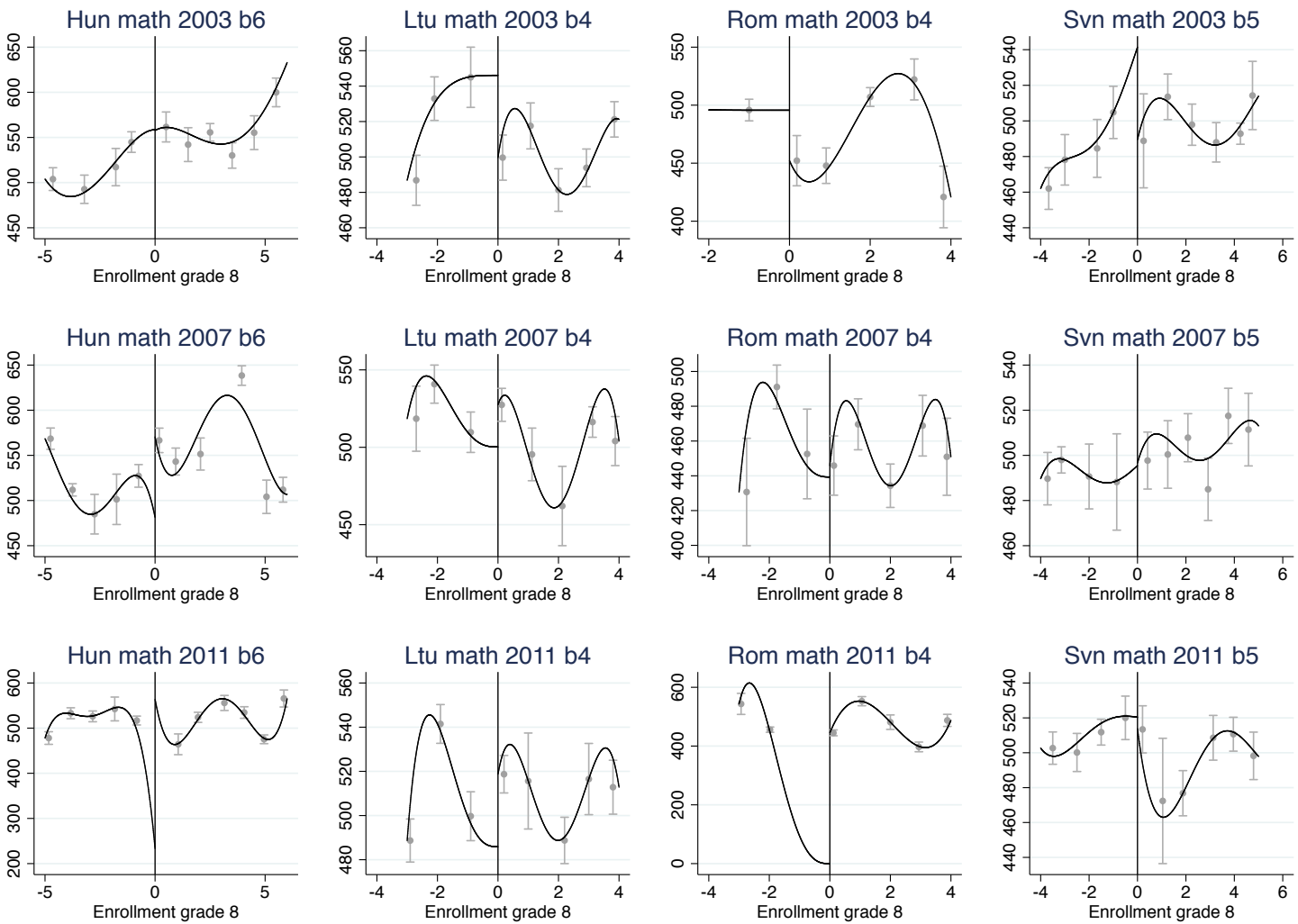


Figure 4.3: RD plot.



Note: The 4th polynomial using the widest bandwidth per country and per year.
The enrollment was centered around each cut-off point. Observations are combined across segments.

TABLES

Table 4.1: Fall sample size in 2003, 2007 and 2011

	2003					2007					2011				
	Math	Phy	Bio	Che	Earth	Math	Phy	Bio	Che	Earth	Math	Phy	Bio	Che	Earth
Hungry															
Student	3302	3301	3301	3301	3256	4111	4111	4059	4111	4109	5178	5178	5178	5156	5178
Teacher	198	158	158	156	154	263	172	179	165	183	242	168	171	157	172
Class	155	155	155	155	153	246	246	244	246	246	251	251	251	250	251
School	155	155	155	155	153	144	144	143	144	144	146	146	146	146	146
Lithuania															
Student	4964	4883	4689	4714	4943	3991	3991	3991	3991	3991	4747	4747	4747	4747	4743
Teacher	214	166	146	158	171	209	152	160	157	163	222	163	160	160	165
Class	258	254	242	240	257	258	258	258	258	258	258	258	258	258	257
School	143	141	133	129	142	142	142	142	142	142	141	141	141	141	141
Romania															
Student	4104	4104	4023	4080	4104	4198	4196	4166	4168	4189	5523	5523	5483	5523	5492
Teacher	178	179	175	180	178	236	189	179	170	188	221	172	176	170	172
Class	178	178	174	177	178	266	266	264	265	266	248	248	246	248	247
School	148	148	146	147	148	149	149	149	149	149	147	147	147	147	147
Slovenia															
Student	3578	3578	3578	3578	-	4043	4025	4043	4043	-	4415	4415	441	4413	4414
Teacher	228	176	176	176	-	459	170	172	157	-	478	194	201	193	196
Class	176	176	176	176	-	260	259	260	260	-	225	225	225	225	225
School	174	174	174	174	-	148	148	148	148	-	186	186	186	186	186

Note: Phy=Physics; Bio=Biology/Life science; Che=Chemistry; Earth= Earth science

Table 4.2: Unweighted descriptive statistics

Variables	2003								2007								2011							
	Hungary		Lithuania		Romania		Slovenia		Hungary		Lithuania		Romania		Slovenia		Hungary		Lithuania		Romania		Slovenia	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Outcomes																								
Math Score	532.45	80.40	507.22	77.00	480.10	88.92	492.41	71.12	522.94	85.18	511.21	79.03	472.25	96.86	503.50	71.13	513.47	85.25	508.87	77.36	469.31	103.83	504.70	70.43
Math_well	3.02	0.78	2.78	0.92	2.78	0.87	2.94	0.78	2.91	0.83	2.76	0.92	2.82	0.86	2.89	0.75	2.83	0.94	2.89	0.90	2.71	0.98	2.82	0.78
Math_quick	2.63	0.94	2.35	0.90	2.67	0.98	2.73	0.87	2.64	0.94	2.43	0.90	2.67	1.00	2.72	0.83	2.60	0.99	2.65	0.88	2.66	0.99	2.70	0.83
Math_enjoy	2.50	1.00	2.51	1.02	2.64	1.01	2.12	0.93	2.25	1.01	2.50	1.02	2.55	1.02	2.02	0.87	2.50	1.01	2.85	0.95	2.74	0.98	2.17	0.91
Math_hard	2.94	0.98	2.65	1.03	2.62	1.07	2.84	0.87	2.92	0.99	2.69	1.03	2.56	1.11	2.93	0.82	2.89	1.01	2.77	1.02	2.48	1.09	2.86	0.86
Math_weak	2.39	1.11	2.76	1.10	2.26	1.09	2.32	1.01	2.36	1.12	2.87	1.11	2.27	1.11	2.37	1.00	2.37	1.12	2.52	1.03	2.22	1.14	2.83	0.93
Physics Score	538.61	75.98	522.21	60.95	475.86	80.97	507.61	58.74	549.11	88.02	510.72	82.03	461.76	95.26	529.31	74.66	533.25	88.85	508.23	79.54	464.70	96.44	531.42	79.21
Physics_well	2.93	0.83	2.73	0.91	2.63	0.92	2.75	0.87	2.76	0.90	2.76	0.90	2.66	0.94	2.57	0.88	2.83	0.96	2.72	0.95	2.63	1.03	2.60	0.88
Physics_quick	2.63	0.93	2.29	0.90	2.52	1.00	2.51	0.91	2.54	0.96	2.36	0.90	2.52	1.01	2.31	0.93	2.65	0.99	2.45	0.93	2.59	1.02	2.39	0.89
Physics_enjoy	2.31	1.01	2.31	1.04	2.43	1.02	2.12	0.96	2.21	1.02	2.42	0.99	2.44	1.05	1.99	0.92	2.62	1.03	2.72	1.03	2.72	1.07	2.13	0.94
Physics_hard	2.99	0.90	2.73	0.96	2.64	0.99	2.78	0.89	2.89	0.94	2.75	0.95	2.62	1.03	2.72	0.90	3.08	0.92	2.71	1.02	2.77	1.05	2.76	0.92
Physics_weak	2.62	1.00	2.81	1.01	2.35	1.03	2.38	0.98	2.57	1.04	2.88	1.00	2.33	1.05	2.42	1.02	2.64	1.03	2.48	0.99	2.37	1.07	2.58	0.95
Biology Score	538.96	70.81	520.35	70.49	474.65	89.23	519.20	69.95	539.35	78.25	534.55	86.35	465.49	91.59	533.10	77.81	526.54	76.57	522.38	79.59	465.89	90.11	530.97	76.79
Biology_well	3.23	0.71	3.28	0.76	3.02	0.81	3.20	0.75	3.17	0.74	3.23	0.74	3.08	0.81	3.06	0.76	3.13	0.82	3.26	0.73	3.11	0.89	2.98	0.75
Biology_quick	2.95	0.88	2.80	0.89	3.03	0.93	3.03	0.83	2.93	0.87	2.77	0.87	2.99	0.95	2.85	0.86	2.96	0.90	2.92	0.82	3.13	0.92	2.82	0.81
Biology_enjoy	2.83	0.99	2.84	1.01	2.93	0.99	2.64	0.97	2.69	1.02	2.84	0.97	2.89	1.02	2.44	0.98	2.95	0.95	3.16	0.85	3.21	0.89	2.51	0.90
Biology_hard	3.16	0.86	3.07	0.95	2.93	1.00	3.08	0.79	3.18	0.86	3.07	0.92	2.98	0.99	3.09	0.82	3.20	0.85	3.14	0.88	3.07	1.00	3.07	0.76
Biology_weak	2.84	0.99	3.29	0.92	2.47	1.06	2.77	0.93	2.84	1.00	3.33	0.86	2.52	1.05	2.68	0.97	2.83	0.98	2.82	0.92	2.67	1.07	3.02	0.82
Chemistry Score	562.07	78.59	536.98	70.76	476.47	95.01	530.28	70.65	545.96	92.92	510.56	90.03	465.97	102.97	548.19	84.22	542.26	82.67	521.95	75.61	477.01	97.18	556.70	82.68
Chemistry_well	2.69	0.88	2.78	0.95	2.63	0.93	2.94	0.84	2.60	0.92	2.81	0.91	2.62	0.97	2.91	0.85	2.67	0.99	2.90	0.95	2.68	1.06	2.88	0.85
Chemistry_quick	2.41	0.94	2.31	0.94	2.52	1.02	2.71	0.90	2.38	0.96	2.40	0.93	2.48	1.04	2.69	0.93	2.48	0.99	2.56	0.94	2.61	1.06	2.70	0.88
Chemistry_enjoy	2.20	1.01	2.52	1.08	2.46	1.05	2.41	0.99	2.11	1.02	2.54	1.03	2.44	1.06	2.39	1.00	2.44	1.06	2.85	1.03	2.71	1.09	2.55	0.98
Chemistry_hard	2.79	0.95	2.74	1.00	2.66	1.01	2.92	0.87	2.77	0.95	2.79	0.97	2.63	1.05	2.98	0.87	2.97	0.95	2.85	1.00	2.77	1.08	2.98	0.86
Chemistry_weak	2.34	1.05	2.85	1.03	2.41	1.07	2.60	0.97	2.38	1.05	2.96	1.00	2.39	1.08	2.71	0.99	2.45	1.04	2.59	0.99	2.42	1.11	2.87	0.91
Earth Score	540.07	77.17	515.68	76.35	471.80	94.27	-	-	540.57	91.49	521.70	93.40	473.52	97.95	-	-	519.53	87.92	521.72	84.32	477.50	90.56	559.41	90.17
Earth_well	3.12	0.78	3.29	0.77	2.98	0.88	-	-	2.93	0.85	3.26	0.78	3.07	0.87	-	-	2.96	0.90	3.24	0.80	3.12	0.92	2.98	0.79
Earth_quick	2.83	0.90	2.80	0.90	2.99	0.97	-	-	2.66	0.94	2.80	0.90	3.00	0.99	-	-	2.76	0.96	2.88	0.88	3.10	0.94	2.86	0.82
Earth_enjoy	2.62	1.01	2.93	0.99	3.00	1.01	-	-	2.39	1.04	2.92	0.97	3.03	1.02	-	-	2.70	1.00	3.15	0.90	3.19	0.93	2.62	0.93
Earth_hard	3.08	0.90	3.04	0.96	2.78	1.03	-	-	3.01	0.91	3.09	0.94	2.89	1.05	-	-	3.14	0.88	3.09	0.94	3.05	1.01	3.10	0.78
Earth_weak	2.78	0.98	3.31	0.90	2.57	1.06	-	-	2.66	1.02	3.33	0.89	2.61	1.09	-	-	2.70	1.00	2.83	0.95	2.66	1.07	3.01	0.83
Class Size (CS)																								
Math_CS	23.05	5.52	26.41	4.17	24.98	4.96	21.32	5.29	22.51	6.18	26.13	4.27	22.40	4.78	17.31	5.22	22.09	5.58	24.78	4.63	24.28	5.00	14.98	5.46
Physics_CS	23.88	5.17	26.40	4.43	25.30	5.07	22.49	3.02	23.76	4.98	26.42	4.98	22.71	4.76	22.56	3.88	23.36	5.15	24.88	4.60	24.08	5.26	22.04	4.13
Biology_CS	23.94	5.04	26.36	4.00	25.37	5.34	22.59	3.08	24.02	5.52	26.26	5.25	22.43	4.46	22.68	3.85	23.57	5.34	24.89	4.61	24.41	4.91	22.36	4.92
Chemistry_CS	23.99	5.12	27.22	4.29	25.15	5.47	22.46	3.10	24.16	5.89	26.41	4.26	22.72	5.39	22.18	4.19	23.43	5.21	24.85	4.60	24.43	4.93	22.04	3.99
Earth_CS	24.29	5.08	27.22	4.29	24.88	5.00	-	-	23.94	5.30	26.15	4.14	22.66	5.42	-	-	23.50	5.19	24.85	4.63	24.52	5.75	21.97	4.07

Note: Each subject score is averaged across five plausible values. M=mean; SD=standard deviation. Earth=Earth science.

Table 4.3: Weighted estimates using IV method in full sample

Variables	2003								2007								2011							
	Hun		Ltu		Rom		Svn		Hun		Ltu		Rom		Svn		Hun		Ltu		Rom		Svn	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Outcomes																								
Math Score	-4.24	6.02	-1.78	1.20	**-5.60***	2.13	-2.73	1.51	-3.51	4.75	-0.52	1.47	-3.69	2.83	-1.11	2.26	3.45	3.16	0.83	1.22	0.62	3.65	0.34	2.26
Math_well	0.01	0.03	-0.01	0.01	0.01	0.01	-0.01	0.01	-0.02	0.03	0.00	0.01	-0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.01	-0.01	0.02	0.01	0.02
Math_quick	0.00	0.04	0.01	0.01	-0.01	0.02	-0.01	0.01	-0.02	0.03	-0.02	0.02	-0.03	0.02	0.03	0.02	0.00	0.02	0.01	0.01	0.00	0.02	0.00	0.02
Math_enjoy	-0.02	0.04	-0.01	0.01	-0.01	0.02	-0.01	0.02	-0.02	0.03	0.01	0.02	**-0.05***	0.02	0.00	0.02	0.00	0.02	0.00	0.01	-0.04	0.03	0.04	0.03
Math_hard	-0.05	0.05	-0.01	0.01	-0.02	0.02	-0.01	0.01	-0.02	0.03	0.00	0.01	0.01	0.02	0.04	0.02	0.01	0.02	0.00	0.01	0.00	0.02	-0.01	0.02
Math_weak	-0.01	0.04	0.01	0.02	-0.01	0.02	0.00	0.02	-0.04	0.04	-0.01	0.02	0.01	0.02	0.01	0.02	0.00	0.02	0.00	0.01	0.00	0.02	-0.01	0.02
Physics Score	-0.37	4.53	-1.03	1.20	**-5.41***	2.44	-1.82	1.12	-1.79	2.45	-0.55	1.72	-1.67	3.99	-0.67	1.59	2.45	2.65	0.63	1.49	-0.63	2.81	0.75	1.26
Physics_well	-0.05	0.05	0.03	0.02	0.02	0.02	0.01	0.02	-0.03	0.02	0.00	0.01	0.03	0.03	0.01	0.01	0.00	0.02	0.02	0.01	0.01	0.02	0.01	0.01
Physics_quick	-0.05	0.05	0.02	0.02	0.00	0.02	0.01	0.01	-0.02	0.02	-0.02	0.01	0.01	0.03	0.01	0.02	0.00	0.02	0.01	0.01	0.01	0.02	0.00	0.01
Physics_enjoy	-0.04	0.04	0.02	0.02	0.04	0.02	0.01	0.02	-0.02	0.02	-0.01	0.01	-0.02	0.03	0.01	0.02	0.01	0.03	0.02	0.01	0.01	0.02	0.01	0.01
Physics_hard	-0.05	0.04	0.00	0.01	-0.02	0.02	0.01	0.01	-0.01	0.02	-0.01	0.01	0.01	0.02	0.03	0.01	-0.02	0.02	0.02	0.01	0.01	0.01	-0.01	0.01
Physics_weak	-0.04	0.04	0.03	0.02	0.02	0.02	0.01	0.01	-0.03	0.02	0.00	0.01	-0.01	0.02	0.01	0.01	-0.01	0.02	0.02	0.01	0.00	0.01	0.00	0.01
Biology Score	-1.88	4.38	-0.04	1.88	-5.38	2.87	-1.87	1.29	-1.38	2.06	0.06	1.42	-0.17	3.97	-1.54	1.43	1.36	2.67	0.50	1.35	-0.63	2.63	-0.28	1.33
Biology_well	0.00	0.03	0.00	0.03	0.01	0.01	0.00	0.01	0.01	0.02	**0.02***	0.01	0.02	0.02	-0.01	0.01	0.02	0.03	-0.01	0.01	-0.01	0.02	0.01	0.01
Biology_quick	0.00	0.03	-0.02	0.03	0.01	0.02	-0.01	0.01	0.02	0.02	**0.03***	0.01	0.01	0.03	-0.02	0.02	0.01	0.02	-0.01	0.01	0.01	0.01	0.01	0.01
Biology_enjoy	0.03	0.04	0.01	0.04	0.03	0.02	0.00	0.02	0.02	0.02	0.03	0.02	0.04	0.03	-0.03	0.02	-0.02	0.03	**-0.03***	0.01	0.00	0.02	0.01	0.02
Biology_hard	0.02	0.03	0.00	0.02	-0.04	0.02	-0.02	0.01	0.00	0.02	0.01	0.01	0.01	0.03	-0.01	0.01	0.00	0.02	0.00	0.01	0.01	0.02	0.01	0.02
Biology_weak	-0.01	0.04	0.00	0.03	-0.01	0.02	-0.01	0.01	0.01	0.02	**0.03***	0.01	0.00	0.02	-0.02	0.02	0.01	0.03	-0.01	0.01	-0.03	0.02	0.01	0.01
Chemistry Score	-5.42	7.93	-0.21	1.14	**-5.09***	2.54	-2.32	1.33	-1.99	3.03	-0.13	2.10	-0.25	3.69	-1.60	1.97	1.87	3.32	0.01	1.04	-0.94	3.23	0.34	1.39
Chemistry_well	-0.03	0.06	0.02	0.01	0.01	0.02	-0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.03	-0.01	0.02	0.04	0.03	**-0.03***	0.01	-0.04	0.03	0.00	0.01
Chemistry_quick	-0.07	0.08	0.01	0.01	-0.01	0.02	-0.01	0.01	0.03	0.02	0.01	0.02	0.00	0.03	0.00	0.02	0.02	0.03	-0.01	0.01	-0.02	0.03	0.00	0.02
Chemistry_enjoy	-0.05	0.07	0.01	0.01	0.00	0.02	-0.01	0.02	0.03	0.03	0.01	0.02	-0.01	0.03	-0.01	0.02	0.01	0.03	**-0.04***	0.01	-0.02	0.03	0.00	0.02
Chemistry_hard	-0.01	0.06	0.00	0.01	-0.01	0.01	-0.01	0.01	0.02	0.02	0.02	0.01	0.00	0.02	0.01	0.02	0.02	0.02	-0.01	0.01	-0.01	0.02	0.00	0.01
Chemistry_weak	0.02	0.06	0.01	0.01	0.01	0.01	0.00	0.01	0.02	0.02	0.02	0.02	0.01	0.02	0.00	0.02	0.03	0.03	0.00	0.01	-0.01	0.01	0.00	0.01
EarthScience Score	-3.92	5.44	-1.35	1.96	**-8.75***	3.70	-	-	-2.89	3.22	0.01	1.62	-4.06	4.99	-	-	2.18	3.97	-0.11	1.27	-0.19	2.29	-0.81	1.48
EarthScience_well	-0.03	0.05	0.00	0.01	0.00	0.02	-	-	-0.01	0.02	-0.01	0.01	-0.01	0.03	-	-	-0.03	0.04	-0.01	0.01	0.02	0.02	0.02	0.01
EarthScience_quick	-0.04	0.06	-0.01	0.02	0.02	0.02	-	-	0.00	0.02	**-0.03***	0.01	-0.05	0.04	-	-	-0.02	0.04	-0.01	0.01	0.01	0.02	0.02	0.01
EarthScience_enjoy	-0.02	0.06	0.01	0.02	0.01	0.02	-	-	0.00	0.03	-0.02	0.02	-0.07	0.04	-	-	-0.01	0.05	-0.02	0.01	0.01	0.02	0.03	0.02
EarthScience_hard	-0.10	0.07	0.01	0.01	-0.03	0.02	-	-	0.00	0.02	-0.01	0.01	0.03	0.03	-	-	0.01	0.03	-0.02	0.01	-0.01	0.02	0.00	0.01
EarthScience_weak	est	0.04	0.02	0.02	-0.04	0.02	-	-	0.00	0.02	-0.01	0.01	0.04	0.03	-	-	-0.02	0.04	-0.02	0.01	-0.01	0.02	0.01	0.01

Note: * p<0.05. Subject score combined estimates across five plausible values. Hun=Hungary; Ltu=Lithuania; Rom=Romania; Svn=Slovenia

Table 4.4: Unweighted estimates using IV method in full sample

Variables	2003								2007								2011							
	Hun		Ltu		Rom		Svn		Hun		Ltu		Rom		Svn		Hun		Ltu		Rom		Svn	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Outcomes																								
Math Score	-4.93	6.37	-1.25	1.52	**-7.35***	3.40	-2.35	1.42	-5.99	9.17	0.51	1.55	-1.65	2.36	-1.22	2.08	4.62	2.81	1.37	1.38	0.21	3.60	0.98	2.31
Math_well	0.00	0.03	-0.02	0.01	0.01	0.02	-0.01	0.01	-0.01	0.03	0.01	0.01	-0.01	0.01	0.02	0.02	0.02	0.02	0.00	0.01	-0.01	0.02	0.00	0.02
Math_quick	0.00	0.04	0.01	0.01	0.01	0.02	-0.01	0.01	-0.02	0.04	-0.01	0.01	-0.02	0.02	0.03	0.02	0.01	0.01	0.02	0.01	-0.01	0.02	0.00	0.02
Math_enjoy	0.00	0.04	-0.01	0.02	0.00	0.03	-0.02	0.02	-0.01	0.05	0.02	0.02	-0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.02	-0.03	0.02	0.06	0.04
Math_hard	-0.05	0.05	-0.01	0.01	-0.02	0.03	-0.02	0.01	-0.02	0.03	0.00	0.01	0.02	0.02	0.03	0.02	0.01	0.01	0.00	0.01	0.01	0.02	-0.01	0.02
Math_weak	-0.01	0.04	0.01	0.01	0.00	0.02	-0.01	0.01	-0.06	0.07	0.00	0.02	0.01	0.01	0.02	0.02	0.01	0.02	0.00	0.01	0.00	0.02	0.01	0.02
Physics Score	-1.14	5.30	-0.27	1.40	-7.85	4.24	-1.60	1.17	-1.25	3.35	0.86	1.73	0.27	3.10	-1.36	1.55	3.53	2.41	1.51	1.57	-1.22	2.33	0.64	1.04
Physics_well	-0.04	0.05	0.02	0.02	0.02	0.03	0.02	0.02	-0.02	0.03	0.00	0.02	0.03	0.03	0.01	0.01	0.00	0.03	0.03	0.02	0.00	0.02	0.01	0.01
Physics_quick	-0.05	0.05	0.02	0.02	-0.01	0.04	0.02	0.01	-0.02	0.02	-0.01	0.02	0.03	0.03	0.00	0.01	0.01	0.03	0.02	0.01	0.00	0.02	0.00	0.01
Physics_enjoy	-0.05	0.05	0.03	0.03	0.06	0.04	0.03	0.02	-0.01	0.03	-0.01	0.02	0.01	0.03	0.01	0.02	0.01	0.03	0.03	0.02	0.01	0.02	0.01	0.01
Physics_hard	-0.06	0.05	0.00	0.02	-0.04	0.03	0.01	0.01	0.00	0.02	-0.01	0.01	0.01	0.02	0.02	0.01	-0.02	0.02	**0.04***	0.02	0.02	0.01	0.00	0.01
Physics_weak	-0.03	0.04	0.03	0.02	0.04	0.03	0.02	0.01	-0.01	0.02	0.00	0.01	-0.03	0.02	0.00	0.01	-0.01	0.02	0.03	0.02	0.01	0.01	0.00	0.01
Biology Score	-2.73	4.85	1.17	2.07	-8.26	4.64	-1.96	1.67	-1.78	3.82	1.21	1.53	2.67	3.63	-1.71	1.49	2.43	2.54	1.36	1.41	-0.63	2.49	0.34	1.14
Biology_well	0.02	0.03	0.00	0.03	0.01	0.02	0.00	0.02	0.02	0.02	0.02	0.01	0.04	0.03	-0.02	0.02	0.03	0.03	-0.01	0.01	0.00	0.01	0.01	0.01
Biology_quick	0.02	0.03	-0.01	0.03	0.01	0.02	0.00	0.02	0.03	0.02	**0.03***	0.01	0.05	0.04	-0.02	0.02	0.02	0.02	-0.02	0.01	0.01	0.02	0.02	0.01
Biology_enjoy	0.05	0.05	0.04	0.04	0.05	0.03	0.00	0.02	0.03	0.03	0.02	0.02	0.06	0.04	-0.03	0.02	-0.01	0.03	**-0.04***	0.02	0.00	0.02	0.01	0.02
Biology_hard	0.03	0.03	-0.02	0.03	-0.07	0.04	-0.02	0.01	0.01	0.02	0.01	0.01	0.03	0.03	-0.02	0.02	0.01	0.02	0.00	0.01	0.02	0.01	0.01	0.01
Biology_weak	0.01	0.04	0.03	0.03	-0.03	0.03	0.00	0.02	0.02	0.03	**0.03***	0.01	0.01	0.02	-0.02	0.02	0.01	0.03	-0.01	0.01	-0.01	0.01	0.01	0.01
Chemistry Score	-9.85	16.84	0.02	1.44	**-8.11***	4.03	-2.28	1.64	-1.88	4.37	1.63	1.83	1.24	2.90	-2.19	2.43	3.13	2.72	0.68	1.20	-1.40	2.73	0.35	1.12
Chemistry_well	-0.05	0.11	0.01	0.02	-0.01	0.03	-0.02	0.02	0.00	0.03	0.03	0.02	0.01	0.02	0.00	0.02	0.05	0.04	-0.02	0.01	**-0.05***	0.02	0.00	0.01
Chemistry_quick	-0.10	0.15	0.00	0.02	-0.03	0.03	-0.01	0.02	0.00	0.03	0.00	0.02	0.00	0.02	0.00	0.02	0.04	0.03	-0.01	0.01	-0.03	0.02	0.00	0.01
Chemistry_enjoy	-0.11	0.16	0.00	0.02	-0.02	0.03	-0.01	0.02	0.00	0.03	0.01	0.02	0.00	0.02	-0.02	0.03	0.02	0.04	**-0.03***	0.02	-0.04	0.03	0.00	0.02
Chemistry_hard	-0.11	0.16	0.01	0.01	-0.02	0.02	-0.01	0.01	0.01	0.02	0.01	0.02	-0.01	0.02	0.02	0.02	0.03	0.02	0.00	0.01	-0.01	0.02	0.00	0.01
Chemistry_weak	0.00	0.08	0.01	0.01	0.01	0.02	0.00	0.01	0.01	0.03	0.02	0.02	0.00	0.02	0.00	0.02	0.05	0.03	0.01	0.01	-0.01	0.02	0.00	0.01
EarthScience Score	-5.97	7.24	-0.28	2.21	-14.36	8.81	-	-	-3.46	6.16	0.87	1.67	-0.15	4.15	-	-	3.10	3.05	0.74	1.38	-0.74	2.30	-0.29	1.28
EarthScience_well	-0.01	0.05	0.00	0.02	-0.01	0.03	-	-	0.00	0.04	-0.02	0.01	0.03	0.03	-	-	-0.02	0.03	-0.01	0.01	0.01	0.02	0.02	0.01
EarthScience_quick	-0.01	0.05	0.00	0.02	0.02	0.04	-	-	0.02	0.04	**-0.03***	0.02	0.01	0.04	-	-	-0.01	0.03	-0.01	0.01	-0.01	0.02	0.01	0.01
EarthScience_enjoy	0.01	0.06	0.02	0.03	0.02	0.04	-	-	0.02	0.05	-0.02	0.02	0.00	0.04	-	-	-0.01	0.04	-0.01	0.02	0.01	0.02	**0.03***	0.02
EarthScience_hard	-0.10	0.08	0.02	0.03	-0.06	0.05	-	-	0.00	0.03	-0.02	0.01	0.04	0.03	-	-	0.01	0.02	0.00	0.01	0.01	0.01	0.00	0.01
EarthScience_weak	0.00	0.04	0.02	0.03	-0.07	0.06	-	-	0.01	0.04	-0.01	0.01	0.04	0.03	-	-	-0.01	0.03	0.00	0.01	0.00	0.01	0.01	0.01

Note: * p<0.05. Subject score combined estimates across five plausible values. Hun=Hungary; Ltu=Lithuania; Rom=Romania; Svn=Slovenia

Table 4.5: IV first stage F-test values and class size correlation in full sample

Weighted	Hun		Ltu		Rom		Svn	
	1st stage	Corr	1st stage	Corr	1st stage	Corr	1st stage	Corr
2003								
Math	2.3	0.4	**30.9**	0.8	**21.8**	0.6	**20.3**	0.5
Physics	3.0	0.5	**21.9**	0.7	**12.4**	0.6	**32.5**	0.6
Biology	3.2	0.4	**11.4**	0.7	**12.8**	0.6	**26.5**	0.6
Chemistry	1.4	0.4	**22.0**	0.6	**27.5**	0.7	**25.9**	0.6
Earth Science	2.9	0.4	**14.7**	0.7	9.6	0.6	-	-
2007								
Math	3.5	0.4	**76.3**	0.8	**20.8**	0.7	**25.7**	0.4
Physics	**16.1**	0.6	**35.7**	0.7	6.4	0.6	**27.4**	0.6
Biology	**13.5**	0.5	**38.1**	0.7	5.9	0.6	**48.0**	0.6
Chemistry	**11.0**	0.6	**27.2**	0.7	7.5	0.6	**14.7**	0.4
Earth Science	9.8	0.6	**42.6**	0.8	3.8	0.6	-	-
2011								
Math	5.5	0.5	**31.8**	0.8	6.8	0.5	5.5	0.2
Physics	8.8	0.5	**26.8**	0.8	**10.0**	0.5	**27.1**	0.6
Biology	4.4	0.5	**36.8**	0.8	9.3	0.5	**21.1**	0.5
Chemistry	4.5	0.5	**40.5**	0.8	9.0	0.6	**25.2**	0.7
Earth Science	3.1	0.4	**28.7**	0.8	7.6	0.5	**30.5**	0.7

Unweighted	Hun		Ltu		Rom		Svn	
	1st stage	Corr	1st stage	Corr	1st stage	Corr	1st stage	Corr
2003								
Math	2.4	0.3	**25.8**	0.7	8.7	0.5	**17.8**	0.4
Physics	2.3	0.4	**20.0**	0.6	5.3	0.5	**48.6**	0.6
Biology	2.8	0.4	**10.8**	0.6	6.1	0.5	**22.3**	0.5
Chemistry	0.5	0.3	**20.4**	0.5	9.4	0.5	**21.6**	0.5
Earth Science	2.2	0.3	**13.8**	0.6	2.9	0.5	-	-
2007								
Math	1.2	0.3	**37.2**	0.7	**16.5**	0.6	**23.9**	0.3
Physics	8.9	0.4	**25.2**	0.6	7.0	0.6	**26.3**	0.5
Biology	4.2	0.4	**30.4**	0.6	5.0	0.5	**32.3**	0.5
Chemistry	5.4	0.4	**24.6**	0.6	8.2	0.5	7.5	0.3
Earth Science	2.6	0.4	**32.2**	0.7	3.4	0.4	-	-
2011								
Math	6.9	0.4	**20.8**	0.7	**10.3**	0.5	5.4	0.2
Physics	9.4	0.4	**16.9**	0.6	**16.0**	0.5	**43.5**	0.6
Biology	4.8	0.4	**19.9**	0.7	**12.4**	0.5	**29.1**	0.5
Chemistry	5.7	0.4	**21.4**	0.7	**16.1**	0.5	**41.8**	0.7
Earth Science	4.5	0.4	**21.6**	0.7	7.9	0.4	**42.5**	0.7

Note: corr=correlation between teacher reported class size and predicted average class size based on maximum class size rule

Table 4.6: RD data details

Country	2003					2007					2011				
		Enrollment Grade 8		Average class size		Enrollment Grade 8		Average class size			Enrollment Grade 8		Average class size		
Hungary															
Segments	Three	Left	Right	Left	Right	Left	Right	Left	Right		Left	Right	Left	Right	
Segment1		[25, 30]	[31, 36]	[25,30]	[15.5,18]	[25, 30]	[31, 36]	[25,30]	[15.5,18]		[25, 30]	[31, 36]	[25,30]	[15.5,18]	
Segment2		[55,60]	[61, 66]	[27.5,30]	[20.3,22]	[55,60]	[61, 66]	[27.5,30]	[20.3,22]		[55,60]	[61, 66]	[27.5,30]	[20.3,22]	
Segment3		[85,90]	[91,96]	[28.3,30]	[22.7,24]	[85,90]	[91,96]	[28.3,30]	[22.7,24]		[85,90]	[91,96]	[28.3,30]	[22.7,24]	
Bandwidth	6			Left	Right			Left	Right				Left	Right	
Sample size (N)				511	600			983	503				687	804	
Small class treatment		**Treat=1 (acs8<24.5)**		[25,30]	[15.5,24]	**Treat=1 (acs8<24.5)**		[25,30]	[15.5,24]		**Treat=1 (acs8<24.5)**		[25,30]	[15.5,24]	
Bandwidth	**5**			Left	Right			Left	Right				Left	Right	
sample size (N)				337	566			844	399				525	720	
Small class treatment		**Treat=1 (acs8<24.5)**		[26,30]	[15.5,23.8]	**Treat=1 (acs8<24.5)**		[26,30]	[15.5,23.8]		**Treat=1 (acs8<24.5)**		[26,30]	[15.5,23.8]	
Bandwidth	**4**			Left	Right			Left	Right				Left	Right	
Sample size (N)				337	498			347	349				450	478	
Small class treatment		**Treat=1 (acs8<24.5)**		[27,30]	[15.5,23.5]	**Treat=1 (acs8<24.5)**		[27,30]	[15.5,23.5]		**Treat=1 (acs8<24.5)**		[27,30]	[15.5,23.5]	
Bandwidth	**3**			Left	Right			Left	Right				Left	Right	
Sample size (N)				248	417			289	277				282	323	
Small class treatment		**Treat=1 (acs8<24.5)**		[28,30]	(15.5,23.3]	**Treat=1 (acs8<24.5)**		[28,30]	(15.5,23.3]		**Treat=1 (acs8<24.5)**		[28,30]	(15.5,23.3]	
Lithuania				Average class size				Average class size					Average class size		
Segments	Six	6[30-180]				6[30-180]					6[30-180]				
Bandwidth	4			Left	Right			Left	Right				Left	Right	
Sample size (N)				464	746			480	497				764	681	
Small class treatment		**Treat=1 (acs8<27)**		[27,30]	[15.5,26.3]	**Treat=1 (acs8<27)**		[27,30]	[15.5,26.3]		**Treat=1 (acs8<27)**		[27,30]	[15.5,26.3]	
Bandwidth	3			Left	Right			Left	Right				Left	Right	
Sample size (N)				291	513			345	429				428	417	
Small class treatment		**Treat=1 (acs8<27)**		[28,30]	[15.5,26.1]	**Treat=1 (acs8<27)**		[28,30]	[15.5,26.1]		**Treat=1 (acs8<27)**		[28,30]	[15.5,26.1]	
Romania				Average class size				Average class size					Average class size		
Segments	Seven	7[30-210]				Four	4[30-120]				Five	5[30-150]			
Bandwidth	4			Left	Right			Left	Right				Left	Right	
Sample size (N)				361	1006			385	501				743	624	
Small class treatment		**Treat=1 (acs8<27)**		[27,30]	[15.5,26.8]	**Treat=1 (acs8<26)**		[27,30]	[15.5,24.8]		**Treat=1 (acs8<26)**		[27,30]	[15.5,25.7]	
Bandwidth	3			Left	Right			Left	Right				Left	Right	
Sample size (N)				225	679			308	494				689	434	
Small class treatment		**Treat=1 (acs8<27)**		[28,30]	[15.5,26.6]	**Treat=1 (acs8<26)**		[28,30]	[15.5,24.6]		**Treat=1 (acs8<26)**		[28,30]	[15.5,25.5]	
Slovenia				Average class size				Average class size					Average class size		
Segments	Four	4[28-112]				Three	3[28-84]				Three	3[28-84]			
Bandwidth	5			Left	Right			Left	Right				Left	Right	
Sample size (N)				369	925			884	539				827	564	
Small class treatment		**Treat=1 (acs8<24)**		[24,28]	[14.5,23.4]	**Treat=1 (acs8<24)**		[24,28]	[14.5,22.5]		**Treat=1 (acs8<24)**		[24,28]	[14.5,22.5]	
Bandwidth	4			Left	Right			Left	Right				Left	Right	
Sample size (N)				227	874			761	454				660	451	
Small class treatment		**Treat=1 (acs8<24)**		[25,28]	[14.5,23.2]	**Treat=1 (acs8<24)**		[25,28]	[14.5,22]		**Treat=1 (acs8<24)**		[25,28]	[14.5,22]	
Bandwidth	3			Left	Right			Left	Right				Left	Right	
Sample size (N)				151	399			274	356				512	259	
Small class treatment		**Treat=1 (acs8<24)**		[26,28]	[14.5,23]	**Treat=1 (acs8<24)**		[26,28]	[14.5,21.6]		**Treat=1 (acs8<24)**		[26,28]	[14.5,21.6]	

Note: acs8=predicted average class size based on maximum class size rule (the instrument of class size used in the full sample analysis).

Table 4.7: RD results in 2003

	Hungary								Lithuania				Romania				Slovenia					
Bandwidth	6		5		4		3		4		3		4		3		5		4		3	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Math Score	4.12	8.30	-0.05	5.79	-2.94	7.26	-13.81	26.81	3.17	5.29	3.27	2.90	2.96	4.99	-0.98	6.81	-0.88	2.70	-5.35	8.04	-2.73	3.80
Math_well	-0.03	0.09	-0.02	0.06	-0.02	0.06	-0.13	0.30	-0.13	0.11	0.04	0.03	0.01	0.03	0.01	0.05	0.02	0.02	0.02	0.04	0.00	0.03
Math_quick	-0.01	0.08	-0.03	0.08	-0.04	0.10	-0.20	0.42	0.00	0.05	0.00	0.03	-0.01	0.04	-0.02	0.06	0.01	0.02	0.02	0.04	0.03	0.02
Math_enjoy	-0.02	0.07	-0.04	0.08	0.00	0.05	-0.15	0.34	-0.07	0.09	0.04	0.05	-0.06	0.07	-0.04	0.09	0.04	0.04	0.06	0.07	0.06	0.04
Math_hard	-0.03	0.06	-0.01	0.04	-0.05	0.08	-0.07	0.15	0.04	0.07	0.00	0.02	0.02	0.05	-0.01	0.07	0.00	0.02	0.02	0.04	0.01	0.03
Math_weak	-0.13	0.21	-0.10	0.14	-0.10	0.16	-0.26	0.54	-0.03	0.04	0.01	0.03	-0.06	0.05	-0.10	0.09	0.01	0.02	-0.04	0.05	0.03	0.03
Physics Score	18.08	51.42	-5.74	18.80	-7.79	25.40	-3.34	11.14	3.11	3.69	**2.29***	0.95	12.20	13.04	8.65	14.93	-0.47	2.84	-4.02	6.63	-7.63	14.68
Physics_well	-0.14	0.42	0.16	0.30	0.20	0.42	0.03	0.06	0.02	0.06	**-0.12***	0.04	0.18	0.26	0.38	0.79	0.01	0.04	0.04	0.08	0.02	0.15
Physics_quick	-0.19	0.55	0.21	0.37	0.27	0.55	0.00	0.06	0.03	0.04	**-0.07***	0.03	0.19	0.22	0.24	0.42	0.00	0.03	0.04	0.07	0.16	0.23
Physics_enjoy	-0.21	0.52	0.27	0.53	0.40	0.92	0.09	0.10	-0.01	0.07	**-0.13***	0.06	0.18	0.22	0.36	0.72	0.01	0.03	0.10	0.10	0.19	0.21
Physics_hard	-0.01	0.13	0.14	0.25	0.13	0.27	-0.01	0.07	-0.01	0.03	**-0.07***	0.02	0.04	0.06	0.01	0.05	0.04	0.03	0.10	0.09	0.06	0.10
Physics_weak	-0.11	0.33	0.22	0.43	0.23	0.48	0.07	0.08	0.05	0.06	-0.06	0.03	-0.03	0.07	0.02	0.09	0.04	0.03	0.08	0.10	0.25	0.29
Biology Score	12.02	17.42	-28.32	168.74	139.41	4264.04	-7.20	12.32	-6.07	17.43	9.09	5.74	2.60	6.45	7.64	9.34	1.23	4.84	-1.21	8.82	-2.83	4.20
Biology_well	0.07	0.12	-0.38	2.57	7.82	922.93	0.01	0.07	0.13	0.51	**0.08***	0.03	0.02	0.05	0.00	0.05	-0.01	0.06	-0.03	0.10	0.01	0.04
Biology_quick	0.01	0.07	-0.02	0.31	-1.04	41.71	0.12	0.18	0.17	0.82	**0.11***	0.03	0.02	0.06	0.02	0.08	-0.03	0.07	-0.04	0.13	0.06	0.06
Biology_enjoy	0.09	0.16	-0.31	2.63	0.59	11.87	0.15	0.23	0.27	1.01	**0.15***	0.07	0.03	0.05	-0.06	0.06	-0.07	0.10	-0.05	0.14	0.10	0.09
Biology_hard	0.12	0.16	-0.10	0.76	0.13	6.10	-0.01	0.04	0.31	0.97	**0.05***	0.02	0.07	0.06	-0.09	0.08	-0.05	0.06	0.00	0.08	0.06	0.06
Biology_weak	0.08	0.15	0.00	0.42	1.07	70.77	0.10	0.17	0.22	0.67	**0.12***	0.04	0.06	0.05	-0.07	0.07	-0.06	0.08	-0.05	0.13	0.06	0.06
Chemistry Score	-18.03	76.01	-3.76	5.37	-3.73	7.74	-30.56	85.81	2.78	3.32	**10.09***	2.78	15.26	11.82	-0.46	20.34	-1.08	4.63	-9.96	16.91	-6.19	7.92
Chemistry_well	-0.14	0.83	-0.01	0.05	-0.02	0.07	-0.37	0.98	0.03	0.07	0.07	0.04	0.07	0.08	0.22	0.34	0.04	0.06	0.12	0.19	0.00	0.07
Chemistry_quick	0.16	0.74	0.03	0.04	0.03	0.05	-0.25	0.69	-0.04	0.08	0.03	0.05	0.12	0.08	0.44	0.57	0.06	0.07	0.10	0.15	0.02	0.06
Chemistry_enjoy	0.27	1.62	0.02	0.05	0.07	0.08	-0.13	0.36	-0.02	0.07	0.07	0.06	0.13	0.09	0.55	0.81	0.05	0.09	0.08	0.18	-0.06	0.08
Chemistry_hard	-0.43	2.12	-0.01	0.04	-0.03	0.07	-0.40	1.07	0.06	0.06	0.03	0.03	-0.02	0.05	-0.04	0.17	0.01	0.05	0.06	0.11	0.06	0.07
Chemistry_weak	-0.36	1.83	-0.01	0.04	-0.02	0.06	-0.26	0.72	-0.07	0.08	0.07	0.05	0.00	0.05	0.11	0.24	0.02	0.05	0.01	0.09	0.02	0.07
Earth Score	-9.31	123.12	0.48	8.02	2.22	16.28	15.49	43.26	0.57	3.03	**10.85***	3.29	6.34	3.88	-1.89	3.61	-	-	-	-	-	-
Earth_well	0.50	6.28	0.10	0.11	0.12	0.27	-0.01	0.12	-0.04	0.05	-0.04	0.05	-0.02	0.03	-0.03	0.03	-	-	-	-	-	-
Earth_quick	-0.17	2.70	0.02	0.07	0.00	0.12	0.15	0.43	0.01	0.06	-0.06	0.06	-0.01	0.03	-0.03	0.03	-	-	-	-	-	-
Earth_enjoy	0.04	1.46	0.06	0.11	0.19	0.45	-0.09	0.22	-0.03	0.07	-0.03	0.07	-0.06	0.03	-0.05	0.02	-	-	-	-	-	-
Earth_hard	0.59	8.06	0.08	0.10	0.04	0.11	-0.03	0.07	-0.05	0.05	0.01	0.02	0.01	0.02	0.00	0.02	-	-	-	-	-	-
Earth_weak	-0.34	4.72	0.03	0.06	-0.05	0.10	-0.03	0.10	0.01	0.03	-0.02	0.03	0.00	0.02	-0.02	0.02	-	-	-	-	-	-

Note: * p<0.05

Table 4.8: RD results in 2007

	Hungary								Lithuania				Romania				Slovenia					
Bandwidth	6		5		4		3		4		3		4		3		5		4		3	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Math Score	-2.01	5.22	-20.42	111.76	-26.35	56.73	-9.25	12.41	-2.48	2.58	1.57	1.67	-1.64	4.46	-3.88	5.33	0.56	5.36	-0.37	10.55	-0.70	5.68
Math_well	-0.02	0.02	-0.16	0.64	-0.10	0.17	-0.04	0.04	0.06	0.03	0.04	0.02	0.01	0.02	0.04	0.03	0.01	0.05	-0.07	0.13	-0.09	0.07
Math_quick	0.00	0.03	-0.18	0.67	-0.13	0.18	-0.07	0.04	0.03	0.04	0.03	0.03	0.03	0.02	0.06	0.03	0.04	0.05	-0.05	0.07	-0.08	0.05
Math_enjoy	0.01	0.03	-0.14	0.56	-0.05	0.09	-0.08	0.06	0.01	0.03	0.03	0.02	-0.01	0.03	0.02	0.03	-0.05	0.07	-0.17	0.19	-0.10	0.07
Math_hard	-0.02	0.02	-0.20	0.85	-0.07	0.12	-0.05	0.04	-0.01	0.03	-0.01	0.02	0.03	0.04	0.03	0.05	0.00	0.05	-0.12	0.14	-0.08	0.06
Math_weak	0.01	0.02	0.02	0.20	-0.04	0.08	-0.04	0.03	-0.01	0.03	0.02	0.03	0.03	0.03	0.04	0.05	0.07	0.06	0.07	0.12	0.01	0.05
Physics Score	-7.79	26.10	-3.84	15.65	-66.92	364.27	-10.55	19.11	-0.29	2.75	3.18	2.17	-1.91	9.51	-1.55	6.19	4.55	9.77	3.93	11.54	4.45	6.95
Physics_well	0.08	0.18	0.11	0.18	0.43	1.61	0.05	0.10	0.05	0.05	0.01	0.02	-0.01	0.06	0.02	0.04	0.05	0.11	0.14	0.24	0.05	0.08
Physics_quick	0.05	0.13	0.11	0.16	0.30	1.07	0.04	0.08	0.08	0.05	0.05	0.03	0.09	0.09	0.04	0.04	0.02	0.11	0.04	0.18	0.02	0.08
Physics_enjoy	0.03	0.14	0.06	0.13	0.29	1.14	-0.03	0.07	0.06	0.04	**0.06***	0.02	-0.03	0.08	-0.04	0.06	0.14	0.21	0.18	0.32	0.05	0.10
Physics_hard	-0.03	0.10	0.00	0.08	0.10	0.45	-0.01	0.06	-0.01	0.03	-0.01	0.02	0.16	0.10	0.04	0.04	0.11	0.16	0.11	0.22	0.11	0.12
Physics_weak	0.03	0.11	0.10	0.15	0.29	0.96	0.01	0.06	-0.01	0.03	0.00	0.02	0.05	0.06	0.05	0.04	0.05	0.10	0.09	0.18	0.06	0.08
Biology Score	9.99	92.90	-3.86	19.64	46.87	186.81	-25.20	45.50	0.32	4.47	-0.27	2.73	88.44	475.12	23.49	55.50	-7.96	11.93	-7.77	11.14	-2.81	6.59
Biology_well	-0.12	2.85	0.11	0.21	-0.17	1.01	0.05	0.08	0.00	0.02	**-0.03***	0.01	-0.34	1.71	-0.21	0.51	-0.30	0.36	-0.31	0.37	-0.19	0.18
Biology_quick	0.21	3.51	0.11	0.19	-0.36	2.13	0.03	0.08	0.01	0.04	**-0.05***	0.02	0.03	0.45	-0.22	0.43	-0.33	0.39	-0.35	0.40	-0.24	0.22
Biology_enjoy	0.37	5.31	0.11	0.21	-0.12	0.61	0.00	0.06	0.04	0.03	-0.02	0.04	-0.02	0.62	-0.10	0.34	-0.38	0.41	-0.40	0.44	-0.29	0.25
Biology_hard	-0.02	0.70	0.09	0.14	-0.28	1.73	0.02	0.05	-0.04	0.04	-0.04	0.02	0.52	2.35	0.20	0.41	-0.25	0.28	-0.26	0.28	-0.20	0.16
Biology_weak	-0.36	5.30	0.17	0.28	-0.34	1.84	0.05	0.07	0.00	0.03	0.00	0.02	0.06	0.58	-0.10	0.21	-0.25	0.28	-0.25	0.28	-0.17	0.15
Chemistry Score	-6.14	14.28	-6.52	13.56	-24.45	34.05	-21.80	23.05	2.37	3.84	-0.68	2.30	1.73	4.51	1.17	5.15	-8.92	26.88	-4.55	19.93	-99.39	2323.57
Chemistry_well	0.08	0.12	0.17	0.17	0.26	0.36	0.12	0.15	-0.02	0.07	-0.03	0.05	0.05	0.03	0.05	0.03	0.11	0.26	0.16	0.39	1.21	53.28
Chemistry_quick	0.06	0.08	0.14	0.13	0.23	0.31	0.11	0.12	-0.02	0.06	-0.05	0.05	0.04	0.03	**0.07***	0.04	0.14	0.32	0.24	0.60	8.62	747.30
Chemistry_enjoy	0.05	0.09	0.11	0.12	0.17	0.29	0.02	0.11	-0.03	0.06	-0.05	0.04	0.06	0.03	0.06	0.04	0.22	0.48	0.33	0.84	6.76	374.51
Chemistry_hard	0.06	0.08	0.08	0.08	0.18	0.24	0.10	0.09	-0.04	0.05	-0.03	0.02	0.05	0.03	0.03	0.04	-0.09	0.25	-0.02	0.13	-2.78	110.31
Chemistry_weak	0.04	0.09	0.12	0.12	0.20	0.28	0.09	0.11	-0.04	0.06	-0.05	0.04	**0.06***	0.03	0.07	0.04	-0.01	0.14	0.07	0.19	0.88	27.39
Earth Score	-6.37	12.38	-2.53	7.19	-10.31	17.26	-16.50	21.28	-1.06	3.05	-0.59	3.08	6.09	4.08	**7.92***	4.02	-	-	-	-	-	-
Earth_well	-0.06	0.08	-0.03	0.05	-0.03	0.07	0.08	0.10	**-0.08***	0.04	**-0.14***	0.02	0.10	0.06	0.07	0.05	-	-	-	-	-	-
Earth_quick	-0.07	0.07	-0.03	0.04	-0.03	0.06	0.07	0.11	-0.06	0.03	**-0.14***	0.02	**0.11***	0.06	0.09	0.05	-	-	-	-	-	-
Earth_enjoy	-0.08	0.10	-0.05	0.05	-0.10	0.11	-0.07	0.11	-0.03	0.06	**-0.15***	0.02	0.08	0.06	0.06	0.06	-	-	-	-	-	-
Earth_hard	-0.03	0.05	-0.02	0.04	-0.01	0.06	0.06	0.07	**-0.07***	0.03	**-0.09***	0.03	**0.10***	0.04	0.08	0.05	-	-	-	-	-	-
Earth_weak	-0.06	0.08	-0.02	0.05	-0.01	0.08	0.08	0.11	**-0.11***	0.04	**-0.13***	0.02	0.07	0.05	0.07	0.05	-	-	-	-	-	-

Note: * p<0.05

Table 4.9: RD results in 2011

Bandwidth	Hungary								Lithuania				Romania				Slovenia					
	6		5		4		3		4		3		4		3		5		4		3	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Math Score	2.15	2.50	0.66	2.70	3.85	2.90	16.74	27.45	-4.07	5.10	2.64	6.28	24.60	20.42	24.09	17.16	11.02	7.92	134.87	948.83	-16.01	12.76
Math_well	-0.01	0.02	-0.01	0.02	-0.01	0.02	0.03	0.08	-0.01	0.03	-0.04	0.09	0.09	0.07	0.08	0.06	0.07	0.06	0.81	4.83	-0.18	0.17
Math_quick	-0.01	0.02	-0.01	0.02	-0.02	0.02	0.01	0.09	-0.07	0.05	-0.14	0.11	0.10	0.08	0.09	0.07	0.06	0.06	1.00	6.23	-0.17	0.16
Math_enjoy	-0.03	0.03	**-0.06***	0.02	**-0.06***	0.03	-0.21	0.16	-0.07	0.05	-0.25	0.15	0.12	0.11	0.10	0.09	0.16	0.12	1.27	8.75	-0.20	0.18
Math_hard	-0.01	0.02	-0.03	0.02	0.00	0.02	0.11	0.14	-0.02	0.05	-0.12	0.14	0.01	0.03	0.03	0.03	0.02	0.04	0.48	3.19	-0.15	0.13
Math_weak	-0.03	0.03	-0.03	0.03	-0.03	0.03	0.00	0.06	-0.03	0.04	-0.04	0.07	0.08	0.07	0.10	0.07	0.08	0.06	0.87	5.85	-0.16	0.15
Physics Score	5.26	6.51	27.36	269.25	2.75	12.34	14.41	29.45	-4.20	4.27	2.37	1.83	-50.95	145.24	-43.57	133.54	43.06	122.95	264.29	3359.48	-23.04	39.99
Physics_well	-0.01	0.08	-0.50	4.73	0.02	0.16	0.05	0.24	0.03	0.03	0.00	0.03	-0.07	0.17	-0.09	0.23	0.54	1.60	4.14	87.08	-0.33	0.64
Physics_quick	0.01	0.08	-0.47	4.20	0.04	0.19	0.11	0.30	0.04	0.04	0.06	0.04	-0.13	0.32	-0.12	0.31	0.34	0.94	3.13	59.49	-0.32	0.65
Physics_enjoy	0.01	0.10	-0.80	8.19	0.10	0.36	0.09	0.29	0.04	0.05	0.01	0.03	-0.08	0.27	-0.10	0.35	0.57	1.68	3.44	60.13	-0.42	0.77
Physics_hard	-0.02	0.05	-0.40	4.14	0.04	0.15	0.11	0.22	0.06	0.03	-0.02	0.03	0.02	0.07	0.00	0.07	0.14	0.39	2.24	57.71	-0.20	0.39
Physics_weak	0.04	0.08	-0.79	7.44	0.18	0.46	0.22	0.49	0.01	0.05	-0.04	0.05	0.03	0.09	0.04	0.13	0.31	0.97	3.53	158.88	-0.21	0.40
Biology Score	1.49	4.42	-39.99	476.90	4.05	10.65	-29.36	270.53	-6.33	13.16	-1.89	7.45	16.94	13.85	21.43	16.74	-15.37	29.04	64.45	274.05	-35.94	134.06
Biology_well	0.09	0.10	0.86	12.42	0.14	0.32	-0.25	1.79	0.14	0.20	0.17	0.10	0.04	0.05	0.06	0.08	-0.11	0.21	0.00	0.22	-0.10	0.49
Biology_quick	0.08	0.08	0.76	9.90	0.15	0.34	-0.35	2.58	0.17	0.23	0.17	0.10	0.02	0.05	0.02	0.07	-0.16	0.29	0.15	0.76	-0.26	0.99
Biology_enjoy	0.07	0.09	0.44	5.47	0.13	0.32	-0.22	1.66	0.24	0.35	0.28	0.18	0.02	0.06	0.02	0.07	0.03	0.14	-0.78	3.40	0.30	1.18
Biology_hard	0.00	0.03	0.26	4.37	0.08	0.17	0.07	0.41	0.05	0.09	0.16	0.10	0.00	0.03	0.01	0.03	-0.06	0.12	-0.09	0.47	0.04	0.26
Biology_weak	0.04	0.07	0.41	6.65	0.15	0.35	-0.34	2.22	0.10	0.12	0.18	0.10	0.00	0.05	0.01	0.05	-0.13	0.25	0.00	0.24	-0.03	0.28
Chemistry Score	1.49	6.29	-431.82	33892.58	1.38	11.95	-20.48	48.75	-4.84	12.10	-10.47	27.05	28.52	32.02	27.03	21.23	30.41	88.21	146.18	2214.32	-28.52	149.53
Chemistry_well	0.14	0.18	9.95	993.88	0.38	0.80	0.27	0.58	0.17	0.33	-0.67	1.21	0.11	0.10	0.13	0.09	0.09	0.29	-0.72	16.02	0.12	0.67
Chemistry_quick	0.11	0.14	3.35	165.53	0.28	0.61	0.23	0.51	0.18	0.31	-0.60	1.12	0.06	0.07	0.09	0.07	0.02	0.20	-0.45	10.40	-0.21	1.21
Chemistry_enjoy	0.19	0.24	16.24	2598.37	0.47	1.03	0.14	0.32	0.21	0.41	-0.89	1.60	0.03	0.08	0.04	0.08	-0.24	0.80	-2.12	38.31	0.21	1.24
Chemistry_hard	0.16	0.18	40.61	14995.92	0.28	0.58	0.04	0.07	0.13	0.23	-0.48	0.94	0.01	0.05	0.02	0.05	-0.04	0.23	-0.27	12.62	-0.13	0.71
Chemistry_weak	0.15	0.19	-1.99	33.93	0.39	0.88	0.09	0.10	0.05	0.07	-0.10	0.23	0.02	0.05	0.03	0.05	0.07	0.27	-0.86	22.00	-0.06	0.51
Earth Score	2.03	2.98	-15.74	52.58	1.31	3.49	5.39	13.19	2.74	6.50	**8.45***	2.18	14.84	9.04	**12.53***	6.06	20.93	28.04	29.08	52.66	26.24	90.85
Earth_well	0.06	0.07	0.32	1.06	0.04	0.07	-0.04	0.08	0.13	0.12	**0.05***	0.02	0.06	0.05	0.07	0.05	0.13	0.17	0.18	0.30	0.29	0.86
Earth_quick	0.10	0.10	0.34	1.01	0.07	0.10	-0.07	0.08	0.10	0.10	0.04	0.03	0.09	0.06	0.10	0.06	0.07	0.14	0.15	0.29	0.33	1.05
Earth_enjoy	0.09	0.10	0.45	1.51	0.09	0.12	-0.06	0.07	0.22	0.23	0.01	0.04	0.09	0.07	0.10	0.06	0.12	0.23	0.05	0.20	0.16	0.63
Earth_hard	0.05	0.05	0.24	0.80	0.04	0.06	-0.11	0.23	0.18	0.18	**0.11***	0.04	0.06	0.04	0.06	0.03	0.04	0.09	0.08	0.16	0.10	0.37
Earth_weak	0.08	0.09	0.43	1.57	0.08	0.10	-0.08	0.11	0.07	0.08	**0.10***	0.04	0.05	0.04	0.06	0.04	0.06	0.12	0.13	0.24	0.14	0.50

Note: * p<0.05

Table 4.10: IV first stage F-test values and correlation of class size in RD sample

RD	2003					2007					2011				
	Math	**Physics**	**Biology**	**Chemistry**	**Earth**	**Math**	**Physics**	**Biology**	**Chemistry**	**Earth**	**Math**	**Physics**	**Biology**	**Chemistry**	**Earth**
Hungary															
Bandwidth 6															
1st Stage	0.2	1.1	0.2	0.1	0.7	0.0	0.7	0.9	0.3	0.1	0.5	1.2	0.0	0.3	0.3
Correlation	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	-0.1	-0.1	0.0	-0.2	-0.2	-0.1	-0.1	-0.1
Bandwidth 5															
1st Stage	0.0	1.4	1.0	0.3	1.3	0.5	0.0	1.0	0.0	0.3	0.2	1.6	0.1	0.5	0.2
Correlation	0.0	-0.1	-0.1	0.0	-0.1	0.2	0.0	0.0	0.0	0.1	-0.2	-0.2	-0.2	-0.2	-0.2
Bandwidth 4															
1st Stage	0.0	1.1	1.2	0.1	1.2	1.1	0.2	0.1	0.0	1.8	5.7	0.0	0.1	1.2	0.0
Correlation	0.0	-0.1	-0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.1	-0.4	-0.2	-0.2	-0.2	-0.2
Bandwidth 3															
1st Stage	0.6	1.5	1.8	0.1	0.0	0.7	0.0	0.1	1.4	4.3	**13.6**	3.2	1.1	1.1	0.3
Correlation	0.0	-0.1	-0.1	0.0	0.0	0.2	0.0	-0.1	0.1	0.1	-0.5	-0.2	-0.4	-0.3	-0.4
Lithuania															
Bandwidth 4															
1st Stage	0.4	0.3	0.6	0.8	0.8	**12.0**	8.9	**23.5**	7.3	**12.4**	9.2	9.2	7.1	6.8	**15.2**
Correlation	-0.1	0.0	-0.1	-0.1	0.0	-0.5	-0.5	-0.5	-0.5	-0.5	-0.4	-0.4	-0.4	-0.4	-0.4
Bandwidth 3															
1st Stage	0.6	0.5	0.1	0.5	1.0	8.2	5.9	**13.3**	**10.0**	9.4	7.5	3.7	4.8	9.9	8.0
Correlation	-0.2	-0.1	-0.2	-0.1	-0.1	-0.5	-0.5	-0.5	-0.5	-0.5	-0.3	-0.3	-0.3	-0.3	-0.3
Romania															
Bandwidth 4															
1st Stage	0.3	0.0	0.2	0.0	0.1	**11.4**	4.4	0.1	6.2	**11.4**	0.3	1.9	0.6	0.6	0.1
Correlation	0.1	0.0	0.0	0.0	0.2	-0.3	-0.3	-0.1	-0.2	-0.2	-0.1	-0.2	0.0	-0.1	0.1
Bandwidth 3															
1st Stage	3.1	0.1	0.0	0.0	0.9	**16.3**	5.1	0.6	9.3	**15.0**	0.1	0.9	0.1	0.2	0.4
Correlation	0.1	0.0	0.0	0.0	0.2	-0.3	-0.3	-0.1	-0.3	-0.2	0.0	-0.1	0.1	0.0	0.1
Slovenia															
Bandwidth 5															
1st Stage	7.8	**25.6**	**31.3**	**25.6**	-	**10.9**	**23.6**	**26.5**	**11.3**	-	4.9	**40.6**	**12.3**	**40.0**	**24.5**
Correlation	-0.3	-0.5	-0.5	-0.5	-	-0.2	-0.5	-0.5	-0.4	-	-0.2	-0.6	-0.4	-0.5	-0.5
Bandwidth 4															
1st Stage	**13.5**	**23.6**	**18.2**	**16.4**	-	**13.0**	**15.5**	**14.8**	8.7	-	4.5	**23.0**	**18.6**	**23.3**	**14.6**
Correlation	-0.4	-0.5	-0.4	-0.4	-	-0.3	-0.5	-0.5	-0.3	-	-0.2	-0.5	-0.4	-0.5	-0.5
Bandwidth 3															
1st Stage	5.4	**21.3**	5.2	3.5	-	5.5	0.5	0.9	0.0	-	**12.8**	**25.4**	**20.9**	**19.5**	**10.9**
Correlation	-0.4	-0.6	-0.4	-0.4	-	-0.3	-0.4	-0.4	-0.2	-	-0.3	-0.5	-0.4	-0.4	-0.5

Table 4.11: T-test values that check local balance of covariates in 2007

Bandwidth	Hungary								Lithuania				Romania				Slovenia					
	6		5		4		3		4		3		4		3		5		4		3	
	Est	Sig	Est	Sig	Est	Sig	Est	Sig	Est	Sig	Est	Sig	Est	Sig	Est	Sig	Est	Sig	Est	Sig	Est	Sig
Math	20.4		36.6		31.0		2.0		-13.3		3.1		-13.6		0.6		9.5		7.4		6.5	
Female	0.1		0.1	*	0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0	
Age	0.0		0.0		0.0		0.0		0.0		0.0		-0.1	*	-0.1		0.0		0.0		-0.1	
Home items	0.2		0.3		0.5		0.2		0.1		0.3		-0.3		0.0		0.2	*	0.2		0.1	
Male teacher	0.2	*	0.2		0.3	*	0.2		-0.1		0.0		-0.1		-0.1		0.0		-0.1		0.0	
Teacher education	-		-		-		-		0.0		0.0		0.0		0.0		0.0		0.0		0.0	
Teacher experience	1.0		1.4		-0.4		-0.8		1.6		-1.8		-1.4		-4.0		0.7		0.4		-0.3	
ED percent	-0.7		-0.8		-0.7		-0.6		0.7		-0.1		0.6		0.2		-0.6		-0.6		-1.0	
Physics	13.3		25.8		23.4		-0.1		-11.0		2.5		-20.7		-7.3		10.7		8.4		5.3	
Female	0.1		0.1	*	0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0	
Age	0.0		0.0		0.0		0.0		0.0		0.0		-0.1	*	-0.1		0.0		0.0		-0.1	
Home items	0.2		0.3		0.5		0.2		0.1		0.3		-0.3		0.0		0.2	*	0.2		0.1	*
Male teacher	0.1		0.1		0.1		0.1		-0.1		-0.2		-0.1		-0.1		0.1		0.1		0.3	
Teacher education	-		-		-		-		0.2		0.2		0.0		-0.1		0.0		0.0		0.1	
Teacher experience	-2.6		-2.7		-1.0		-1.6		3.9		3.6		-1.9		-0.5		1.2		0.0		5.0	
ED percent	-0.7		-0.8		-0.7		-0.6		0.7		-0.1		0.6		0.2		-0.6		-0.6		-1.0	
Biology	12.0		24.1		20.9		-0.7		-10.0		6.8		-19.4		-5.0		13.5	*	12.5		16.1	
Female	0.1		0.1	*	0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0	
Age	0.0		0.0		0.0		0.0		0.0		0.0		-0.1	*	-0.1		0.0		0.0		-0.1	
Home items	0.2		0.3		0.5		0.2		0.1		0.3		-0.3		0.0		0.2	*	0.2		0.1	*
Male teacher	0.2	*	0.3	*	0.2		0.3		0.1		0.0		-0.1		-0.1		-0.1		-0.1		-	
Teacher education	-		-		-		-		-0.2		-0.2		0.0		0.0		-		-		-	
Teacher experience	1.5		0.6		-0.4		2.1		-5.9		-4.3		12.6	*	14.0	*	-0.1		2.0		2.7	
ED percent	-0.7		-0.8		-0.7		-0.6		0.7		-0.1		0.6		0.2		-0.6		-0.6		-1.0	
Chemistry	11.9		27.1		20.5		-6.5		-10.9		1.5		-20.0		-5.5		13.1		9.1		7.2	
Female	0.1		0.1	*	0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0	
Age	0.0		0.0		0.0		0.0		0.0		0.0		-0.1	*	-0.1		0.0		0.0		-0.1	
Home items	0.2		0.3		0.5		0.2		0.1		0.3		-0.3		0.0		0.2	*	0.2		0.1	*
Male teacher	0.2		0.2		0.1		-0.1		-0.2		-0.2		0.2		0.2		0.0		0.0		-	
Teacher education	-		-		-		-		-0.2		-0.1		0.1		0.1		0.1		0.1		0.1	
Teacher experience	-5.2		-4.8		-1.7		-2.4		-4.0		-3.4		6.3		6.9		-1.0		2.0		5.4	
ED percent	-0.7		-0.8		-0.7		-0.6		0.7		-0.1		0.6		0.2		-0.6		-0.6		-1.0	
Earth Science	13.9		26.1		24.8		2.0		-7.6		7.8		-20.0		-3.7		-	-	-	-	-	
Female	0.1		0.1	*	0.0		0.0		0.0		0.0		0.0		0.0		-	-	-	-	-	
Age	0.0		0.0		0.0		0.0		0.0		0.0		-0.1	*	-0.1		-	-	-	-	-	
Home items	0.2		0.3		0.5		0.2		0.1		0.3		-0.3		0.0		-	-	-	-	-	
Male teacher	0.0		-0.2		-0.2		-0.2		0.0		0.0		0.2		0.1		-	-	-	-	-	
Teacher education	-		-		-		-		0.2		0.2		0.0		-0.1		-	-	-	-	-	
Teacher experience	-4.3		-4.1		-6.6		-0.2		2.6		-3.9		2.6		3.5		-	-	-	-	-	
ED percent	-0.7		-0.8		-0.7		-0.6		0.7		-0.1		0.6		0.2		-	-	-	-	-	

Note: * p<0.05

Table 4.12: Class size result summary of statistical significant estimates

Weighted full sample					
	Est	SE	ES	N	% Bias
	2003				
Romania					
Math score	-5.602	2.134	0.063	4104	25.32
Physics score	-5.410	2.439	0.067	4104	11.62
Chemistry score	-5.090	2.537	0.054	4080	2.30
Earth score	-8.747	3.703	0.093	4104	17.00
	2007				
Romania					
math_enjoy	-0.048	0.023	0.047	4046	7.70
Lithuania					
Biology_well	0.023	0.010	0.031	3956	14.59
Biology_quick	0.034	0.012	0.039	3924	32.07
Biology_weak	0.032	0.012	0.037	3900	28.63
Earth_quick	-0.030	0.013	0.033	3923	15.98
	2011				
Lithuania					
Biology_enjoy	-0.029	0.012	0.034	4690	19.93
Chemistry_well	-0.028	0.010	0.029	4701	30.41
Chemistry_enjoy	-0.028	0.010	0.028	4694	30.41

Unweighted full sample					
	Est	SE	ES	N	% Bias
	2003				
Romania					
Math score	-7.346	3.395	0.083	4104	9.39
Chemistry score	-8.114	4.034	0.085	4080	2.54
	2007				
Lithuania					
Biology_quick	0.034	0.014	0.039	3924	17.49
Biology_weak	0.032	0.013	0.037	3900	21.17
Earth_quick	-0.032	0.015	0.036	3923	8.03
	2011				
Lithuania					
Physics_hard	0.036	0.016	0.035	4671	13.48
Biology_enjoy	-0.041	0.018	0.048	4690	13.63
Chemistry_enjoy	-0.031	0.016	0.031	4694	0.85
Romania					
Chemistry_well	-0.048	0.023	-0.045	5313	4.90
Slovenia					
Earth_enjoy	0.032	0.016	0.034	4344	3.13

Unweighted RD sample					
	Est	SE	ES	N	% Bias
(bandwidth 3)	**2007**				
Lithuania					
Biology_well	-0.028	0.014	0.042	611	4.68
Biology_quick	-0.049	0.023	0.060	607	7.14
Earth_well	-0.144	0.025	0.192	607	66.50
Earth_quick	-0.137	0.024	0.155	606	64.94
Earth_enjoy	-0.150	0.015	0.156	603	79.86
Earth_hard	-0.092	0.026	0.097	604	44.23
Earth_weak	-0.132	0.019	0.150	601	71.52

Note: the estimates are all statistically significant at 0.05; Est=Estimate; SE=Standard error
ES=Effect size (the absolute value of estimate over the standard deviation of the dependent variable)

REFERENCES

REFERENCES

Akabayashi, H., & Nakamura, R. (2014). Can Small Class Policy Close the Gap? An Empirical Analysis of Class Size Effects in Japan. *Japanese Economic Review, 65*(3), 253-281.

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

Altinok, N., & Kingdon, G. (2012). New Evidence on Class Size Effects: A Pupil Fixed Effects Approach*. *Oxford Bulletin of Economics and Statistics, 74*(2), 203-234.

Ammermüller, A., Heijke, H., & Wößmann, L. (2005). Schooling quality in Eastern Europe: Educational production during transition. *Economics of Education Review, 24*(5), 579-599.

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association, 90*(430), 431-442.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association, 91*(434), 444-455.

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.

Areepattamannil, S., Freeman, J. G., & Klinger, D. A. (2011). Influence of motivation, self-beliefs, and instructional practices on science achievement of adolescents in Canada. *Social Psychology of Education, 14*(2), 233-259.

Asadullah, M. N. (2005). The effect of class size on student achievement: evidence from Bangladesh. *Applied Economics Letters, 12*(4), 217-221.

Biggs, J. (1998). Learning from the Confucian heritage: so size doesn't matter? *International Journal of Educational Research, 29*(8), 723-738.

Bingley, P., Myrup Jensen, V., & Walker, I. (2005). The effects of school class size on length of post-compulsory education: some cost-benefit analysis.

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and Instruction, 21*(6), 715-730.

Blatchford, P., Chan, K. W., Galton, M., Lai, K. C., & Lee, J. C. K. (2016). *Class Size: Eastern and Western Perspectives*: Routledge.

Blatchford, P., Moriarty, V., Edmonds, S., & Martin, C. (2002). Relationships between class size

and teaching: A multimethod analysis of English infant schools. *American Educational Research Journal, 39*(1), 101-132.

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

Bressoux, P., Kramarz, F., & Prost, C. (2009). Teachers' Training, Class Size and Students' Outcomes: Learning from Administrative Forecasting Mistakes*. *The Economic Journal, 119*(536), 540-561.

Breton, T. R. (2014). Evidence that class size matters in 4th grade mathematics: An analysis of TIMSS 2007 data for Colombia. *International Journal of Educational Development, 34*, 51-57.

Browning, M., & Heinesen, E. (2007). Class Size, Teacher Hours and Educational Attainment*. *The Scandinavian Journal of Economics, 109*(2), 415-438.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *Stata Journal, 17*(2), 372-404.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014a). Robust data-driven inference in the regression-discontinuity design. *Stata Journal, 14*(4), 909-946.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014b). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica, 82*(6), 2295-2326.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs. *R Journal, 7*(1), 38-51.

Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2016). Interpreting Regression Discontinuity Designs with Multiple Cutoffs. *Journal of Politics, 78*(4), 1229-1248.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics, 126*(4), 1593-1660.

Cheung, H. Y., & Chan, A. W. (2008). Relationships amongst cultural dimensions, educational expenditure and class size of different nations. *International Journal of Educational Development, 28*(6), 698-707.

Chingos, M. M. (2013). Class size and student outcomes: Research and policy implications. *Journal of Policy Analysis and Management, 32*(2), 411-438.

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.

De Giorgi, G., Pellizzari, M., & Woolston, W. G. (2012). Class size and class heterogeneity. *Journal of the European Economic Association, 10*(4), 795-830.

De Paola, M., Ponzo, M., & Scoppa, V. (2013). Class size effects on student achievement: heterogeneity across abilities and fields. *Education Economics, 21*(2), 135-153.

Dee, T. S., & West, M. R. (2011). The non-cognitive returns to class size. *Educational Evaluation and Policy Analysis, 33*(1), 23-46.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding Self-Report Bias in Organizational Behavior Research. *Journal of Business and Psychology, 17*(2), 245-260.

Eurydice. (2005). *Key Data on Education in Europe 2005*: Brussels: Eurydice.

Eurydice. (2009). *Key Data on Education in Europe 2009*: Brussels: Eurydice.

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

Finn, J. D. (2002). Small classes in American schools: Research, practice, and politics. *Phi Delta Kappan, 83*(7), 551-560.

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

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21*(2), 97-109.

Finn, J. D., Gerber, S. B., & Boyd-Zaharias, J. (2005a). Small classes in the early grades, academic achievement, and graduating from high school. *Journal of educational psychology, 97*(2), 214-223.

Finn, J. D., Gerber, S. B., & Boyd-Zaharias, J. (2005b). Small classes in the early grades, academic achievement, and graduating from high school. *Journal of educational psychology, 97*(2), 214.

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis, 35*(4), 437-460.

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly Journal of Economics, 128*(1), 249-285.

Galton, M., & Pell, T. (2012). Do class size reductions make a difference to classroom practice? The case of Hong Kong primary schools. *International Journal of Educational Research, 53*, 22-31.

Gary-Bobo, R. J., & Mahjoub, M.-B. (2013). Estimation of Class-Size Effects, Using "Maimonides' Rule" and Other Instruments: the Case of French Junior High Schools. *Annals of Economics and Statistics*(111/112), 193-225.

Hargreaves, L., Galton, M., & Pell, A. (1998). The effects of changes in class size on teacher–pupil interaction. *International Journal of Educational Research, 29*(8), 779-795.

Heinesen, E. (2010). Estimating Class-size Effects using Within-school Variation in Subject-specific Classes*. *The Economic Journal, 120*(545), 737-760.

Hörner, W., Döbert, H., Kopp, B. V., & Mitter, W. (2007). *The education systems of Europe*: Springer.

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

Imbens, G., & Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *Review of Economic Studies, 79*(3), 933-959.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-635.

Kallai, E., & Maniu, M. (2004). Input Efficiency in Publicly Provided Education: The Case of Romania. *Babes-Bolyai University, Cluj-Napoca, Romania*.

Konstantopoulos, S., & Shen, T. (2016). Class size effects on mathematics achievement in Cyprus: evidence from TIMSS. *Educational Research and Evaluation, 22*(1-2), 86-109.

Konstantopoulos, S., & Traynor, A. (2014). Class Size Effects on Reading Achievement Using PIRLS Data: Evidence from Greece. *Teachers College Record, 116*(2), 1-29.

Krassel, K. F., & Heinesen, E. (2014). Class-size effects in secondary school. *Education Economics, 22*(4), 412-426.

Krueger, A. B. (1999). *Experimental estimates of education production functions*. Retrieved from The Quarterly Journal of Economics.

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature, 48*(2), 281-355.

Leuven, E., Oosterbeek, H., & Rønning, M. (2008). Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway*. *The Scandinavian Journal of Economics, 110*(4), 663-693.

Li, W., & Konstantopoulos, S. (2016). Class Size Effects on Fourth-Grade Mathematics

Achievement: Evidence From TIMSS 2011. *Journal of Research on Educational Effectiveness, 9*(4), 503-530.

Lindahl, M. (2005). Home versus school learning: A new approach to estimating the effect of class size on achievement. *The Scandinavian Journal of Economics, 107*(2), 375-394.

Marsh, H. W., Trautwein, U., Ludtke, O., Koller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76*(2), 397-416.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis, 21*(2), 127-142.

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. (2001). The long-term effects of small classes in early grades: Lasting benefits in mathematics achievement at grade 9. *The Journal of Experimental Education, 69*(3), 245-257.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1), 3-15.

Shen, T., & Konstantopoulos, S. (2017). Class size effects on reading achievement in Europe: Evidence from PIRLS. *Studies in Educational Evaluation, 53*, 98-114.

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 & Economic Statistics, 20*(4).

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

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

van de Mortel, T. F. (2008). Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, 25*(4), 40-48.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and Paradox: Measuring Students' Non-Cognitive Skills and the Impact of Schooling. *Educational Evaluation and Policy Analysis, 38*(1), 148-170.

Wößmann, 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.