UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

MODELACIÓN ESTADÍSTICA DEL IMPACTO DE CONTEXTOS ACADÉMICOS EN LAS CREENCIAS Y EL DESEMPEÑO DE ESTUDIANTES EN CHILE

TESIS PARA OPTAR AL GRADO DE DOCTOR EN CIENCIAS DE LA
INGENIERÍA, MENCIÓN MODELACIÓN MATEMÁTICA
EN COTUTELA CON LA UNIVERSIDAD GRENOBLE ALPES

VALENTINA SOFÍA GIACONI SMOJE

PROFESORES GUÍA:
PASCAL BRESSOUX
PATRICIO FELMER AICHELE

MIEMBROS DE LA COMISIÓN:
BIEKE DE FRAINE
CHRISTIAN MONSEUR
PASCAL PANSU
LEONOR VARAS SCHEUCH

Este trabajo ha sido parcialmente financiado por CONICYT-Beca Doctorado Nacional 2013

SANTIAGO DE CHILE
2017

ABSTRACT
STATISTICAL MODELING OF THE IMPACT OF ACADEMIC ENVIRONMENTS ON STUDENT'S BELIEFS AND ACHIEVEMENT IN CHILE

This PhD thesis is dedicated to the statistical modeling of the impact of academic environments on student's beliefs and achievement in Chile. We contribute to the field of educational effectiveness with a statistical discussion regarding how to combine multilevel models with methods for selection bias and missing data and two empirical studies. The statistical discussion was used to take methodological decisions in the empirical studies. The first empirical study evaluates the impact of science courses on students' beliefs. The second empirical study is about school effects on students' trajectories in mathematics and reading scores.

In the statistical part, we analyze linear adjustment and propensity score matching to address selection bias. Regarding the missing data problem, we considered multiple imputation techniques. Each of these methods is compatible with multilevel models. However, the problem of addressing selection bias and missing data simultaneously with hierarchical data is not resolved. We present a statistical discussion that classifies and analyzes strategies to combine the methods.

The first empirical study regards the influence of Life and Non-life science courses in secondary students' epistemic and self-efficacy beliefs related to sciences. We compared students that took summer science courses with a control group in a post and follow-up beliefs questionnaire. We found positive effects of Life courses and courses with laboratory work, controlling for confounding variables. The results show differences between Life and Non-life scientific disciplines that should be explored.

The second empirical study concerns school effects on trajectories of Chilean students. It has two aims. The first aim is to describe the characteristics of the trajectories in mathematics and reading scores and the variation explained by primary and secondary schools. The second aim is to measure the effect of public schools in comparison with voucher schools on students' trajectories in mathematics and reading scores. We used a longitudinal national database which included measures for the same students at 4th, 8th and 10th grade. Multilevel growth models were used to model the trajectories. We found effects of secondary and primary schools on intercepts and slopes. In addition, we found negative effects from public education, which became not significant after controlling for school' socioeconomic composition and selection practices. The results illustrate the stratification between the public system and voucher system and the need to study inside each system which schools are more efficient.

Keywords: Multilevel modeling, Propensity score matching, Multiple imputation, Science disciplines, School effects, Students' trajectories.

RESUMEN
MODELACIÓN ESTADÍSTICA DEL IMPACTO DE CONTEXTOS ACADÉMICOS EN LAS CREENCIAS Y EL DESEMPEÑO DE ESTUDIANTES EN CHILE

Esta tesis de doctorado está dedicada a la modelación estadística del impacto de entornos académicos en las creencias y desempeño de estudiantes en Chile. Contribuimos al campo de la efectividad educativa con una discusión estadística y dos estudios empíricos. La discusión estadística trata sobre cómo combinar modelos multinivel con métodos para controlar el sesgo de selección y datos perdidos. Esta discusión estadística es usada para tomar decisiones metodológicas en los estudios empíricos. El primer estudio empírico evalúa el impacto de cursos de ciencias sobre las creencias de los estudiantes. El segundo estudio empírico trata de los efectos de las escuelas en las trayectorias de puntajes en matemáticas y lectura de estudiantes.

En la parte de estadística, se describe y analiza el uso de ajuste lineal y de la técnica de puntajes de propensión para controlar el sesgo de selección. Respecto a los problemas de datos perdidos, consideramos técnicas de imputación múltiple. Cada uno de estos métodos es compatible con los modelos multinivel. Sin embargo, enfrentar problemas de sesgo de selección y datos perdidos con datos jerárquicos es aún un problema abierto. Por esto, se presenta una discusión que clasifica y analiza estrategias para combinar estos métodos en presencia de datos jerárquicos.

El primer estudio empírico se refiere a la influencia de las disciplinas científicas que estudian objetos vivos y objetos no vivos en las creencias epistémicas y de autoeficacia relacionadas con las ciencias de los estudiantes de secundaria. Se compararon alumnos que asistieron a cursos de ciencias con un grupo de control al final de los cursos y cuatro meses después. Se encontraron efectos positivos del trabajo en laboratorio y de las disciplinas que estudian objetos vivos. Este estudio muestra que hay diferencias entre las disciplinas que estudian objetos vivos con las disciplinas que estudian objetos no vivos que aún deben ser exploradas.

El segundo estudio empírico se refiere a los efectos de las escuelas en las trayectorias de puntajes en matemáticas y lectura de estudiantes. El primer objetivo es describir las características de las trayectorias en matemáticas y lectura y la varianza explicada por las escuelas primarias y secundarias. El segundo objetivo es medir el efecto de las escuelas públicas en comparación con las escuelas particulares subvencionadas sobre las trayectorias de los estudiantes en matemáticas y lectura. Se utilizó una base de datos nacional longitudinal que incluye medidas para los mismos estudiantes en 4to básico, 8vo básico y 2do medio. Se aplicaron modelos multinivel de crecimiento para modelar las trayectorias. Los resultados obtenidos muestran que las escuelas de secundaria y primaria tienen efectos en las medias y pendientes de las trayectorias. Además, se encontraron efectos negativos de la educación pública, que se vuelven no significativos después de controlar por la composición socioeconómica y prácticas de selección de las escuelas. Los resultados ilustran la estratificación entre el sistema público y el sistema particular subvencionado y la necesidad de estudiar dentro de cada sistema cuáles son las escuelas más eficientes.

Palabras Clave: Modelos multinivel, Apareamiento de puntajes de propensión, Imputación múltiple, Disciplinas científicas, Efectos de escuelas, Trayectorias de estudiantes.

RÉSUMÉ
MODÉLISATION STATISTIQUE DE L'IMPACT DES ENVIRONNEMENTS ACADÉMIQUES SUR LES CROYANCES ET LA RÉUSSITE DES ÉLÈVES AU CHILI

Cette thèse de doctorat est consacrée à la modélisation statistique de l'impact des environnements académiques sur les croyances et la réussite des élèves au Chili. Nous contribuons au domaine de l'efficacité éducative avec une discussion statistique et deux études empiriques. La discussion statistique questionne la façon de combiner les modèles multiniveaux avec des méthodes pour le biais de sélection et pour les données manquantes. Cette discussion statistique sera utilisée pour prendre des décisions méthodologiques dans les études empiriques. La première étude empirique consiste en une évaluation d'intervention de l'impact des cours de sciences sur les croyances des étudiants. La deuxième étude empirique concerne l'effet des écoles sur les trajectoires des scores de mathématiques et de lecture des élèves.

Dans la partie statistique, nous avons décrit et analysé les méthodes d'ajustement linéaire et d'appariement des scores de propension pour modéliser le biais de sélection. En ce qui concerne les problèmes de données manquantes, nous avons analysé la méthode d'imputation multiple. Chacune de ces méthodes est compatible avec les modèles multiniveaux. En revanche, l'utilisation combinée de ces méthodes pour des données hiérarchiques n'est pas résolu. Nous présentons alors une discussion statistique qui analyse et classe des stratégies pour combiner ces méthodes.

La première étude empirique concerne l'influence des disciplines scientifiques qui s'intéressent à des objets vivants et non-vivants sur les croyances épistémiques et le sentiment d'auto-efficacité des étudiants de secondaire. Nous avons comparé, pour ces croyances, les étudiants qui ont suivi des cours de sciences à un groupe contrôle sur deux temps de mesure, à la fin des cours et 4 mois après. Nous avons constaté un effet positif du travail en laboratoire et des disciplines qui s'intéressent à des objets vivants (en contrôlant les variables confondues). Cette étude met en lumière des différences entre les disciplines qui s'intéressent à des objets vivant et des objets non-vivant qui devront être explorées.

La deuxième étude empirique concerne l'effet des écoles sur les trajectoires des scores en mathématiques et en lecture des élèves. Le premier objectif est de décrire les caractéristiques des trajectoires et la variance expliquée par les écoles primaires et secondaires. Le deuxième objectif est de mesurer l'effet du type d'école, publique ou voucher (privée avec un financement de l'état), sur les trajectoires. Nous avons utilisé une base de données nationale longitudinale qui comprenait des mesures pour les mêmes élèves en 4ème, 8ème et 10ème années. Des modèles de croissance multiniveaux ont été utilisés pour modéliser les trajectoires. Nos résultats montrent que les écoles secondaires et primaires ont un effet sur les interceptes et les pentes des trajectoires. Par ailleurs, nous avons constaté un effet négatif de l'école publique, qui est devenu non significatif lorsque nous avons contrôlé la composition socio-économique de l'école et ses pratiques de sélection. Ces résultats illustrent la stratification entre le système public et le système voucher ainsi que la nécessité de questionner l'efficacité des écoles pour chaque système.

Mots clés: Modèles multiniveaux, Appariement des scores de propension, Imputation multiple, Disciplines scientifiques, Effets-écoles, Trajectoires des élèves

Agradecimientos

Esta tesis se escribió en inglés y fue realizada en Chile y en Francia. Es por esto que estos agradecimientos están escritos en tres idiomas, ya que espero puedan expresar de mejor manera mi gratitud a todas las personas e instituciones involucradas.

First, I would like to thank Professor Patricio Felmer and Professor Pascal Bressoux for their support, encouragement and wise advices along all this PhD.
Profesor Felmer muchas gracias por haberme apoyado en todo lo que quise hacer, por ser siempre positivo y valorizar lo que hago, incluso frente a situaciones complejas. Su entusiasmo, confianza y consejos fueron fundamentales. También por su mirada aguda y por la sabiduría que puso en mi investigacion.
Professeur Bressoux, je voudrais vous remercier de votre courage de m'avoir accepté comme étudiant de doctorat, pour votre support constant et inconditional, pour vos conseils qui ont toujours donné naissance à beaucoup des apprentissages et des améliorations a mon projet de doctorat.

I would like to thank to the two reviewers, professors Bieke de Fraine and Christian Monseur for your useful comments and remarks that open new directions to continue my statistical learning and improve my research. I would like to thank also the members of the jury, professors Pascal Pansu and Leonor Varas, for accepting to evaluate my thesis and give me your advice and vision regarding my research.

This thesis received the support of several institutions. First, I would like to thank to CONICYT national scolarship (CONICYT-PCHA/Doctorado Nacional/2013 folio 21130684). I would like to thank also to the institutions that have supported this research and my development as a researcher. Thank you to the Department of Mathematics Engineering of the University of Chile and the Laboratory of Educational Sciences of University Grenoble Alpes. In addition, this thesis received partial funding from the basal projects from the Center for Advanced Research in Education, and the Center of Mathematical Modeling, all of them from the University of Chile.

La segunda parte de la tesis fue realizada gracias al apoyo de la Escuela de Verano de la Universidad de Chile (EdV), sus directivos, administrativos, profesores y alumnos. En particular me gustaría agradecer a Alejandra Ávila y Lissette Astorga por su apoyo en este proyecto y por las enriquecedoras discusiones que tuvimos y por lo que pudimos aprender en conjunto. Para financiar la recolección de datos

en la EdV contamos con el apoyo del Proyecto Basal FB0003 del Programa de Investigación asociativa de CONICYT.

En la tercera parte de esta tesis se utilizó como fuente de información las bases de datos de la Agencia de Calidad de la Educación. Agradezco a esta institución el acceso a la información. Todos los resultados del estudio son de mi responsabilidad y en nada comprometen a dicha institución.

Para la realización de los análisis estadísticos, esta tesis fue parcialmente apoyada por la infraestructura de supercómputo del NLHPC (ECM-02,Powered@NLHPC) del Centro de Modelamiento Matemático.

Agradezco a todos los funcionarios del Departamento de Ingeniería Matemática de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, especialmente a Silvia Mariano, Eterin Jaña, Luis Mella, Oscar Mori y Georgina Paredes.

También me gustaría agradecer a investigadores que en diversos momentos de mi tesis me aconsejaron y ayudaron a continuar. En particular a Paula Flores Zuñiga, Emilio Vilches, Elisa Giaconi, Valeska Grau, Juan Pablo Valenzuela, Cristian Bellei, Andrew Hart, Juan Carlos Maureira, Sandra Astete y Jean-Marc Montanier.

En este camino he compartido con mucha gente maravillosa y estoy muy agradecida por ello. A mis colegas de ARPA: muchas gracias por todos los buenos momentos y lo que hemos podido desarrollar juntos. Para mi ha sido muy importante ser parte del equipo. En especial gracias a Pepi, Armando, Yoya, Carlos, Sergio, Lisa, Farzaneh y Luz.
Tambien me gustaría agradecer a mis amigos chilenos y franceses: Sandra, JM, Pedro, Pita, Riffo, Anne, Johan, Dani e Ignacio. A mis compañeros del doctorado en Chile, en especial a Evelyn, Roberto, Hugo y Rodrigo. A mis amigas Cata C., Cata Z., Trini y Dani Olivares, muchas gracias. También a mi familia de Burdeos, fue maravilloso reencontrarme con ustedes, en especial a Isabelle, Michotte y Hubert. Merci beaucoup à tous mes amies-potes-collègues de France, j'ai passé des moments magnifiques de ma vie en France grâce à vous. En particulier, merci à Jessica, Fernando, Ludivine, Olga, Laurent, Émeline, Carole, Céline, Rene, Jeremy, Isabelle et Madame Lescure. Aussi, un grand merci à mes amis grenoblois Lolita, Christian et Charlotte.

Finalmente me gustaría agradecer a mi familia, que durante todo este camino me han acompañado y dado su cariño infinito. Especialmente a mi mamá, papá, Eli, noni y tía Pitina quienes me visitaron, me acompañaron diaramente por el internet, siempre han estado conmigo. Los quiero mucho. Por último, infinitas gracias a mi amor Emilio Vilches, por su apoyo incondicional en esta aventura.

Contents

List of Tables

List of Figures

Introduction

Chapter 1

General Introduction

This thesis is devoted to the statistical modeling of the impact of academic environments on student's beliefs and achievement in Chile. Academic environment is a very general term that can refers to classes, courses, families, schools, etc. What they have in common, is that students nested in the same academic environment tend to have similar characteristics. In educational sciences we are often interested in the relation between these environments and individual outcomes. When the objective is to statistically model the effect of these environments, multilevel models are a flexible tool which allows us to properly consider the clustered structure and to investigate the effect of aggregated and integral cluster variables (Bressoux, 2007). Also, multilevel models include several extensions that permits to model longitudinal data and complex clustering structures (Bressoux, 2010).

Multilevel models allow us to answer causal questions about the effect of a treatment on clustered data. Nevertheless, very often, it is not possible to do randomized studies and we need to use observational studies. This raises selection bias problems. Usually, confounding variables which are correlated with the treatment and with the outcome are the source of selection bias. In this thesis, we will consider two methods to reduce selection bias: linear adjusting and propensity score matching. Both strategies work under the hypothesis that all confounding variables are observed.
Linearly adjusting for confounding variables is a classical approach in social sciences. Propensity score matching implies balancing the treated and control samples on the confounding variables, and then comparing the outcomes across the balanced samples. Moreover, these strategies can be used simultaneously, giving more robust conclusions (Stuart, 2010). We choose these methods because they are very popular in the social sciences and address the selection bias problem in different ways.
In addition, missing data is a common problem. Simple methods such complete-case analysis and single imputation are not recommended because they can produce biased estimates and loss of power. We deal with missing data by using multiple imputation. This method imputes several values for each missing value. We choose it because it is a recommended method when the missing data distribution depends on observed values (Lang & Little, 2016; Peugh & Enders, 2004).

Even if the methods for selection bias and missing data are compatible with multilevel models, the problem of addressing both issues simultaneously with hierarchical data is not resolved. This is why we present a statistical discussion regarding how to combine multilevel modeling with selection bias and missing data problems. In addition, we present two empirical studies from educational sciences that illustrate the application of the statistical methods.
The first empirical study regards the influence of Life and Non-Life science courses in secondary students' epistemic and self-efficacy beliefs related to sciences. The aim of the study is to measure the effect of science summer courses about life objects (e.g. Anatomy) and about non-life objects (e.g. Mathematics) on science epistemic and science self-efficacy beliefs.
The second empirical study concerns school effects on academic trajectories of Chilean students and it has two aims. The first aim is to describe the characteristics of the achievement trajectories in mathematics and reading of Chilean students and the influences of primary and secondary schools on those trajectories. The second aim is to measure the effect of public schools in comparison with voucher schools in students' trajectories in mathematics and reading test scores.

The empirical studies illustrate the use of the strategies described in the statistical part. They have clustered data, selection bias and missing data. The need for multilevel models is clear, because the research questions regard the causal effects of academic environments, that is clustered structures. In the first study, the environments are sciences courses and in the second study schools. In addition, in both studies there is a considerable amount of missing data. Equally important is the fact that both studies are observational. Thus, there is selection bias. Considering the summer science courses the selection bias is moderate. However, for the study regarding the comparison of voucher and public schools there is a large amount of selection bias, produced by several selection mechanisms that operate in the Chilean school system (e.g. family selection of schools, schools' fees). In particular, the empirical studies point out the need to combine the statistical methods.

The thesis is organized in three parts. In Part I, we present the statistical part where we analyze how to combine multilevel models with linear adjusting, propensity score matching and multiple imputation. It is presented before the empirical studies because it also justifies the methodological choices made in those studies.
First, we present an introduction chapter to motivate the study of multilevel models, linear adjustment, propensity score matching and multiple imputation. Then, we present a chapter with the theoretical framework where we describe each method individually. In particular, we focus in the assumptions and in the possibilities offered by each method.
In the following chapter, we present the main discussion regarding how to simultaneously modeling selection bias and missing data in the context of multilevel data. In order to develop this discussion, we analyzed and classify recent literature. For multilevel and selection bias we describe how linear adjustment and propensity score matching can be implemented. When linear adjustment for confounding variables is used in the context of multilevel models, more issues emerge regarding the hy-

potheses and there are more strategies to control for selection bias (Cheslock & Rios-Aguilar, 2011; Hill, 2013). Regarding propensity score matching, several possibilities allows us to consider the clustering. These possibilities are classified and discussed. For multiple imputation applied to multilevel data, we describe recent results and implementation issues. Regarding the combination of multilevel models with selection bias methods and multiple imputation we propose three strategies and describe their advantages and disadvantages.
Finally, the last chapter summarizes the proposed strategies and highlights open statistical problems.

In Part II of the thesis we present study the impact of life and non-life sciences courses on secondary students' epistemic and self-efficacy beliefs related to sciences. We present the study as follows. In a first chapter we give the rationale of the research and a literature review. There are two main points that motivate the study. The first point is related to the relevance of epistemic and self-efficacy beliefs based on its impact on achievement, motivational beliefs and learning strategies (Deng, Chen, Tsai, & Chai, 2011; Mason, Boscolo, Tornatora, & Ronconi, 2013; Tsai, Jessie Ho, Liang, & Lin, 2011) and that their development can be seen as an educational goal in itself (Lederman, 2007). The second point regards the usefulness of explore the effect of life and non-life disciplines. This discipline' distinction has not received attention in science education, but we think that can be valuable to understand several processes. In the literature review we describe science epistemic and self-efficacy beliefs and previous results regarding the influence of academic climates.
In the following chapter we present the methodology. To develop the study, we applied a beliefs questionnaire in 50 sciences courses developed by the Summer School of the University of Chile. This institution offers summer courses to secondary students in order to approach the students to rich learning experiences in the university. We applied the questionnaire previous to the courses, at the end of the courses and in a follow-up questionnaire about science epistemic and self-efficacy beliefs. The control group was defined as the group of applicants that did not enroll in the courses. Socio-demographic and academic variables were used to adjust for selection bias. Multilevel and linear regression were used to model contextual effects and linearly adjust for selection bias. Multiple imputation was use to handle missing data.
Then, we present a chapter with the results. We detail the psychometric analysis done to the beliefs questionnaire and the results of the estimated multilevel and regression models used for the estimation of the effect of scientific disciplines, laboratory work and compositional beliefs variables. Finally, we present a discussion to interpret the results and relate them to previous research.

In Part III of the thesis we present the study regarding school effects on students' trajectories in mathematics and reading.
In the introduction we explain why it is relevant to study student trajectories and compare public and voucher schools in Chile. Studying trajectories improves the estimation process because it accounts for the measurement error, allowing us to study the shape of the change and the rate of the growth (Bressoux, 2010). Also studying the Chilean system raises academic and political interest because it is a

very stratified and segregated system, which has a large amount of voucher schools (Bellei, 2008; Valenzuela, Bellei, & De los Ríos, 2014).
After the introduction, we present a theoretical framework where we describe school effects and their definition, the Chilean educational context and studies that have compared voucher and public education in Chile. This serves as a precedent for our results and justifies methodological choices.
In the following chapter, we describe the methodology. We detail the sample, variables and the statistical models. The sample came from a national data base that includes standardized tests scores and background student variables as well as school level data. We analyzed mathematics and reading scores from the same students at 4th, 8th and 10th grade. In order to model the hierarchical structure, we used multilevel growth models and a cross-classified structure to account for student mobility between schools. Linear adjusting, propensity score matching and multiple imputation techniques were applied for adjust to selection bias and missing data.
The following chapter describes the results. First, we analyze how the variance of the test scores is distributed between the occasion, student and school level and we describe average students' trajectories for different groups. Then, we compare trajectories of students from the voucher system with students from the public system using different strategies to control for selection bias.
The last chapter of Part III presents a discussion that puts together the results, considers their implications for Chilean policy and proposes further lines of research.

The final chapter of the thesis is a discussion which synthesize and connects the main of results of the three parts of the thesis. In addition, it offers research perspectives.

Part I

Multilevel modeling dealing with selection bias and missing data in educational sciences

Chapter 2

Introduction

Research in social sciences, in particular in educational sciences, needs to address several specific statistical issues that arise because of the type of data that is collected and the type of research questions that are addressed. In this chapter, we will describe statistical methods to study data that have a hierarchical or multilevel structure in the context of research questions about causal effects where there are selection bias and missing data.

Multilevel data are relevant in the context of educational sciences because they arise very often and their modeling allows us to study unique research questions. For example, questions about the effect of different courses or about school effects are answered with data that have a hierarchical structure because students are nested in courses and schools. In addition, multilevel data includes longitudinal data, because repeated measures can be seen as nested in the individual, generating another level.

Regarding the modeling of the hierarchical structure of the data, multilevel linear models allow us to study and model specific characteristics of the data. For example, we can model how the variance is distributed across different levels and measure the effect of variables at different levels (Bressoux, 2007). They provide flexibility to model longitudinal data and complex clustering as in cross-classified and multiple membership models. An example of the relevance of these models in educational sciences is the research on school effectiveness (Reynolds et al., 2014).

In addition to the modeling of the structure, we have to focus on research questions. In this part of the thesis, we want to describe techniques to answer what is the effect of a treatment on an outcome. In several cases, these questions cannot be addressed with randomized studies, and have to be done with observational studies. In this context, two methodological problems appear:

1. Selection bias
2. Missing data

Selection bias is defined as the bias produced by non random assignment of the treatment or non random sampling. For example, an important question is which type of schools is more effective: private or public schools. Usually it is not possible, neither ethical, to assign students randomly to different types of schools. But, we cannot just compare students from different schools, because they can differ systematically on other variables, which impedes to measure the effect of the school' type.

Two methods to reduce selection bias will be described and analyzed: linear adjusting and propensity score matching. We will describe both methods under the Rubin framework for causal inference (Imbens & Rubin, 2015). We choose linear adjusting because it is one of the most popular methods in social sciences. Also, with multilevel models it is straightforward to model the multilevel structure of the data and at the same time linearly adjusting for control variables to reduce selection bias.
On the other hand, we also analyzed propensity score matching because it is based on balancing the treatment and control samples, which is an alternative strategy compatible with regression analysis because it is possible to use both techniques simultaneously. This method allows us to do a parallel with randomized studies and to evaluate whether the samples are comparable or not without using the outcome, which is more scientifically sound (Rubin, 2007). In addition, propensity score analysis has become very popular. Thoemmes and Kim (2011) made a review of social science articles that used propensity score. They found that the number of articles using propensity score increased exponentially and that, in the articles they review, 39.5% were in educational sciences.

Regarding missing data, it refers to the problem where there is missing information of the units in the sample in one or more variables. Missing data can cause biased estimates and loss of statistical power. The relevance of the missing data problem depends on the missing data mechanism and the amount of missing data. Simple methods for treating missing data are not recommended. In addition, usually there are several patterns of missing data in variables with different characteristics. This is why we choose multiple imputation, which is a method recommended in the literature for data missing at random (Lang & Little, 2016; Peugh & Enders, 2004) and, at the same time, it is very flexible to model several types of variables. Furthermore, it can be combined with different types of analysis, in particular with multilevel models.

Even if all the methods for selection bias and missing data are compatible with multilevel models, the problem of addressing both issues simultaneously with hierarchical data is not resolved. From a theoretical level, assumptions of each method should be adapted to the multilevel case. A main point is that the presence of selection bias and missingness in multilevel data can be related to variables that vary at different levels and that take different roles in the model, for example outcome variables, group identification variables, etc. This raises new questions and motivates this part of the thesis, which aims at describing methods for modeling multilevel data and answering causal questions with the presence of missing data and selection bias. We provide a theoretical discussion about how to combine the methods. This discussion is also useful to understand the methodological decisions that were taken in the empirical studies presented in part II and part III of the thesis.

Chapter 3

Theoretical Framework: Description of multilevel models and methods for selection bias and missing data

In this chapter we describe multilevel models for different types of hierarchical data, regression analysis and propensity score analysis for overcoming selection bias and multiple imputation for addressing missing data problems. First, we describe the statistical notation and terminology and then we describe the statistical models.

3.1 Preliminaries and statistical notation

In this section, we describe general assumptions and the statistical notation and terminology, which are summarized in Table 3.1.
Throughout this chapter, we will assume that we have a continuous dependent or outcome variable, and we want to estimate the effect of a binary treatment. We will name the treatment values *active treatment* and *control treatment.* Explanatory variables or covariates can be of different natures (categorical, continuous, ordinal, etc.).

The dependent or outcome variable is denoted y. We want to measure the effect of the treatment on y. This is the observed outcome, but for the analysis related to selection bias, it is also necessary to define the potential outcomes. The potential outcomes are the values of the outcome in the two possible treatments. We will denote them y^1 and y^0 , where y^1 is the value of the outcome in the case that the unit received the active treatment and y^0 is the value of the outcome in the case the unit received the control treatment.
The treatment assignment variable will always be denoted z, where $z_i = 1$ if unit i

received the active treatment and $z_i = 0$ if unit i received the control treatment, y can be defined as:

$$y_i = z_i y_i^1 + (1 - z_i) y_i^0 \quad \forall\ i$$

Also in the framework of selection bias, we distinguish in the explanatory variables the treatment assignment variable z and the control variables or covariates denoted by $X = (x^1, \dots, x^k)$ (see Table 3.1). In this framework, we will assume that we know a priori that the covariates X are not affected by the treatment assignment. These covariates allow us to: explain some variation in outcomes; describe the sample and model the assignment mechanism (Imbens & Rubin, 2015, p. 16).
For clarity, in linear models we will express the multiplication of the parameters with the covariates as vector multiplications. For example, if the parameters of a linear regression are $\vec{\beta} = (\beta_1, \dots, \beta_k)^t$, we will use the following expression:

$$X\vec{\beta} = \beta_1 x^1 + \cdots + \beta_k x^k$$

.

For the description of models concerning missing data, we will denote $Y = (Y_{rs})$ the matrix of dimensions $N \times (k+2)$ with all the observed data. N is the sample size and $k+2$ is the total number of observed variables. Each row is a unit and each column is an observed variable. Y contains the dependent variable y, the treatment variable z and the explanatory variables X. We denote $M = (M_{rs})$ the matrix with missing indicators, that is, $M_{rs} = 1$ if Y_{rs} is missing and $M_{rs} = 0$ if Y_{rs} is observed.

In the notation for multilevel models, subscripts will be used to clarify the level at which each variable changes. For example, if we used the letter i as the index to identify students, and j as the index to identify schools. Then, x_{ij} is the value of a variable for student i who belongs to school j, implying that x varies at the student level. If instead, we define x_j, it implies that x varies at the school level only. The same goes for random effects, which will be denoted with Greek letters.

For estimating multilevel models, particularly to estimate random effects at higher levels of the data structure, it is necessary to define which unit belongs to which cluster or group. In order to do this in the estimation process, we use identification or grouping variables for clusters, for example school and class identifiers. These variables are very relevant, especially for the missing data discussion in the next chapter. Consequently, we include them in Table 3.1. In general, the effect of higher level variables in outcomes at the first level are called contextual effects. However, there is a terminology that will be used for higher level variables in multilevel models, that allows to distinguish different types of contextual effects (Diez, 2002):

1. Aggregated or derived variables: Variables that are defined at higher levels, but are calculated with variables from lower levels. Usually they can characterize the mean level, for example in a study of student nested in science courses, an aggregate variable could be the mean of the students grades in each course. The effect of these variables in outcomes at the first level are called compositional effects (Diez, 2002).

2. Integral variables: Variables that are defined at the cluster level and are a cluster level construct. They are not aggregated variables. For example, in a study of students nested in science courses, an integral variable could be the type of discipline of the course (Diez, 2002).

Table 3.1: Description of the variable notation and terminology in each model.

Variable notation	Variable description		
	Multilevel models	Missing Data	Selection Bias
y	dependent	dependent	outcome
y^1	dependent	dependent	outcome
y^0	dependent	dependent	outcome
z	explanatory	explanatory	z treatment
$X = (x^1, \dots, x^k)$	explanatory	explanatory	X control or co-variates
$Y = (y, z, X)$		observed variables	
M		missing data pattern in Y	
ε, μ, etc.	random effects		
$IDschool$, $IDclass$, etc.	grouping	grouping	

3.2 Modeling the data structure: Multilevel models for cross-sectional and longitudinal data

In this section, we describe multilevel models. In particular, we describe the multilevel models used and discussed in this thesis. We present the general multilevel model, cross-classified models, multiple membership models and longitudinal models. Multilevel models are introduced as an extension of the linear regression model. The hypotheses of regression models are stated because they are important in the section about selection bias.

3.2.1 Multilevel models as extensions of the regression model

Multilevel models are linear models that allow us to model a dependent variable measured in units nested in groups. The typical example is students nested in classes or schools. The base of multilevel models is the linear regression model. In fact, multilevel models can be seen as an extension of this model and they share several assumptions. This is why, this section starts with a very brief description of the multivariate regression model and its hypotheses. The regression model proposes linear relations between dependent and explanatory variables. Its expression is presented in equation (3.1), where β_0 to β_k are fixed coefficients that define the linear relations and ε_i is the error term.

$$\begin{aligned} & y_i = \beta_0 + \beta_1 x_i^1 + \cdots + \beta_k x_i^k + \varepsilon_i \\ & \varepsilon_i \sim N(0, \sigma^2) \ \ \forall i \end{aligned} \tag{3.1}$$

In order to have a well specified model and to obtain unbiased and consistent estimators, the following hypothesis should be attained (Bressoux, 2010, p. 101):

- H1: Linearity and additivity

$$\mathbb{E}(y|X) = \beta_0 + \beta_1 x_i^1 + \cdots + \beta_k x_i^k \tag{3.2}$$

- H2: y is continous and not bounded. $X = (x^1, \ldots, x^k)$ are fixed and measured without error.
- H3: The expectancy of the error term given X is 0 for all i.

$$\mathbb{E}(\varepsilon_i|X_i) = 0 \ \ \forall i \tag{3.3}$$

- H4: Homoscedasticity of the error terms

$$\text{var}(\varepsilon_i|X_i) = \sigma^2 \tag{3.4}$$

- H5: Independence of the error terms

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j \tag{3.5}$$

- H6: Errors have a normal distribution

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{3.6}$$

- H7: Explanatory variables are linearly independent. It is not possible to calculate one as a linear combination of the others.

These hypotheses for the linear regression model permit to introduce multilevel models, as the ones that relax the hypothesis of independence of the errors (H5) and model the effect of groups or environments. Relaxing this hypothesis is very important in educational settings. When the data come from students nested in classes and schools, equation (3.5) is not plausible, because observations of students that belong to the same school or the same class are not independent.

In Table 3.2 several models are described in the case of data with two levels (for example students nested in schools) to illustrate the flexibility and potential of multilevel modeling. The simplest one is called the empty model where the dependent variable is modeled with a fixed effect β_0, which is the general intercept, and with random effects ε_{ij} and μ_j at the student and school level respectively. The empty model is very important because it allows us to know how the variance is distributed among the different levels. The other models in the table include random intercepts

and also random slopes.
In order to completely specify the models, the distribution of the random effects has to be defined. Usually, it is assumed that they are normally distributed. If there are several random effects at the same level, the assumption is that they have a joint normal distribution.

Multilevel models are not only an adjustment to take account of the non-independence of the observations. The possibility to model random intercepts and slopes and the effect of explanatory variables at different levels permit to address new research questions. In particular, multilevel models allow us to model the effect of the environments in the student level variables (Bressoux, 2007). This feature will be exploited largely in this thesis, notably in parts II and III where the research questions are about the effect of environments in student level outcomes. The first question regards the effect of scientific courses on epistemic and self-efficacy beliefs and the second question is about school' effects on student' trajectories in mathematics and reading scores.

Table 3.2: Multilevel models with increasing complexity.

Model	Random effects distributions	Description
$y_{ji} = \beta_0 + \mu_j + \varepsilon_{ij}$	$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$ $\mu_j \sim N(0, \sigma^2_\mu)$	Empty model
$y_{ji} = \beta_0 + \beta_1 x^1_{ij} + \cdots + \beta_k x^k_{ij} + u_j + \varepsilon_{ij}$ $u_j = \alpha_1 x_j + \mu_j$	$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$ $\mu_j \sim N(0, \sigma^2_\mu)$	Model with random intercepts μ_j. It includes student-level and school level explanatory variables
$y_{ji} = \beta_0 + \beta_{1j} x^1_{ij} + \cdots + \beta_k x^k_{ij} + \mu_j + \varepsilon_{ij}$ $\beta_{1j} = \nu_j$	$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$ $(\mu_j, \nu_j)^\intercal \sim N(0, \Sigma)$	Model with random slopes ν_j and random intercepts μ_j.
$y_{ji} = \beta_0 + \beta_{1j} x^1_{ij} + \cdots + \beta_k x^k_{ij} + \mu_j + \varepsilon_{ij}$ $u_j = \alpha_1 x_j + \mu_j$ $\beta_{1j} = \alpha_2 x_j + \nu_j$	$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$ $(\mu_j, \nu_j)^\intercal \sim N(0, \Sigma)$	Model with random slopes ν_j and random intercepts μ_j. It includes student-level and school level explanatory variables for the intercepts and the slopes.

Naturally, higher levels can appear. For example, students nested in schools which are nested in municipalities. The models with more levels are analogous to the two-level model. There are several other extensions to these models, which allow to relax hypotheses and model different features of the data. In the following sections, we are interested in describing three specific extensions:

1. Multilevel growth models
2. Cross-classified and Multiple-membership models

Multilevel growth models are important because they allow us to model longitudinal data. Cross-classified and multiple-membership models allow us to model complex clustering structures that are not strictly nested. In particular, they are useful to model student mobility in multilevel growth models of students' trajectories (Sun & Pan, 2014). The general multilevel model and their extensions are relevant for the applied problems addressed in parts II and III of this thesis.

3.2.2 Multilevel growth models

Multilevel growth models permit to model observations for the same units that vary along time or along a growing variable. These models have two important characteristics that distinguish them from the classical multilevel model (Bressoux, 2010, pp. 359):

1. Repeated measures for a unit are modeled as nested in the unit. For example, several measures of students' reading scores can be modeled with a two-level model. The first level correspond to the students' repeated measures and the second level correspond the students.
2. The shape of the change along time, or along the growing variable, is modeled.

This first point implies that it is not necessary to have the same quantity of measures by person because, as in any multilevel model, the clusters can have different sizes. Also, it is not necessary to have equally spaced measures. The model includes the growing variable for each person, which can vary freely.

The second point is fundamental, because it implies that just modeling longitudinal data does not imply that we have a growth model. We have to explicitly model the shape of the growth using the growing variable. For example, in the research question from part II of this thesis there are three repeated measures. However, it is not growth modeling because we made separate models for the effects of the courses in the post and follow-up measures and we used the first measure as a control variable. The research question from part III of the thesis is an example of growth modeling, because we want to know how are the students' trajectories and how these trajectories are affected by different variables.

Equation 3.7 presents a simple growth model. To illustrate, we can consider that it is a model over student outcomes. In this equation, for each student i the model assigns the same slope β_1 and a student-specific intercept μ_i, which is a random effect.

$$y_{ti} = \underbrace{\beta_1 t + \beta_0 + \mu_i}_{\text{Modeling linear growth}} + \varepsilon_i + \varepsilon_{ti}. \tag{3.7}$$

The flexibility of multilevel growth models allows us to add random slopes for each student, as in equation (3.8). We can model the random slopes and intercepts with explanatory variables. This enables the study of the rate and shape of the growth. This information informs about processes and cannot be obtained with cross-sectional models. In equation (3.8), we are modeling a linear relation between $y_{\cdot i}$ and t. However, having more measures allows to model more complex curves as polynomials, piece-wise functions, etc.

$$\begin{aligned} y_{ti} &= \beta_{1i} t + \beta_0 + \mu_i + \varepsilon_i + \varepsilon_{ti} \\ \beta_{1i} &= \underbrace{\alpha_0 + \alpha_1 x_i^1 + \nu_i}_{\text{Modeling the rate of the change}} \end{aligned} \tag{3.8}$$

Modeling repeated measures with multilevel models poses challenges and also provides advantages. A first challenge is the need to have comparable measures across time (Hox, 2010, p. 79). A second challenge is that the time scale has to be carefully chosen to produce interpretable parameters and, at the same time, to avoid estimation problems. Regarding the advantages, a first one is the possibility of modeling different growth curves for each individual, which is more aligned with individual development. A second advantage is the possibility of measuring the effect of clusters on individual development and add easily time-variant variables as covariates (Hox, 2010, p. 98).

3.2.3 Cross-classified models and multiple-membership models

There are cases where the grouping of the units is not strictly hierarchical, and other types of structures appear. We will describe here two of these structures: cross-classified and multiple membership.

A typical example of cross-classified structure is students that belong to schools and neighborhoods (see Goldstein (2011, ch. 12) and Hox (2010, ch. 9)). In this case, there is no strict hierarchical structure because students from different neighborhoods can belong to the same school and students from different schools can belong to the same neighborhood. What is present here is a cross-classified structure, where there are two dimensions that produce clustering, but they are not hierarchically related.
Regarding multiple-membership structures, these are structures where one unit can belong to more than one cluster. For example, if the units are students and the clusters are friendship groups (see (Goldstein, 2011, ch. 13)).

A very important case where cross-classified and multiple membership structures can be used is repeated data. An example is the study about the Chilean student trajectories in part III of this thesis. In this study, we analyze students' trajectories in mathematics and reading scores using measures at 4th, 8th and 10th grade. Because students can change school, three types of clustering appear: the school in 4th grade, the school in 8th grade and the school in 10th grade. For this example, the empty cross-classified growth model is expressed in (3.9). It uses a similar notation as in Hox (2010, ch. 9), where the parenthesis (jkl) represent the school indexes for the

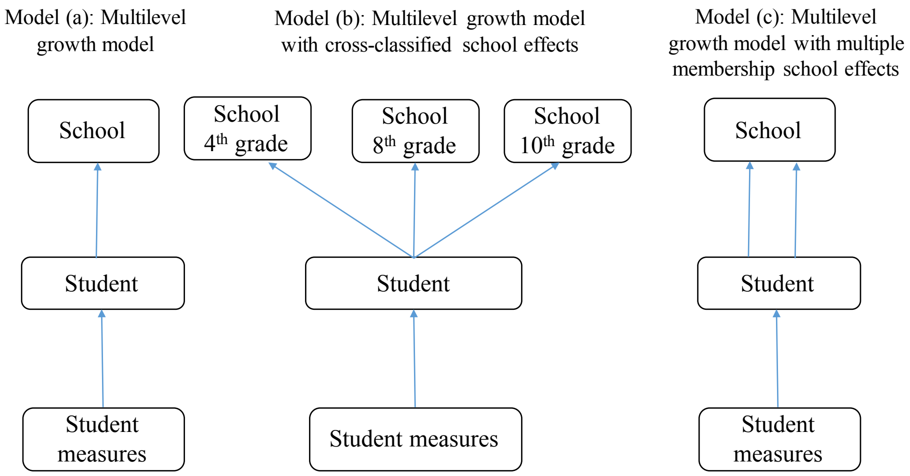

Model (a): Multilevel growth model
Model (b): Multilevel growth model with cross-classified school effects
Model (c): Multilevel growth model with multiple membership school effects
School
School 4th grade
School 8th grade
School 10th grade
School
Student
Student
Student
Student measures
Student measures
Student measures

Figure 3.1: Multilevel growth models

school in 4th grade, 8th grade and 10th grade respectively. The parenthesis indicates that the units i, j and k are not necessarily nested. In particular, equation (3.9) is an example of the modeling of school effects with a cross-classified model where each clustering defines a different random effect.

$$y_{ti(jkl)} = \beta_0 + \beta_{1i(jkl)}t + \underbrace{\mu_j^{4th} + \mu_k^{8th} + \mu_l^{10th}}_{\text{schools effects}} + \varepsilon_{i(jkl)} + \varepsilon_{ti(jkl)} \tag{3.9}$$

In the case of multiple membership, the approach is different. Here, we use weights to model that a student can belong to different clusters, in this case schools. For example, if a student attended School A in 4th grade and attended School B between 8th and 10th grade, their weights could be $\frac{1}{3}$ for School A, $\frac{2}{3}$ for School B, and 0 for the rest of schools. An example of a multiple membership model is described in equation (3.10).

$$y_{ti} = \beta_0 + \beta_{1i}t + \underbrace{\sum_{j=1}^{N_J} \omega_{ij}\mu_j}_{\text{school effects}} + \varepsilon_i + \varepsilon_{ti} \tag{3.10}$$

Using notation from Browne, Goldstein, and Rasbash (2001), Figure 3.1 represent the multilevel growth model, in the case there is no cross-classification of schools (a), in the case there is cross-classification of schools (b) and in the case of multiple membership (c). It is possible to define random intercepts and slopes in the same way as with regular multilevel models. The importance of these models is that they allow to properly account for the non-nested clustering.

3.3 Modeling selection bias: Linear regression and Propensity Score models

3.3.1 Causal Estimation

To estimate the effect of a treatment, we need to know the difference in the value of an outcome between the units in the case they received the active treatment and the same units in the case they received the control treatment. That means answering the question: What would have happened if the unit received another value of the treatment?. In order to know this, we should see the outcome in case the unit was in the treatment and in case the unit was in the control group, that means observe both potential outcomes.
The problem is that a unit can only receive one treatment (the factual), so the value of the outcome in the case she received the other treatment (the counterfactual) is missing. This is the fundamental problem of causal inference (Rubin, 1974).
These ideas are part of the conceptual framework for causal inference developed by Rubin (Imbens & Rubin, 2015). An important issue is that this framework is useful if the counterfactual can be well defined (Cameron & Trivedi, 2005, pp. 34), which implies it has well defined treatments. Imbens and Rubin (2015, pp. 4-5) discuss the relevance of having a well defined treatment variable, where the alternative to the active treatment has to be the control treatment. For example, if the active treatment is taking an aspirin, the control treatment can be defined as not taking an aspirin and the potential outcomes in both cases are well defined, so the causal effect can be defined. In many cases, to define the possible treatments is not obvious and the treatments and the population under study have to be characterized carefully to be able to do causal inference.

If the active and control treatment are well defined, we can state that, since it is impossible to observe the outcome of a unit in the control and in the treatment group, it is necessary to assign the unit to one group, the active or the control treatment. If the assignment mechanism is not random, it can introduce bias to the estimation of the effect of the treatment. In this case, observed or unobserved factors can produce systematic differences between the treatment groups. We will consider the selection bias produced as a result of not having a randomized treatment assignment. The term selection bias is also used for non-random sampling, but in this section we assume that we have random sampling.

In this context, the best way to study the effect of a treatment is doing randomized studies (Imbens & Rubin, 2015; Rubin, 1974). In these studies, the assignment to each possible value of the treatment is done in a random way, were we know the probability of assignment (Imbens & Rubin, 2015, p. 40). Random assignment can balance observed and unobserved variables, if the sample is large enough.

Nevertheless, in educational sciences there are several cases were it is not possible to assign randomly individuals to the treatment and control group. In these cases, treatment effect estimation has to be done in observational studies. In this section, we use the conceptual framework from Imbens and Rubin (2015) to describe the statistical techniques to estimate treatment effects in observational studies, where selection bias appears naturally.

We consider the case of an observational study, where we have a random sample from each treatment group, where N_1 is the sample size of the active treatment group, N_0 is the sample size of the control group and the total sample size is N. The potential outcomes are y_1 and y_0. Our objective is to estimate the average treatment effect (ATE) defined as:

$$ATE = \mathbb{E}(y_1) - \mathbb{E}(y_0)$$

It is important to note that the ATE is defined trough the potential outcomes, and that the natural sample estimation that we would like to do is the difference of means:

$$\widehat{ATE}_1 = \frac{1}{N}\sum_{i=1}^{N} y_{1i} - \frac{1}{N}\sum_{i=1}^{N} y_{0i}$$

But this estimation is impossible, because for each unit i, we can only observe one of the potential outcomes. This mean that the quantity that we can estimate is:

$$\widehat{ATE}_2 = \frac{1}{N_1}\sum_{z_i=1} y_{1i} - \frac{1}{N_0}\sum_{z_i=0}^{N} y_{0i}$$

If there is a random experiment, $\widehat{ATE}_2$ is an unbiased estimator of ATE. The problem is that it can produce biased estimates if the treatment assignment mechanism is not random. However, if there are measured pre-treatment variables X, which accomplish that the potential outcomes are independent of the treatment conditional on X, we can estimate ATE (Rosenbaum & Rubin, 1983). The precise assumptions over the pre-treatment variables are stated in equations (3.11) and (3.12). These assumptions allow us to estimate ATE with propensity score analysis and, adding some linearity hypothesis, with regression analysis (Rosenbaum & Rubin, 1983).

$$(y_1, y_0) \perp z|X \tag{3.11}$$

$$0 < \mathbb{P}(z_i = 1|X_i) \ \forall i \tag{3.12}$$

Assumption from equation (3.11) is called the ignorability assumption or uncounfoundedness. This assumption cannot be tested and we have to consider what we know from our empirical problem to evaluate its plausibility (Wooldridge, 2010, p.

910). Assumption from equation (3.12) implies that all the units have non-zero probability of being in the active and in the control treatment. This assumption is natural because it is not possible to define a potential outcome if there is no possibility that the unit gets a defined treatment and is also referred as the need to have common support. Equations (3.11) and (3.12) form the strong ignorability assumption (Rosenbaum & Rubin, 1983).

Another important assumption is the stable unit treatment value assumption (SUTVA), which says that the treatment assigned to one unit does not have influence in the outcomes of the other units (Imbens & Rubin, 2015, p. 10). If SUTVA does not apply, we should consider more treatments and the problem is intractable. For example, consider that the outcome of unit 1 depends on where the unit 2 was assigned. In this case, we should consider four potential outcomes which are all the treatment assignment combinations for unit 1 and 2 . But this prevents us from estimating the treatment effect, because in some sense each assignment of units is a different treatment. In this case, the counterfactuals and potential outcomes are not well defined.

In the following subsections, we describe how to estimate the ATE with regression based models and propensity score analysis. If the strong ignorability assumption is met both analyses can give an unbiased estimate of the ATE.

3.3.2 Regression Models and Selection Bias

The main idea for using linear regression for estimating treatment effects, is that linearly adjusting of the effect of control variables in the outcome can isolate the effect of the treatment variable. In this section, we will describe regression analysis used for treatment effect estimation in the classical regression framework and in the potential outcomes framework. In order to make causal claims through regression analysis, hypotheses have to be done over the variables. First, we need a treatment variable z, where the potential outcomes are correctly defined and there is an understanding of the assignment mechanism in order to control for the proper covariates. In addition, we have to accomplish the regression hypotheses in order to have appropriate estimations. The basic assumption for using regression analysis in the context of observational studies, is that we can correct for selection bias adding the pre-treatment variables to the regression equation. In the simplest case, where there is no selection bias, the regression model to estimate the average effect of a treatment z in an outcome y is defined as follows:

$$y_i = \beta_0 + \beta_1 z_1 + \varepsilon_i \tag{3.13}$$

If there is selection bias, we try to adjust for it controlling by relevant covariates, as follows:

$$y_i = \beta_0 + \beta_1 z_1 + \vec{\beta}_2 X_i + \varepsilon_i \tag{3.14}$$

In these models the average treatment effect is β_1, and the estimated average treatment effect is $\hat{\beta}_1$, the estimated value of this coefficient. To claim that $\hat{\beta}_1$ is an unbiased estimator of the treatment effect, there are two possibilities to argue and define the necessary hypotheses.

In the classical regression framework, we can ask that the usual hypotheses for linear regression be met (Bressoux, 2010, p. 101). There are two hypotheses that are especially critical in causal effect estimation. The first one is hypothesis H3 (equation 3.3) that states $\mathbb{E}(\varepsilon_i|z, X_i) = 0$. This hypothesis implies that $\mathrm{cov}(\varepsilon_i, (z_i, X_i)) = 0$ and this latter property is enough for having unbiased estimates in the regression coefficients (Wooldridge, 2010, p. 54). The hypothesis H3 is stronger than $\mathrm{cov}(\varepsilon_i, (z_i, X_i)) = 0$, but is better because it assures us that $\mathrm{cov}(\varepsilon_i, g(z_i, X_i)) = 0$ with any function g. This is very useful for including nonlinear functions of X_i in the regression (Wooldridge, 2010, p. 18) .
If a variable x is correlated with the error term ε, we say that it is endogeneous, if it is uncorrelated, we say that it is exogeneous (Wooldridge, 2010, p. 54). The endogeneity appears in three forms: omitted variables , measurement error and simultaneity (Wooldridge, 2010, p. 54,55). Usually, the most important for treatment effect estimation are the omitted variables that are correlated with the outcome y and the treatment assignment z, which can produce $\mathrm{cov}(\varepsilon_i, z_i) \neq 0$ and consequently biased estimates. The second relevant hypothesis is H1 regarding linearity and additivity. In particular, this model assumes that linearly adjusting for the confounding variables is enough. Nevertheless, this is a strong supposition because there can be another types of relations, more complex that cannot be modeled with a linear model.

Linear regression can also be studied from the potential outcomes framework. In this framework, using Corollary 4.3 from Rosenbaum and Rubin (1983) we can estimate the treatment effect using linear regression. Rosenbaum and Rubin (1983) stated the corollary in terms of balancing scores. A balancing score is any score that accomplish equation (3.15).

$$X \perp z|b(x) \tag{3.15}$$

In particular $b(X) = X$ is a balancing score. Therefore, we can infer that if the strong ignorability assumption is true and the conditional expectation of the potential outcomes is linear in X. Then, the estimator from equation (3.17) is an unbiased estimator for the average treatment effect if the units in the study are a random sample (Rosenbaum & Rubin, 1983).

$$\mathbb{E}(y^g|z = g, X) = \alpha_g + \vec{\beta}_g X \; g = 1, 0 \tag{3.16}$$

$$\widehat{ATE}' = \hat{\alpha}_1 - \hat{\alpha}_0 + (\hat{\vec{\beta}}_1 - \hat{\vec{\beta}}_0) \left(\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \right) \tag{3.17}$$

The estimator $\widehat{ATE}'$ is not exactly what is generated by linear regression, but if we add the assumption that $\vec{\beta}_1 = \vec{\beta}_0$, it is equivalent.

In summary, regression analysis has as a limitation the strong hypothesis of linearity. Regarding the other hypotheses, they are related because the strong ignorability assumption and hypothesis H3 from classical regression analysis ask for the treatment to not be confounded with unmeasured variables. In conclusion, it is important to identify empirically which are the covariates that define or influence the treatment assignment and measure them.
Nevertheless, regression analysis also presents advantages. For example, there can be an interest in understanding the effect of the covariates and in this case estimating $\vec{\beta}_2$ from equation (3.14) is informative. In addition, extensions of the linear regression model allow to work easily with outcome variables with different distributions, for example logistic regression and with clustered data as in multilevel models.

3.3.3 Propensity Score Analysis

Propensity score analysis is a technique that is based on balancing the samples that received different treatments to be able to compare them. The basic idea is that we can correct for selection if we make the different treatment groups comparable on the pre-treatment variables or covariates. After the samples are equivalent, the effect of the treatment can be easily estimated, comparing the outcome variable in the balanced samples.
Usually, there are several pre-treatment variables or covariates that need to be balanced in the sample, which make a challenge to use them all. This problem is solved, because Rosenbaum and Rubin (1983) proved that if the strong ignorability assumption is true it is enough to balance the sample on one variable: the propensity score. The propensity score is the probability of belonging to the active treatment conditional on the covariates, defined as $\mathbb{P}(z = 1|X)$.

Rosenbaum and Rubin (1983) showed that, under the strong ignorability assumption, the mean difference between the samples balanced through the propensity score was an unbiased estimate of the average treatment effect. Also that balancing the samples through propensity score produced that treated and control groups have the same distribution of the pre-treatment variables X.

To balance the samples of the different treatments, the first step is to estimate the propensity score; the second step is to balance the samples to achieve the same propensity score distribution; and then to compare the outcomes in the balanced samples. The first step is finding a proper model for estimating the propensity score and explicit which variables were included and why. Then, researchers have to check that a good balance was achieved and define the region of common support, that is where the distribution of the propensity scores for the treatment groups overlaps, that means that have units with comparable propensity scores, and finally report how outcomes of the balanced samples where compared (Thoemmes & Kim, 2011).
To estimate the propensity score, it is necessary to use models that explain the probability to belong to the active treatment with pre-treatment variables. Typical choices are logit and probit regression models (Thoemmes & Kim, 2011). Neverthe-

less, there exist other models to estimate this probability, for example Generalized Boosted Models (GBM), which have several appealing properties such as: selecting the variables to estimate the propensity score maximizing the achieved balance and avoiding distributional assumptions as probit and logit (McCaffrey, Ridgeway, & Morral, 2004), (McCaffrey et al., 2013).
To balance the samples, it is possible to do: matching, weighting, stratification and covariate adjustment (Clark, 2015). Matching strategies imply select units in the active treatment and control group that have the same or very similar propensity scores. When it is the same it is exact matching and when it is similar is approximate matching (Thoemmes & Kim, 2011). There are several decisions in order to choose the matching strategy. Thoemmes and Kim (2011) distinguish the following factors: the number of treated units matched to one control unit, the matching (exact or approximate) and the algorithm (optimal or greedy). In addition, it is possible to match the propensity scores and observed covariates that are especially relevant (Stuart, 2010).
Weighting implies giving different weights to each unit in order to have weighted balanced samples. The weights are calculated with the propensity scores and the typical weighting strategy is inverse probability weighting (Clark, 2015). Stratification (or subclassification) implies the division of the sample in strata according to the propensity score. After the division, control and treated units are compared in each stratum.
Finally, covariate adjustment uses the propensity score as a control variable in a regression model. This strategy is reasonably because the propensity score can be seen as a data reduction of all the control variables. In addition, it gives unbiased estimates under strong ignorability assumption and linearity assumption (Rosenbaum & Rubin, 1983, corollary 4.3). However, it is not recommended because it adds assumptions (Thoemmes & Kim, 2011).

An important point is that propensity score can be combined with linear adjustment. This is a recommended procedure because it is enough that one of the methods be properly defined to have well estimated treatment effects. Also it has been demonstrated that combined work best (Stuart, 2010).

3.4 Modeling Missing Data: Multiple imputation

3.4.1 Missing data mechanism

Missing data is an unavoidable problem in most applied research. The simplest methods to address this problem are complete case analysis and single imputation. Complete case analysis implies erase the units that have missing values in one or more variables. The advantage is that it is very easy to implement and easy to combine with other statistical models, which can be applied without extra data management. The main disadvantages are the loss of information and statistical power and that

depending on the missing data mechanism, it can produce biased results (Little & Rubin, 2002). The loss of information is especially critical in studies with several measured variables, where it is more probably for a unit have a missing value in one of the variables.
Another simple method is to do single imputation, that means impute the missing value with a plausible value and then make the same data analysis that would be done with complete data. There are several techniques to generate the single imputations, the simplest is to impute missing values with the mean of the variable. Single imputation strategies are not recommended because they can bias the results, specially variance and covariance estimators. In addition, these strategies do not account for the uncertainty of the missing values and treat the imputed values as regular values (Allison, 2002, p. 28).

The adequate method to handle missing data depends on several factors such as: the missing data mechanism, the amount of missing data, the type of data and the research question. The most important factor is which is the missing data mechanism. These mechanisms were defined by Rubin (1976). If they are not taken into consideration there can be biased estimates. This section is strongly based on Little and Rubin (2002, p. 11,12). In order to define the missing data mechanisms, we use the notation from Table 3.1 and we set some extra statistical notation. We denote Y_{mis} and Y_{obs} the parts of the Y matrix with missing and observed values respectively. The vector ϕ is a vector of unknown parameters. To define the missing data mechanism, the key function is the distribution of the matrix of missing indicators M, conditional on Y and on ϕ: $f(M|Y,\phi)$. The missing data mechanism is defined through this distribution (Little & Rubin, 2002, p. 11,12):

- Missing completely at random (MCAR):

$$f(M|Y,\phi) = f(M|\phi) \ \forall \ Y, \phi \tag{3.18}$$

- Missing at random (MAR):

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \ \forall \ Y_{mis}, \phi \tag{3.19}$$

- Missing not at random (NMAR): When $f(M|Y,\phi)$ depends on Y_{obs} and Y_{mis}

In the case of data MCAR, the distribution of missing data does not depend on the observed variables. This implies that in this case complete-data analysis would not bias the results. In the case of MAR, the distribution of missing data depends only on observed values. This gives the chance that, using these observed values, we can solve the missing data problem. Finally, in the case of data NMAR, the distribution of missing data depends on Y_{mis}, which by definition is information that is not available.

3.4.2 Multiple imputation

The multiple imputation technique is based on imputing several values for each missing value. Each imputed value belongs to a different complete data-set. The rationale behind this idea is that by imputing several values, it is possible to capture the uncertainty added by the missing data and, at the same time, it allows to use the complete data set for the analysis. To do the multiple imputations, it is necessary to generate random draws with the distribution of the missing values. In the case the missing data mechanism is MAR, we can estimate these distributions with the observed data.
If there are m imputed values or imputations for each missing value, there will be m complete data sets. The analyses are done in each of the complete data sets and then combined to have a unique result. This permits to avoid the problems of loss of power and bias from complete-case analysis. Also, it avoids the bias produced by single-estimation methods.

After multiple imputing the data, the statistic of interest θ and its standard error are estimated in each complete data set. We denote this estimators $\hat{\theta}_i$ and $\hat{\sigma}_i$ respectively with $i = 1 \ldots m$. Then, these m estimated values are combined according to the Rubin rules to get a unique estimator for the estimate and its variance, which permits to do inference. The Rubin rules for the estimation of θ and σ are given, respectively, in (3.20) and (3.21) (Rubin, 1987, p. 76).

$$\hat{\theta}_{MI} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i \tag{3.20}$$

$$\hat{\sigma}_{MI} = \frac{1}{m}\sum_{i=1}^{m}\hat{\sigma}_i + \left(1+\frac{1}{m}\right)\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta}_{MI})^2 \tag{3.21}$$

In the last formula, the first term is the average variance and the second term term is the variance between imputations. With the last term, the method includes the uncertainty produced by imputing missing values. Rubin rules also include formulas to calculate t-statistics for doing inference regarding the parameters. They are generic and they are not associated to a specific imputation model. Nevertheless, in order to do inference, they should be applied to estimators which are normally or asymptotically normally distributed (Carpenter & Kenward, 2013, p. 41).

In order to create the imputations, it is necessary to estimate the distribution of the missing values and take m random samples of this distribution. For accomplishing this, it is necessary to define an imputation model for the variables with missing data.
If the missing data mechanism is MAR, the imputation model is a model for the conditional density $f(Y_{mis}|Y_{obs})$. The imputation model is defined by the researcher and it should: respect the nature of each variable, consider all relevant predictors and capture the essential characteristics of the data. If the aim of the research is the application of a substantive model for interest, this model with the imputation model should be congenial, that means that the imputation model should include the

relations from the substantive model (Carpenter & Kenward, 2013, p. 41), (Enders, Mistler, & Keller, 2016).

Chapter 4

Simultaneously modeling selection bias and missing data in the context of multilevel data

In this chapter, we present a discussion about solving selection bias problems and missing data problems in the context of multilevel modeling. The methods for selection bias are linear adjustment for confounding variables and propensity score matching. Propensity score matching will be used as a pre-treatment of the data in order to apply a multilevel model as the final analysis. The method for missing data is multiple imputation.
Fundamental aspects to consider are how the hypotheses and the implementations of the methods can be extended considering the multilevel structure of the data.

4.1 Selection bias and multilevel data

In this section, we discuss how to estimate causal effects when there are selection bias problems in the context of multilevel modeling. Multilevel models can be used for the propensity score model and to model the outcome. Arpino (2010, 84-85,112-120) elaborates the relevance of considering the multilevel dimension in the estimation of causal effects in three points: cluster-heterogeneity of the treatment effect, the multilevel nature of the selection process and potential violations of the SUTVA assumption. Cluster-heterogeneity refers to treatments that can have different effects within each cluster. This can can be modeled using random slopes for the treatment variable. The multilevel nature of the selection process is important, because it can depend on cluster characteristics and interactions between cluster and individual variables.
Other advantages to estimate causal effects with multilevel models are that it is possible to account for the data characteristics and adjusting for unmeasured covariates (Feller & Gelman, 2015). Finally, when they are used for modeling the outcome,

they allow to have treatment and outcomes varying at different levels. For example, in the study of part III of this thesis, the outcomes are the intercepts and slopes for each student trajectory.

Nevertheless, using multilevel models raises new challenges because it is necessary to take more modeling decisions and check hypotheses at more levels. For example, it is important to choose between fixed or random effects; the role of aggregate variables is complex and hypotheses regarding the random effects distribution and exogeneity should be met at all levels.

To organize the discussion, we classify causal effect estimation in multilevel settings in two cases: *intra-cluster* and *inter-cluster* treatment assignment. Equations (4.1) and (4.2) illustrate each case. In both equations, z denotes the treatment, y denotes the outcome and u_j are random effects or fixed effects. We used as examples two-level multilevel models, but the ideas can be applied to models with more levels. The *intra-cluster* case is when inside the cluster there are control and treated units, which is illustrated in equation (4.1). Thoemmes and West (2011) relate this design to multisite randomized studies, where there are treatment and control units in different sites, like hospitals or schools.
The *inter-cluster* case is when in each cluster there are only treated units or only control units, which is illustrated in equation (4.2). In this case, we are concerned with cluster effects. Zubizarreta and Keele (2016) relate this design to clustered-randomized studies.

1. Intra-cluster design
$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 z_{ij} + u_j + \varepsilon_{ij} \\ u_j &= \alpha_0 + \mu_j \end{aligned} \tag{4.1}$$
2. Inter-cluster design
$$\begin{aligned} y_{ij} &= \beta_0 + u_j + \varepsilon_{ij} \\ u_j &= \alpha_0 + \alpha_1 z_j + \mu_j \end{aligned} \tag{4.2}$$

We will use this classification to describe how the hypotheses from linear regression that are relevant for unbiased estimation of treatment effects are extended for the multilevel case. Finally, we will discuss how linear adjusting for confounding variables and propensity score analysis can be implemented in these two cases.

4.1.1 Hypotheses from the selection bias methods in a multilevel context

In chapter 3, we saw that a fundamental hypothesis for having unbiased treatment effect estimation is to have exogenous explanatory variables (hypothesis H3, equation (3.3)). From the causal effect framework, the relevant hypothesis was the ignorability assumption (equation (3.11)). In this section, we will focus more on how the exogeneity hypothesis is extended and achieved in multilevel models, because this analysis is our ultimate goal.

In order to have non biased estimates, the hypothesis H3 has to be extended to the case of multilevel models. In this case, it applies to the residual term and to all the random effects (Cheslock & Rios-Aguilar, 2011). If there is correlation between an explanatory variable with any random effect it will produce biased estimates. To exemplify, we define the model from equation (4.3). This is a two level model with a residual term ε_{ij}, random intercepts μ_j and explanatory variables $X = (x^1, x^2)$ varying at different levels.

$$y_{ij} = \beta_0 + \beta_1 x^1_{ij} + \beta_2 x^2_j + \mu_j + \varepsilon_{ij} \tag{4.3}$$

The hypothesis H3 extended to this model as follows:

- H3: The expectancy of the error term and all the random effects given X is 0 for all i.

$$\mathbb{E}(\varepsilon_{ij}|X_{ij}) = 0 \ \ \forall i \tag{4.4}$$

$$\mathbb{E}(\mu_j|X_{ij}) = 0 \ \ \forall i \tag{4.5}$$

In this context, a very important cause of endogeneity is omitted variable bias (Kim & Swoboda, 2010). Regarding how to address endogeneity problems, Arpino (2010, 97-102) considers two types of endogeneity and describes how they can be handled. If the endogeneity came from the correlation of a covariate with the residual term, it is called *first level endogeneity.* This occurs when equation (4.4) is not true. If the endogeneity came from correlation between covariates with the random effect, it is called *second level endogeneity.* This occurs when equation (4.5) is not true. In the following list we describe some possible solutions:

1. First level endogeneity with first level covariate: This problem is analogous the endogeneity problem in linear regression. It can be addressed with the same methods, for example instrumental variables or simultaneous equations. Econometrics books describe this problem in detail (e.g. Wooldridge (2010)).
2. First level endogeneity with second level covariate: This problem can be solved defining fixed effects at the second level instead of random effects. This will prevent us from including second level covariates. Thus, it produces a loss of information. However, the estimators at the first level would be correct. Kim and Frees (2006) and Kim and Swoboda (2010) analyze this problem for cases with several levels and propose solutions.
3. Second level endogeneity with first or second level covariate: When the random effect is correlated with a variable that varies at the first level, Arpino (2010, 97-102) shows that including in the model the cluster average of the covariate eliminates the correlation with the random effect. This is especially important when the endogeneous variable is the treatment assignment variable.
   Castellano, Rabe-Hesketh, and Skrondal (2014) discuss the case when also the

cluster average of the covariate is endogenous. An example is when the effect of good practices in schools is correlated with the mean of the students' socioeconomic status (SES). In this case, the average SES is endogenous, and its coefficient is not the causal peer effect. Castellano et al. (2014) offers techniques to give bounds to the estimates.
Also, there is the option of modeling fixed effects. In this case, there are no random effects and there are not endogeneity problems. But again, this option has as disadvantage that we cannot add second level variables (Hanchane & Mostafa, 2012). If the treatment variable is at the cluster level, it is not possible to estimate the treatment effect. There are also other options as instrumental variables, for example the work of Manzi, San Martín, and Van Bellegem (2014).

In these analyses we did not consider random slopes, but Kim and Frees (2006) and Kim and Swoboda (2010) consider a general case with random slopes and any number of levels.

In educational sciences, specially in the research on school effects, having stratified systems produces endogeneity problems. Hanchane and Mostafa (2012) argue that stratified systems produce correlation between student level variables and omitted school level characteristics. They compared educational systems with different levels of stratification and found evidence for second level endogeneity in the more segregated systems. They showed the relevance of including the school mean of the endogenous student level variables or model school fixed effects.

These issues are complex to connect because they are developed in two different approaches in multilevel models. On the one hand, the classical approach in educational sciences considers that random effects are very important because they allow us to model heterogeneity between groups and consider the groups as a sample. The problem is that the exogeneity hypothesis should be true for the random effects. On the other hand, the econometric approach highlights much more the accomplishment of the hypotheses to have non biased estimators. It recommends more the use of fixed effects (Cheslock & Rios-Aguilar, 2011). Nevertheless, Hill (2013) discuss that the modeling of fixed effects is not a panacea, and they should not be used to avoid properly adjusting for relevant confounding variables.
Castellano et al. (2014), Hanchane and Mostafa (2012), Cheslock and Rios-Aguilar (2011) and Hill (2013) connect both approaches and allow us to understand the factors to consider what to decide.

Another line to address the problem of selection bias is to evaluate the ignorability assumption. That assumption is not posed in terms of the residual term of the random effects at different levels, but in terms of the potential outcomes. The ignorability assumption can be assumed to be true at the individual or cluster level, thus defining the type of analysis that has to be done. The first possibility is to try to emulate a classical randomized experiment. In this case, the ignorability assumption should be true at the individual level. The second option is to try to emulate a clustered randomized study and match clusters, as in Hansen, Rosenbaum, and Small (2014). This should be the case if the treatment assignment depends only on cluster level variables.

Regarding the SUTVA assumption, the plausibility of this assumption depends on the level of the treatment assignment. For example, Stuart (2007) argues that in studies where the unit is the class and students in the treated classes can interact with students in control classes there can be spillover effects, and SUTVA may be violated. Nevertheless, in the case of interventions at the school level it is more plausible. We will assume that the SUTVA hypothesis is plausible. For modeling possibilities in intra-cluster treatment assignment when SUTVA is not appropriate see (Arpino, 2010, 115-120). When the assignment is inter-cluster and the results are interpreted at the group level, there can be interference between units inside each cluster but not interference between clusters (Hill, 2013, pp. 203) .

4.1.2 Modeling multilevel data and selection bias

In this section we discuss the estimation of causal effects controlling for selection bias trough linear adjustment for confounding variables, implementing propensity score matching and using both techniques simultaneously.

Linear controlling

Addressing selection bias in the case of linearly adjusting for confounding variables is straightforward trough multilevel models because it is easy to add the treatment variable and the control variables at the corresponding level. Equations (4.6) and (4.7) represent examples of linear controlling for the intra-cluster and inter-cluster case. X is the matrix with the control variables at the individual level and X' is the matrix with the control variables at the cluster level. The treatment assignment variable is z and y denotes the outcome.

1. Intra-cluster design

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_{1j} z_{ij} + X_{ij}\vec{\beta_2} + u_j + \varepsilon_{ij} \\ u_j &= \alpha_0 + X'_j \vec{\alpha_1} + \mu_j \end{aligned} \tag{4.6}$$

2. Inter-cluster design

$$\begin{aligned} y_{ij} &= \beta_0 + X_{ij}\vec{\beta_1} + u_j + \varepsilon_{ij} \\ u_j &= \alpha_0 + \alpha_1 z_j + X'_j \vec{\alpha_2} + \mu_j \end{aligned} \tag{4.7}$$

In order to estimate causal effects of the treatment, the major challenge is to determine if the exogeneity hypothesis is plausible (equations (4.4) and (4.4)). Of course, this depends on each specific research problem and related variables, but it is important to consider all the possible sources of endogeneity: within each level and between levels. These analyses can determine if it necessary to integrate the cluster mean of first level variables and the definition of fixed or random effects.

In the case of intra-cluster treatment assignment, it is possible to define μ_j as random effects and control by cluster-level variables or as fixed effects. In equation (4.6), we defined random effects and we controlled by cluster level variables .
In the case of inter-cluster treatment assignment, it is not possible to define fixed effects at the cluster level because it would be impossible to identify the effect of the treatment variable. In this case, cluster level covariates X' should eliminate all the source of endogeneity.
There can be complex designs where the strategies can be combined, for example in studies with students nested in schools which are nested in municipalities. If the treatment is at the school level, the study corresponds to an inter-cluster design regarding the school clustering, but to a intra-cluster design regarding the municipality. To use all the advantages of multilevel models we could define random effects at the school level and fixed effects at the municipality level.

Propensity score matching

In the case of propensity score matching, it is necessary to analyze intra-cluster and inter-cluster treatment assignments separately. For doing propensity score matching, the first step is to estimate the propensity score; the second step is to balance the samples to achieve the same propensity score distribution; and then to compare the outcomes in the balanced samples. In each of these steps it is necessary to determine how to take into account the multilevel structure.
Matching techniques are not designed to model the causal effect and it is necessary to define a model for the outcome (Stuart, 2010) . We consider that always a multilevel model is the final analysis for the outcome. In addition, we restrict to the case where one treated unit is matched to one control unit (1:1 matching). In this case, it is straightforward to use a multilevel model for the outcome model in the matched data. Other matching possibilities may imply the use of weights and they are more complex to combine with multilevel models.

With respect to intra-cluster treatment assignment, first it is necessary to define the propensity score model. In this case, the probability to being assigned to the active treatment can be modeled in several ways. A first decision is using only individual level covariates or include also cluster level covariates. A second decision is choosing between ignoring the clustering, random effects or fixed effects.
Arpino and Mealli (2011) considered the situation when there is an unobserved cluster-level variable that influences the treatment assignment. They found that the fixed effect models performed very well, better than the random effect model that did not include the unobserved variable. Nevertheless, they did not use a multilevel model for the outcome. Thoemmes and West (2011) and Leite et al. (2015) found that including the multilevel dimension in the propensity score was important, especially when the intra-class correlation is large. Nevertheless, if all the variables

relevant to treatment assignment were considered, they did not found important differences in bias reductions according to the specification of the propensity score. These can be explained because they also used a multilevel model for the outcome. This protected them from errors in the multilevel specification of the propensity score (Leite et al., 2015). Another aspect to consider is if the selection process is different in each cluster. This can be included in the propensity score model using random slopes (Leite et al., 2015).

Regarding the matching process, there are several issues to consider. Thoemmes and West (2011) defined two cases that we depict in Figures 4.1a and 4.1b and describe in the following paragraphs.

Case 1 is where the cluster level is a fundamental aspect of the design. There can be variations on the treatment implementation and interference between units in the same cluster. In this case, the idea is to emulate a multisite randomized trial. The selection model can vary in each cluster.

Case 2 is when the clustering is a characteristic of the data, but is not the main aspect of the study. The treatment is implemented without variation.

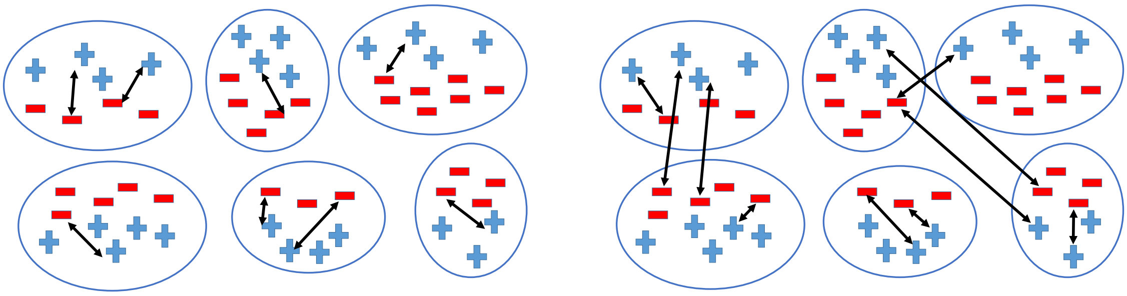
The distinction between the two cases is relevant for choosing a matching strategy. In Case 1, it is necessary that balance is achieved in each cluster, and it implies matching units inside each cluster. Therefore, this strategy provides an exact matching for the observed and unobserved cluster variables. In addition, it is necessary to have enough sample sizes of treated and control units and common support in each cluster.

In Case 2, the clustering is a feature of the study, but it approximates a randomized experiment when the assignment is regardless cluster-membership. Here the matching is done between treated and control individuals that can belong to different clusters, nevertheless level 1 and level 2 variables have to be considered.

The final step is to estimate the treatment effect comparing the outcome in the matched samples trough a multilevel model, for example as in equation (4.1). This is a good option because it protects from non include the multilevel aspect in the propensity score model. Also, this is a more robust strategy because selection bias can be addressed through matching and linear adjustment.

In the case that the treatment assignment is inter-cluster, first it is necessary to decide how the treatment assignment should be modeled. We depict the possible strategies in Figure 4.2. A first option is to do the matching at the individual level including in the propensity score model only individual level covariates, which correspond to Figure 4.2a. A second option is to match individuals including in the propensity score model individual and cluster level covariates as in Figure 4.2b. A third option is to do the propensity score model and the matching at the cluster level as in Hansen et al. (2014). Here, a parallelism is done with a clustered-randomized study. In this case only cluster level variables can be considered. This strategy is depicted in Figure 4.2c. The fourth and final option is to match clusters and individuals, which allows to measure the treatment effect in comparable individual units matched in comparable clusters. Examples of this case are the study from Zubizarreta and Keele (2016)[1] and Wang (2015). This strategy is represented in

[1] They did not use propensity score matching, but cardinality matching which maximizes the



(a) Case 1 intra-cluster (Thoemmes & West, 2011)

(b) Case 2 intra-cluster (Thoemmes & West, 2011)

Figure 4.1: Strategies for matching clustered data in the case of intra-cluster assignment. Figure (a) is when the clustering is extremely relevant and the selection process can differ across clusters. Figure (b) is when the the clustering is a feature of the design, but is not necessary to match within clusters.

Figure 4.2d.

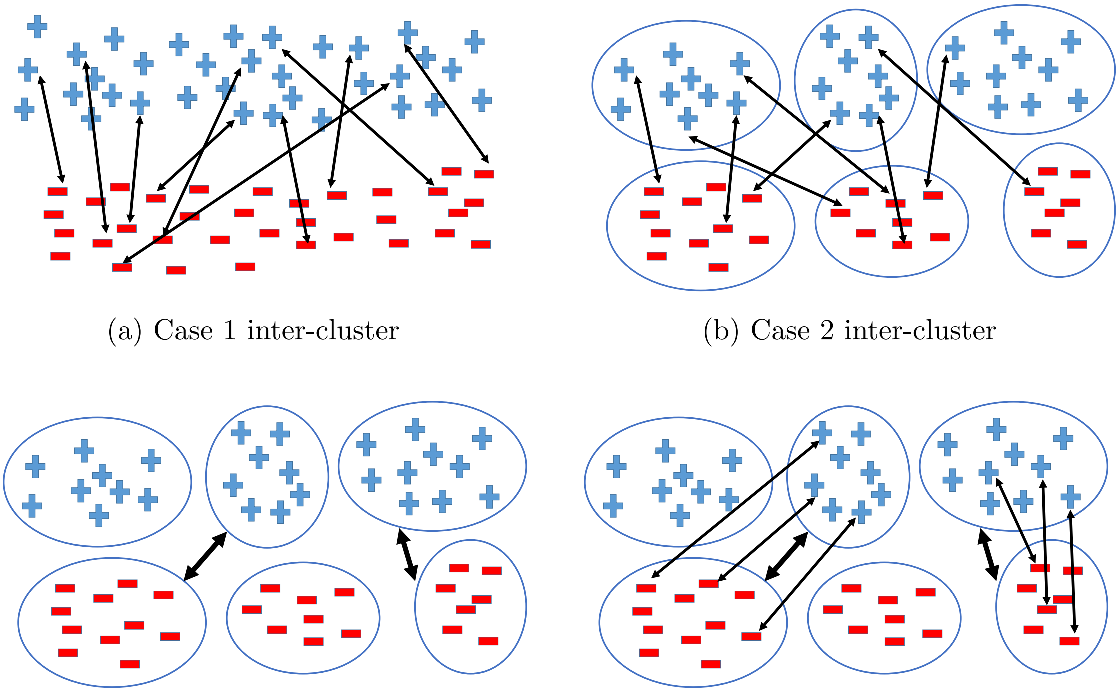
In a randomized study, it is clear if the design matched clusters or individuals. Nevertheless, in an observational study it is not obvious. Not always the treatment assignment can be related to a clustered randomized study. Figure 4.2a correspond to cases where the treatment assignment is clearly individual and the treatment implementation is clustered. For matching the individuals, we only need a regular propensity score model with individual level variables. For modeling the outcome, a multilevel model is necessary. The study from Bellio and Gori (2003) is an example of this strategy.

Figure 4.2c correspond to cases when it is clear that the treatment assignment is between clusters. For example, Hansen et al. (2014) studied the effect of massive floods in children health. They matched flooded and non flooded villages because being in a flooded village was not an individual decision and the treatment assignment was at the cluster level.

Finally, there are cases were it is less clear and treatment assignment can depend on individual and cluster level variables. They are depicted in Figures 4.2b and 4.2d. Studies comparing public and voucher schools in Chile are a good illustration of the problem. There is family selection for the school and school selection of the students, which implies that the treatment assignment is at the individual level and can depend on individual and cluster level variables. This could lead to match students, schools or both. For example, Zubizarreta and Keele (2016) matched individuals and schools. Their strategy allows to study comparable students that attended to comparable schools, but produced a very large sample reduction at the student and school level.

Another example is the study in Part III of this thesis. In this study, propensity score matching was done according to strategy from Figure 4.2a. However, we included cluster level variables in the multilevel model. We choose this strategy because the clustering was cross-classified and too complex to match schools. We decided to model individual variables in the propensity score, in order to compare students

number of individuals and clusters matched.



(a) Case 1 inter-cluster

(b) Case 2 inter-cluster

(c) Case 3 inter-cluster

(d) Case 4 inter-cluster

Figure 4.2: Strategies for matching clustered data in the case of inter-cluster assignment. Figure (a) is when the treatment assignment depends on individual variables. Figure (b) is when the treatment assignment depends on individual and cluster variables that are integrated in the propensity score model, but only individuals are matched. Figure (c) is when the treatment assignment depends only on cluster variables and there is the need to match clusters. Finally, Figure (d) is when the treatment assignment depends on individual and cluster variables and individuals and clusters are matched.

with similar background characteristics, but control for the school level covariates in the multilevel outcome model. This strategy allowed us to compare models with and without school level variables. The school level variables were socioeconomic school composition and selection practices. The interest of this comparison is estimating the effect of the school context and practice (type A effects) and the effect of only the school practice (type B) of public and voucher schools (Raudenbush & Willms, 1995).

For any of the strategies depicted in Figure 4.2, the use of a multilevel model for the outcomes is fundamental and can be an important help to overcome specification errors in the propensity score model.

4.1.3 Discussion regarding the modeling of multilevel data and selection bias

Linear adjustment and propensity score matching focus the researcher on different aspects of the data. On the one hand, linear adjustment in multilevel models offers

advantages because there are several strategies to avoid endogeneity problems and we can think how the variables can interact at different levels. On the other hand, propensity score matching prompt us to focus on the selection process. For example, in intra-cluster designs considering if the selection process is different inside each cluster and in inter-cluster designs considering if it is at the individual level, cluster level or both. Also, a main interest of propensity score matching is the evaluation of common support that increase internal validity of the conclusions.
Nevertheless, propensity score matching is not a method for modeling the outcome (Stuart, 2010). The combination of propensity score matching with multilevel models is a better strategy for properly taking into account the multilevel structure and for adjusting for selection bias with matching and linear adjustment.

Treatment effect estimation with multilevel data can imply a wide variety of research designs, type of treatment assignments, number of levels analyzed, etc. We organized the literature and it recommendations using the intra-cluster and inter-cluster treatment assignment classification.

4.2 Multiple imputation and multilevel data

The imputation techniques should preserve the characteristics of the data that are relevant for the analysis model. This means that the multiple imputation model and the analysis model should be congenial (Carpenter & Kenward, 2013). Also, if units within clusters are more similar between them, this should be captured by the imputation model (Carpenter & Kenward, 2013, pp. 205). If there are random slopes, contextual effects, several levels and complex nesting structures as cross-classified or multiple-membership models, ideally the imputation model should include them. Also, it should respect the nature of the variables and consider their role in the main model. In addition, the missing data mechanism can depend on variables at different levels (Enders et al., 2016). Simulation studies show that ignoring the multilevel structure can lead to biased estimates, especially in random effects parameters (Enders et al., 2016; Lüdtke, Robitzsch, & Grund, 2017; Van Buuren, 2011). Another aspect is that missing data at the cluster level can produce loss of all the units in a cluster if it is not imputed, which may imply a major loss of information (Van Buuren, 2011)

The use of multiple imputation with multilevel data has received several recent developments. In this section, we discuss implementation issues, which are very relevant for the applied researcher. The main hypothesis of multiple imputation is the MAR assumption. It is stated in terms of the distribution f of the missing data. The main issue is how to allow the models to consider the multilevel nature of the data in the modeling of this distribution. This brings us to the implementation of the multiple imputation model.

There are three main techniques to do imputation of multilevel variables: joint modeling, chained equations and fixed effects (Enders et al., 2016). In joint modeling the joint distribution of the variables is modeled (Goldstein, 2011, pp. 304-313). In chained equations the distribution of each variable conditional on the others is modeled. Finally, in fixed effects the idea is to use a single level model with dummy variables for the clusters.

Enders et al. (2016) compared single level imputation, joint modeling for multilevel imputation, chained equations for multilevel imputation and fixed effects. They found that joint modeling and chained equations work well for two-level random intercept models, but none of the methods can account for all the possible complexities. Using theory and simulation studies, they showed that joint modeling performed better for contextual effects and not well for random slopes. Also that chained equations performed well for random slopes but not for contextual effects. Also that joint-modeling allows to model categorical multilevel variables and chained equations do not.

Regarding fixed effects, adding the multilevel structure in the imputation models through fixed effects is better than ignore it or use complete case analysis (Enders et al., 2016; Van Buuren, 2011). However, it can produce biased estimates, specially for the random effects parameters (Drechsler, 2015). In addition, fixed effects in the imputation model can produce bias in the fixed part if there is a high amount of missing data and low intra-class correlations (Drechsler, 2015). In conclusion, if there is interest in estimating random effect parameters, using fixed effects in the imputation model can be detrimental.

A specific aspect is the role of identification variables and it has not received much attention in the literature. In multilevel models, it is necessary to identify for each unit the cluster membership with an identification variable. There are no strategies to impute this value when is missing (Van Buuren, 2011). It is very complex to determine the distribution of an identification variable and become with a reasonable imputed value. This is a very important problem that can arise in longitudinal studies for example. In the case of the study of the part III of this thesis, 53% of the sample was lost because of this issue.

To the best of our knowledge, there is no statistical software that can handle all possible aspects of multilevel data as several levels, cross-classified and multiple-membership structures, categorical variables, actualization of aggregated variables with lower level variables. Most of the available software packages to estimate multilevel imputation models allows maximum two level models. In the case of having three levels because the data is longitudinal, organizing the data in wide-format can avoid a level. Nevertheless, there are several cases where it is necessary to model three levels or more. For example, in a study with repeated measures of students which are nested in classes nested in schools.

Some technical literature gives advice regarding how to implement multilevel imputation for cross-classified structures. And also states that extend their proposed model for data with several levels is easy (for example Carpenter and Kenward (2013, pp. 225-226)). Nevertheless, for the applied researcher it can be cumbersome to code his own methods. There is the need of automated software to deal with this.

In summary, the imputation model should integrate the multilevel structure of the data with all its complexity. Research in educational sciences is shifting to studying more longitudinal data, there is access to bigger and more complex databases. We hope that soon all the possibilities in the software will be developed in order to preserve the exact nature of the data. In the meantime, we need studies to evaluate the loss of not properly model some aspects of the data.

4.3 Selection bias, multiple imputation and multilevel data

In the previous section we described strategies to model multilevel data and selection bias produced for non-random treatment assignment with linear adjustment and propensity score analysis. We described some relevant aspects to evaluate the endogeneity hypothesis and the treatment assignment considering the clustered nature of the data. In addition, we described implementation issues of how to use multiple imputation with multilevel data.

For selection bias models, the exogeneity and the ignorability assumptions are done over complete data. If there is missing data, it is important to understand if these assumptions are still valid and how they can be modeled. The condition for doing valid treatment effect estimations over imputed data is stated in the following equation, named latent ignorability (Frangakis & Rubin, 1999; Hill, 2004; Mattei, 2009)

$$\mathbb{P}(Z = 1|X, M, y^1, y^0) = \mathbb{P}(Z = 1|X) \tag{4.8}$$

Equation (4.8) implies that the treatment assignment depends on X which has observed and not observed values. The important thing is that having this hypothesis and a correct multiple imputation model permits to estimate the treatment effect on the imputed data. Thus far, we need to know the models for the assignment mechanism and the missing data mechanism.

Equation (4.8) is general enough to work in the case of multilevel data, but special attention should be given to the multilevel multiple imputation model. It should preserve the relations between the covariates and the treatment variables. If the missing data mechanism interacts differently with the treatments, the imputation model should include interactions with the treatment variable. For example, if in the treated groups non response is related to a covariate and in the control group it does not, an interaction between the covariate and the treatment variable should be included. The same occurs with the clustering. If the treatment assignment is different for each cluster, the multilevel imputation models should preserve the structure of random intercepts and slopes. Is important to note that the imputation model should include the outcome variable, the treatment variable and the covariates relevant for the MAR assumption. In a study regarding how to combine propensity score weighting and multiple imputation Leyrat et al. (2017) found more bias if in the imputation model the outcome is not included.

From a practical point of view, there are several possibilities to mix the analyses to adjust for selection bias, missing data and the clustered structure of the data. Here, we propose starting with multiple imputation, because that allows to do the rest of the analyses using complete data-analysis. We consider that in all the strategies a multilevel model for the outcome is the final objective. The strategies are depicted in Figure 4.3.

Strategy 1 is simply doing multiple imputation and then estimate a multilevel model for the outcome, where we linearly control by confounding variables. This strategy does not include propensity score analysis. It is important because, in some cases, propensity score matching can be unfeasible or too complex. For example, when the sample sizes are very unequal and there are not several matches. Also, if there are several treatments, it is very simple to add a treatment variable that has more than two values (Hill, 2013).

Strategies 2 and 3 consider how to combine multiple imputation and propensity score analysis. We defined them considering the work of Hill (2004) and Mitra and Reiter (2012). These strategies replicate the within and across approach defined by Mitra and Reiter (2012).
Strategy 2 correspond to the within approach. It implies to do m multiple imputations. Then, in each completed data set, estimate the propensity score and match the treatment and the control group. This will lead to m matched data sets, which can have different sample sizes and units. Finally, estimate the treatment effect in each matched set and pool the m estimated treatment effects.
Strategy 3 correspond to the across approach. It implies to do m multiple imputations. Then, in each completed data set estimate the propensity score and pool the propensity scores through the m imputations. Then, match the pooled propensity scores. This will lead to one set of matched units. Selecting the matched units in each imputed data set produce m matched data sets, which have the same sample size and the same units. Finally, estimate the treatment effect in each matched set and pool the m estimated treatment effects.

The advantage of Strategy 2 is that it allows to use observed covariates with missing data. For example, if the matching is done with the propensity score but exact matching is need in a covariate. In general, any strategy that implies work with observed covariates can be implemented. Having this possibility is appealing because using covariates can approximate a blocked randomized study. When we only match propensity scores, we approximate a completely randomized study that is less efficient than a blocked randomized study (King & Nielsen, 2016).
One disadvantage is that it can be a burden to the researcher to do the matching m times, specially if balance has to be evaluated in each imputed data. The researcher should choose between use the same model for each imputed data set or different models. The other disadvantage is that the m matched data sets can be of different sizes and include different units. If the treatment effect is going to be estimated with

a mean difference, there should not be further problems. But, if we used multilevel models, there are parameters that depend a lot on the sample size, for example the deviance. These parameters are not comparable between the different imputed data sets.

The advantages of the third strategy are several. First, there is only one propensity score by unit and the matching is done once. Also the balance can be evaluated in pooled estimates of standardized mean differences. Pooled estimates of mean differences are less dependent on individual imputations and should be more reliable and produce better modeling decisions. Finally, for larger samples it is more computationally efficient because the matching is done only once (Leyrat et al., 2017). A disadvantage is that if we want to do exact matching in a covariate that has missing data it is not possible because it changes in each imputed data set. A possibility could be using single imputation in that covariate (Stuart, 2010). This would allow to do one matching using the pooled propensity score and the covariate.

Regarding the efficiency in decreasing the bias, in the study from Mitra and Reiter (2012) the within approach produced less bias reductions than the across approach. The difference in bias reduction was more important when the treatment assignment depended on variables with missing data. In Mitra and Reiter (2012), they did not include the outcome in the multiple imputation model. This was done because sample balancing should be done without using the outcome when using propensity score analysis. Nevertheless, in multiple imputation is recommended that all the variables that are related to the missing mechanism should be included. Recently Penning de Vries and Groenwold (2016) replicated the results of Mitra and Reiter (2012) including the outcome in the imputation model and found advantages for the bias reduction in the within approach. Also, advantages for the within approach were founded by Leyrat et al. (2017) in a study that combined propensity score weighting and multiple imputation.

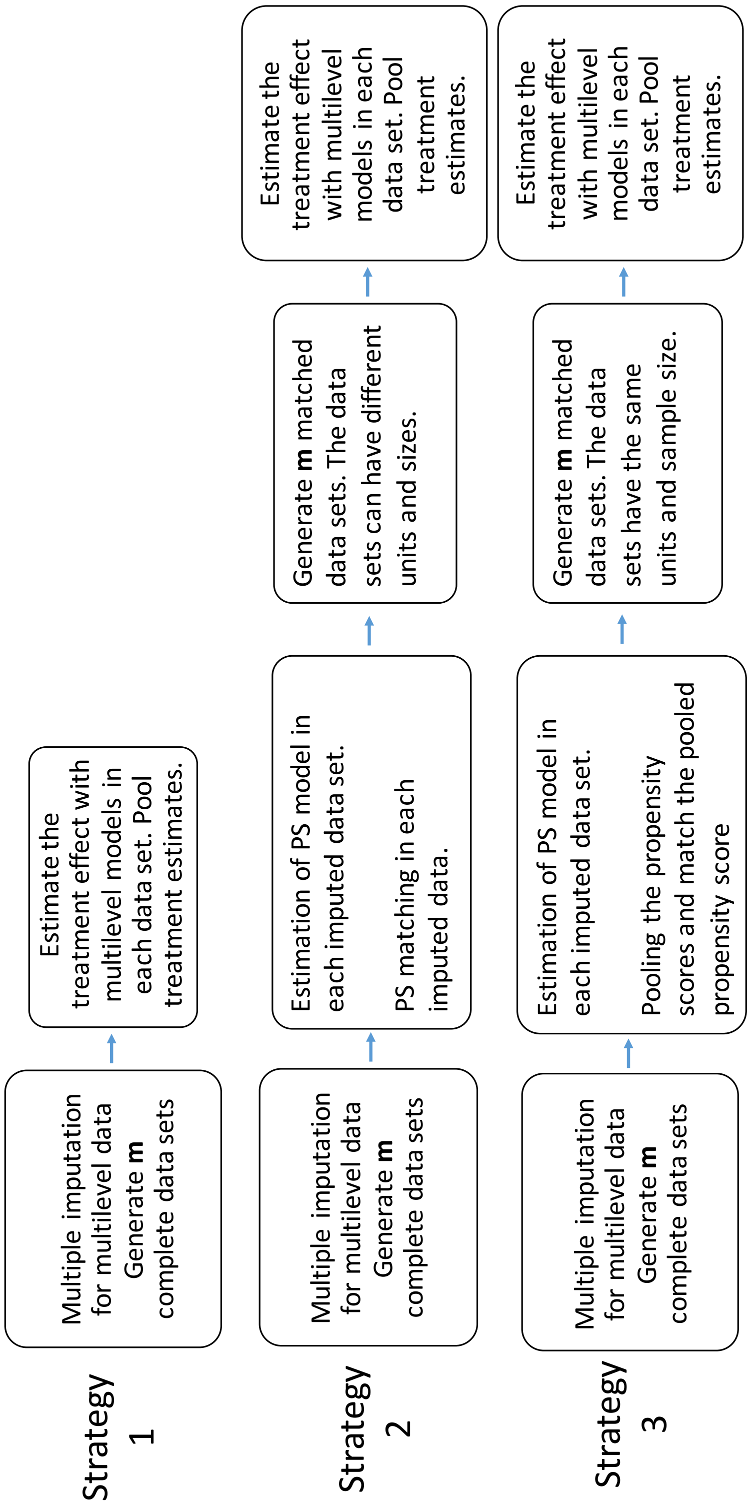

Strategy 1
Multiple imputation for multilevel data Generate m complete data sets
Estimate the treatment effect with multilevel models in each data set. Pool treatment estimates.
Strategy 2
Multiple imputation for multilevel data Generate m complete data sets
Estimation of PS model in each imputed data set.
PS matching in each imputed data.
Generate m matched data sets. The data sets can have different units and sizes.
Estimate the treatment effect with multilevel models in each data set. Pool treatment estimates.
Strategy 3
Multiple imputation for multilevel data Generate m complete data sets
Estimation of PS model in each imputed data set.
Pooling the propensity scores and match the pooled propensity score
Generate m matched data sets. The data sets have the same units and sample size.
Estimate the treatment effect with multilevel models in each data set. Pool treatment estimates.

Figure 4.3: Statistical strategies to address selection bias and missing data using multilevel models

Chapter 5

Summary and future research

In educational research there are very complex examples of clustered data, selection bias and missing data. Multilevel models became more complex and it appears the need of cross-classified and multiple membership structures. In this part of the thesis we described how to use multilevel models having problems of selection bias and missing data. For selection bias we considered linear adjustment at several levels and propensity score matching, were the final outcome model is multilevel.

Regarding modeling multilevel data and selection bias, first we precise the distinction between intra-cluster treatment assignment and inter-cluster treatment assignment. This distinction is fundamental to start a strategy to control for selection bias. In our analysis, we considered that always the model for the outcome is a multilevel model. Therefore, the major aspect that we described was how the hypothesis of exogeneity is extended. We also describe how to deal with endogeneity problems using cluster means of individual level variables or fixed effects. Furthermore, we also discuss how to analyze the treatment assignment and the relevance of the clustering to make a modeling strategy for the propensity score.

In a randomized experiment it is clear whether the treatment assignment was done at the individual level or at the cluster level. In an observational study it is not obvious and the researcher has to understand at which level should do the matching. We described (Thoemmes & West, 2011) distinction for intra-cluster treatment assignment and extended it for the inter-cluster treatment assignment. This can be useful for applied researchers.

On the one hand, modeling the outcome with multilevel models and using linear adjusting for selection bias implies that several hypotheses should be met. The most relevant for the estimation of the treatment effect is to check the exogeneity hypothesis at different levels and think in the interactions between levels. On the other hand, propensity score matching make us reflecting about the assignment mechanism and can also include the multilevel structure of the data. Use them combined is a stronger strategy.

Also, multilevel models allow us to control by unobserved cluster characteristics trough the definition of fixed effects in the propensity score and the outcome model.

There are studies that can have an intra-cluster and inter-cluster treatment assignment. In this case these ideas can be combined. For example, an intervention at the classroom level. Considering the school clustering, the intervention is intra-cluster. In this case we should think if the treatment assignment is the same at each school and we can model school fixed effects. From the classroom level, it is inter-cluster and it is necessary to think if students or classes should be matched. To develop a full picture, additional studies are needed that illuminate which are the best decisions regarding the better matching strategy (as depicted in Figures 4.1 and 4.2) when we are also using linear adjusting for individual and cluster-level variables in the multilevel model for the outcome. When modeling individual outcomes only, combine matching and linear adjusting is a better strategy (Stuart, 2010). From this point of view, in the inter-cluster case using matching and linear adjusting for the cluster level is a better strategy also. Nevertheless, there appear issues of sample size, because usually at the cluster level the samples are significantly smaller than at the individual level. Also matching at the cluster level can produce a loss of information regarding the heterogeneity of the treatment implementation. Finally, the decision can depend on the type of effect that we want to measure, for example in school effects we may need to distinguish between type A and type B school effects (Raudenbush & Willms, 1995).
In summary, other issues should be studied regarding how to define the best strategy, for example sample sizes (total sample size, number of cluster, cluster sizes) and the type of treatment effect that we want to estimate.

Regarding multiple imputation and multilevel data, we described latest results and developments. Several simulation studies show that it is important to include the multilevel aspect in the imputation models. Software packages have not developed all the possibilities for the modeling. However, we think that soon this will be done in order to have congeniality between the outcome models and the imputation model.

Regarding combining multiple imputation, propensity score matching and multilevel models, we described three possible strategies.
Strategy 1 only considers multilevel models and multiple imputation. It is useful when propensity score matching is not feasible because of a small control group or several treatments. When matching is not recommended, propensity score weighting can be a good option, nevertheless we did not analyze this in the thesis. Regarding dealing with propensity score weighting and multilevel models we refer to Li, Zaslavsky, and Landrum (2013) and Leite et al. (2015). Regarding mixing propensity score weighting and multiple imputation a reference is Leyrat et al. (2017).
Strategies 2 and 3 include the combination of multiple imputation with propensity score matching. Strategy 2 has as advantages that, according to simulation results, it produces better bias reduction (Penning de Vries & Groenwold, 2016). Also it is possible to use the covariates to improve the matching. Nevertheless, it is a burden for the researcher.
Strategy 3 has several advantages. One advantage is that it has a simpler implementation than Strategy 2. Also, Strategy 3 produces imputed data sets with the same units and sample sizes. Finally, balance can be evaluated in pooled standardized

differences, which should be less sensitive to individual imputations. The problem is that, according to simulation results, this strategy produce less bias reduction than the within approach and we cannot integrate the value of a covariate for the matching process. However, there are possible solutions to these issues. A possibility to achieve the same bias reduction is using simultaneously propensity score and linear adjustment. In addition, we can use single imputation to include in the matching process observed covariates with missing data.

We analyzed the possibility of including covariates on the matching process considering recent critics to propensity score matching. These critics propose that propensity score matching could increase imbalance, inefficiency, model dependence, and bias. The reason is that propensity score matching tries to reproduce a completely randomized experiment which is less efficient than a fully blocked randomized experiment (King & Nielsen, 2016).
We think that using single imputation in the matching procedure could help to use Strategy 3 and try to reproduce a fully blocked randomized experiment. This should not harm because one of the mayor problems of single imputations is bias in variances, which in general are not used for the matching. Then, we would select the matched units in each multiple imputed data set and estimate the multilevel models. Nevertheless, this issue needs further study.

Regarding future research, it is not clear how to decide whether matching individuals, clusters or both in observational studies. In particular, this should be explored in two cases: when only matching is used and when matching is combined with a multilevel model for the outcome. It is important to determine which factors are more relevant and we need more methodological advice to understand this.
For example, to compare public and voucher schools in Chile, there is freedom for the families to choose schools and freedom for vouchers schools to choose students. The assignment mechanism is very complex, and it make sense to match students and also schools or at least control by school level variables. In this way, propensity score matching and linear adjustment provide a protection to possible bad defined models.
Also the strategies delineated in Figure 4.3 could use another type of propensity score adjustment such as weighting or stratification. Again, there are practical questions that appear regarding the combination of propensity score and multiple imputation. If propensity score weighting is used, Do the weights should be pooled? If we use stratification, The stratification should be done in the propensity score for each imputed data set or over the pooled propensity scores? Regarding propensity score weighting Leyrat et al. (2017) showed that Strategy 2 leaded to the best estimates in terms of bias reduction. But we think that more research is needed regarding this issue.

Complexities of mixing models came from the different origins and theoretical frameworks that make difficult to understand the different languages. For example Cheslock and Rios-Aguilar (2011) illustrate differences from multilevel models from the usual framework in educational sciences and the econometric theory. Also, propensity score analysis tries to emulate randomized experiments with observational data, and it use the literature of randomized experiment and design of experiments. There are examples that connect the different approaches (e.g. Cheslock and Rios-Aguilar (2011) and Castellano et al. (2014)). However, we need more studies that connect the literature of multilevel models, which are used extensively in education, with the literature on econometrics and treatment effect estimation.

Part II

The influence of Life and Non-Life sciences courses on secondary students' epistemic and self-efficacy beliefs related to sciences

Chapter 6

Introduction and Literature Review

6.1 Introduction

Studies with secondary students have shown that more sophisticated science epistemic beliefs are related with better achievements, better motivational beliefs and better learning strategies (Deng et al., 2011; Mason et al., 2013; Tsai et al., 2011). In addition, having proper scientific epistemic beliefs is a science education goal in itself (Lederman, 2007, pp. 832). The picture is similar with science self-efficacy, with studies showing that these beliefs are one of the most important predictors of science achievement and the choice of scientific careers (Britner & Pajares, 2006; Larose, Ratelle, Guay, Senécal, & Harvey, 2006). Thus, it is important to know which kind of academic climate can produce more availing scientific epistemic and self-efficacy beliefs.
In this study, three characteristics of the academic climate will be studied: the type of scientific discipline; scientific instruction with and without laboratory work and the compositional effects of beliefs and achievement.

Biglan (1973) found three dimensions that distinguish disciplines. The first dimension distinguish disciplines that have a single paradigm, which were labeled hard, or disciplines that include several paradigms, which were labeled soft. The second dimension distinguish the applied and pure disciplines. The third dimension distinguish between disciplines that study living objects or life disciplines and disciplines that study non living objects or non-life disciplines. We think that this later distinction is relevant for science education. For example, in the study of Leslie, Cimpian, Meyer, and Freeland (2015), they found a relation between the percentage of women and the perception of that innate talent is fundamental for succeed in the discipline. They did not distinguish between life and non-life disciplines. However, in their results regarding scientific disciplines, it can be seen that life disciplines tend to have lower perceptions of the need of innate talent and a larger percentage of women than the non-life disciplines. We think that the distinction between life and non-life scientific disciplines can enlighten some processes and we measure its effect on science epistemic and self-efficacy beliefs.

Regarding the effect of the type of discipline, studies about general epistemic beliefs have established that students from soft domains are more sophisticated than students from hard domains (Muis, Bendixen, & Haerle, 2006). This has been explained with the nature of hard disciplines that can be characterized as having a single paradigm, having greater consensus regarding the questions, the content and the methods (Biglan, 1973).
The same argument can be used in scientific epistemic beliefs, and would predict that students from scientific disciplines, which are classified as hard disciplines, should be more naïve than the ones from non-science disciplines in their beliefs about sciences. But, an incongruity arises because we also expect that the exposition to scientific disciplines should change beliefs about the nature of science to more sophisticated views, because having more educational experiences produces more sophisticated epistemic beliefs (Hofer & Pintrich, 1997) and should give a better understanding about the nature of science. Conflicting empirical results illustrate this incongruity. For instance, it has been found that in some scientific epistemic beliefs dimensions, students from science majors were more naïve than students from non-science majors and vice versa (Chai, Deng, Wong, & Qian, 2010; Liu & Tsai, 2008). To enlighten this discordance, it is necessary to understand the characteristics of the scientific disciplines that are being delivered to the students. However, studies analyzing the effect of different scientific disciplines in science epistemic beliefs are scarce and they also show some conflicting results. Also, it is not clear if the differences between disciplines are an effect of the exposition to the disciplines or an effect of self-selection, for example if more naïve students choose scientific domains (Trautwein & Lüdtke, 2008).

Therefore, in this study, the effects of science summer courses about scientific disciplines with different characteristics (life objects and non-life objects) were tested, a control group was defined and control variables were measured in order to adjust for self-selection.
Another significant aspect of science instruction and scientific knowledge development is laboratory work, where we tested if the effect of the scientific disciplines can be explained by the use of laboratory work. This is also a contribution to the discussion regarding if the use of laboratory work can change scientific epistemic beliefs (Deng et al., 2011) and motivational variables (Itzek-Greulich et al., 2017).
The last goal of this study is analyze if there are peer effects. Theoretical models propose that the beliefs of the students' peers influence epistemic beliefs (Feucht, 2010). Nevertheless, this hypothesis has not been tested with quantitative studies measuring compositional effects. In the case of science self-efficacy beliefs, we will test if the splashdown effect differ by scientific discipline. This concept, defined by Stake and Mares (2005), states that after assisting to a challenging science course with high ability peers, the self-confidence tends to decrease or do not change at the end of the courses. But when the same student comes back to his/her school,

he or she reevaluates his or her capacities in comparison with regular peers and its confidence increases.

In summary, the research questions are:

1. What is the impact of courses of different science domains (Life - Non-life) on high school students' science epistemic and science self-efficacy beliefs at the end of the courses and some months later controlling for self-selection bias?
2. How this impact is explained by using instruction with laboratory work?
3. Are there compositional effects of the science courses on high school student's scientific epistemic and self efficacy beliefs?

6.2 Literature Review

6.2.1 Science Epistemic and Self-Efficacy Beliefs

Beliefs about science or scientific epistemic beliefs have been studied in two main lines of research. The first line of research came from the literature of personal epistemology and follows the work started by Perry Jr (1999) in the seventies about the development of beliefs about knowledge and knowing. The work of Perry opened a line of research on epistemic beliefs that has been evolving, starting with very large qualitative studies looking for developmental trends to quantitative studies that considered different dimensions of epistemic beliefs (Hofer & Pintrich, 1997). The quantitative and multidimensional research in epistemic beliefs started with the seminal work done by Schommer (1990). Hofer and Pintrich (1997) postulated that the main areas of epistemic beliefs are beliefs about the nature of knowledge and beliefs about the nature of knowing, excluding beliefs about learning or ability. In its theoretical framework, an example of a proposed dimension is certainty of knowledge, were a naïve point of view is the belief that knowledge is static and does not change and a sophisticated point of view is that knowledge evolves. Another example is source of knowledge, a dimension about the role of the authority and the self as owners of the truth. In this dimension, the naïve extreme would be believe that knowledge came always from authorities and outside the self and a more sophisticated extreme is that knowledge can origin from the self and through reason (Hofer & Pintrich, 1997). These dimensions have been evolved with empirical results (e.g. Hofer (2000)) and the development of new questionnaires. Until the year 2000 most of the research in epistemic beliefs focused primarily in general epistemic beliefs, during those years started the interest in domain-specific epistemic beliefs, which is a natural continuation because disciplines have different epistemic assumptions and characteristics (Muis et al., 2006). In this context, the research about science epistemic beliefs usually implies an adaptation of the dimensions of general epistemic belief in relation to sciences. For example, Hofer (2000) developed an epistemic beliefs questionnaire adaptable to different domains. Other example is the work of

Conley, Pintrich, Vekiri, and Harrison (2004) who developed questionnaire specific for science but based in the theoretical framework of Hofer and Pintrich (1997). The second line of research regarding beliefs about science is framed in the Visions of the Nature of Science (VNOS). This line of research is specific to science education. Lederman (2007), in his review about the topic, proposed several characteristics of the nature of science matching the usual dimensions proposed in the personal epistemology literature (for example, scientific knowledge is tentative and subject to change). But VNOS considers other characteristics as, for example, conceiving science as socially and culturally embedded. In this study we will consider studies from both lines of research, but with emphasis on the personal epistemology line regarding science epistemic beliefs.

This study also focuses in the effect of science courses in science self-efficacy beliefs. These beliefs are about the perceived capacity to succeed and learn in sciences (Britner & Pajares, 2006) and they are key variables to understand the success in learning and school achievement (Schunk & Pajares, 2009). In particular Science self-efficacy beliefs are very important because they are related with science achievement and the choice of a scientific career (Britner & Pajares, 2006). Usually, science self-efficacy is one of the most important predictors of achievement.
The relation between science self-efficacy and epistemic beliefs has primarily been tested with cross-sectional studies, for example (Chen & Pajares, 2010; Kizilgunes, Tekkaya, & Sungur, 2009; Mason et al., 2013; Tsai et al., 2011). All the proposed models said that epistemic beliefs influence science self-efficacy beliefs (and not the reverse relation). The mechanism to explain these relations are different, for example Buehl (2003) propose that epistemic beliefs influences the perceptions that people have about a task. Consequently, influence the perceptions that people has about their ability to accomplish the task, their motivation and the reasons to do it. In particular for self-efficacy beliefs, in order to evaluate our capacity regarding a subject domain, the visions that we have about a domain should influence this evaluation (Buehl, 2003). Nevertheless, it is not evident if that imply that we should expect a positive relation (were more sophisticated epistemic beliefs are related with more science self-efficacy) and which dimensions of epistemic beliefs should be related to self-efficacy beliefs. The empirical results are in general supporting that more sophisticated epistemic beliefs are related with higher self-efficacy beliefs.

6.2.2 Science epistemic beliefs, self-efficacy beliefs and change: Influence of academic climates

In this subsection we describe the effect of different academic climates on student scientific epistemic and self-efficacy beliefs.

Studies about the contextual effect of different disciplines

Few studies have investigated the effect of different scientific disciplines in science epistemic beliefs. We describe all the articles about the subject that we found.

Chai et al. (2010) and Liu and Tsai (2008) found that students from science majors were more naïve in some of VNOS dimensions than students from non-science majors. The interpretation is that science major students are more exposed to epistemic climates where knowledge is objective and universal. These studies have compared scientific and non-scientific domains. Regarding the effect of different scientific disciplines in science epistemic beliefs, we only found three research articles. In the first one, Miller, Montplaisir, Offerdahl, Cheng, and Ketterling (2010) compared VNOS dimensions before and after two different science courses, however students from one course were non-science majors, and from the other were science majors. So, it is not clear if differences are the effect of course subjects or the major. In the second article, Thoermer and Sodian (2002) compared VNOS dimensions between undergraduate and graduate students in Physics, Biology and Chemistry. They found an effect of the discipline, were the most sophisticated students came from physics. They provide as a possible interpretation that in physics the change of theories is more manifest and the relation between experiments and theory is more evident. The last study we found is by García and Mateos (2013), who studied a sample of university professors from Biology, Chemistry, Mathematics and Physics. They found that professors had advanced beliefs about the nature of science, but Physics professors tended to have more naïve beliefs, in opposition to Thoermer and Sodian (2002) results described above. It should be noted however that the instruments they use are very different.

Regarding science self-efficacy beliefs, the discussion has been more focused on the specificity of self-efficacy for different disciplines than on the effects of different disciplines on self-efficacy (e.g. Bong (2001); Schunk and Pajares (2009)). However, some studies have found differences in self-efficacy between disciplines. For example, Larose et al. (2006) classified a sample of students from three college programs (general science, technological related to Physics, technological related to Biology) according to their science self-efficacy trajectories (increasing, decreasing and stable). In the increasing group, there were proportionally more students from the Biology technological program than in the other groups (Larose et al., 2006).

Looking at the research that has been done, it is not clear how the exposure to different scientific domains change scientific epistemic and self-efficacy beliefs and it is not clear if the effect came only from self-selection or from the influence of disciplines. Another aspect that has not been researched is if the change last or not.

Studies about the contextual effect of laboratory work and peer effects

Regarding the effect of laboratory work on scientific epistemic beliefs, there are many studies from the VNOS framework. Some of these studies compare the effect of explicit teaching of VNOS, the effect of experimental work, inquiry based learning and the effect of teaching history of science (Lederman, 2007). As a summary from the reviews by Lederman (2007) and Deng et al. (2011), explicit VNOS instruction with laboratory work is more efficacious, but implicit VNOS instruction through practical laboratory work did also produce changes. Research from personal epistemology has been more focused on comparing traditional and constructivist types of instruction. Muis and Duffy (2013) showed that more constructivist instruction produces more sophisticated beliefs. They interpreted this result as a process of enculturation that is facilitated by teacher and student interactions. In fact, theoretically, we expect that learners' epistemic beliefs influences between them (Feucht, 2010; Muis & Duffy, 2013). Regarding the effect of peers on self-efficacy beliefs, there are studies describing the "big-pound small-fish effect", that is how the peers' capacities can influence the self-efficacy beliefs (Marsh, 1987). There is evidence of this effect on science self-efficacy in the work by Stake and Mares (2005).

6.2.3 Science university courses for high school students

In this study, we consider courses belonging to the Summer School of the University of Chile. This institution offers summer courses to secondary and primary students, lasting between one and three weeks. In Chile, the school year runs from March to December and summer vacations are during January and February. The summer school courses take place during January in Santiago.

The courses belong to the following knowledge domains: Physics, Mathematics, Engineering Sciences, Biology, Chemistry, Biomedical Sciences, Social Sciences, Humanities, Visual Arts and Artistic Expression. To enter the program, students' applications are selected by school grades. The courses are paid but they are not expensive and there are scholarships available. In this study, only science related courses were considered, totaling 49. These courses can be classified regarding whether their objects of study are alive (N=29) or non-alive (N=20). For simplicity of language, we will name them Life courses and Non-life courses respectively. In the Non-life courses the subjects are Mathematics (N=10); theoretical and experimental Physics (N=7); Engineering (applied Computer Science, Electronics or Physic, N=7); Chemistry (N=2); Astronomy (N=2) and Geology (N=1). In the Life courses the subjects are Human Health (from Microbiology to Specific Organ Systems and their Diseases, N=12); Biochemistry and Biotechnology (N=4); Biology (N=2) and the study of Ecosystems or Animals (N=2).

The courses are done with a university approach in a university environment, where teachers have academic experience and they are experts in their areas. There is a focus on "learning by doing". For example, all science courses have practical activities together with theoretical classes. In the Life courses, practical activities usually mean laboratory work, while in the Non-life objects this depends on the subject. For example, in Mathematics the practical activity is problem solving sessions, in Chemistry courses it is laboratory work and in Engineering courses it is building artifacts (for example robots).

These courses are different from the usual science courses in Chilean schools, which are very traditional, with little student interaction and little scientific inquiry (Cofré et al., 2010). Some subjects are completely different from high school subjects, as Geology for example, or have a very different approach than the school one, as Mathematics. Regarding the practical laboratory work, the approach is more professional and specialized than in the school. For example, the health-related courses run laboratory work at the Faculty of Medicine Laboratories. The courses are challenging and they differ from real university courses in the length and the consequences that grades have for students. Finally, the student's peers in Summer School are different from peers in school because the students in the Summer School choose to study during the summer, they are motivated and diverse.

The classification of course disciplines in Life and Non-life was done considering the work of (Biglan, 1973). There the author found three empirical dimensions to classify disciplines: Hard - Soft, Pure - Applied and Life - Non life. The classification hard-soft has been used in studies of personal epistemology (e.g. (Muis et al., 2006)), but for this study it was not pertinent because all science courses are classified as hard. The classification pure-applied was suitable, but in many courses this classification was very ambiguous. Finally, the third dimension Life - Non-life was pertinent because it was feasible to classify the courses and it allows to do an interesting distinction of disciplines. Also using this well defined classification of science disciplines would allow the present study to be replicable.

Chapter 7

Methodology

7.1 Sample

The study sample is composed of N=994 students that applied for a science course in the Summer School. However, not all students answered the questionnaire in the pre, post and follow-measure. The sample sizes for each measurement time are detailed in Table 7.1. The applicants could take two paths: enroll to a science course and be part of the treatment group (N=782) or do not enroll and be part of the control group (N=212). This control group is composed by two kinds of students: selected students that finally did not take the courses and not selected students. Being not selected depends on student grades and on available places in courses. Thus, there are not selected students with similar grades in very demanded courses. The students in the sample are mostly secondary students. The percentages of students in 8th, 9th, 10th, 11th and 12th grade are 6.4%, 16.2%, 29.9%, 37.4% and 10.1%, respectively. Regarding gender, 58.0% of them were girls. These students can be considered as motivated since all of them are willing to spend part of their summer vacations taking a course in the university. They school grades are above the national grade point average.

Table 7.1: Sample sizes for each measurement time.

	Group			
Measurement time	Life	Non Life	Control	Total
Pre	216	297	182	695
Post	322	385	66	773
Follow-up	116	157	41	314

7.2 Variables: Measures of Sciences Epistemic Beliefs and Self-efficacy beliefs

The sciences epistemic beliefs questionnaire used in this study was developed by Conley et al. (2004). It has 26 items measuring the following dimensions: Development, Justification, Certainty and Source. Development has 6 items that measure beliefs about the developing nature of science, a sample item is "Ideas in sciences sometimes change". We notice that in some studies this dimension is also called certainty (e.g. Hofer (2000)). Justification has 9 items and measure beliefs about how individuals justify knowledge in science, specifically regarding the use of experiments, a sample item is "A good way to know if something is true is to do an experiment". Certainty has 6 items and measure the belief that in science there exists a unique answer, a sample item is "All questions in science have one right answer". Finally, Source has 5 items and measures the beliefs that truth in science come from external authorities, a sample item is "Only scientists know for sure what is true in science".
This questionnaire was validated by Conley et al. (2004) and by Tsai et al. (2011), in samples of primary and secondary students respectively, showing good psychometric properties. The original English questionnaire was translated to Spanish and presented to colleagues for meaning appropriateness of each item. The science self-efficacy beliefs questionnaire has 6 items adapted from the mathematics self-efficacy scale used by Tuohilampi et al. (2015), a sample item is "I am sure that I can learn sciences". These items were randomly ordered in the questionnaire. The items of both questionnaires have a Likert scale from 1 (Strongly disagree) to 5 (Strongly agree). The items in English and in Spanish and the final questionnaire can be found in appendix A.1.

7.3 Procedure

Record variables and the pre-beliefs questionnaire were measured in October 2014, during the student application period. After the courses finished, the beliefs questionnaire was applied two times: a post-measure just after the courses ended in January 2015 and a follow-up measure applied between April and May 2015. The application protocol for the beliefs questionnaire is detailed in Table 7.2.
We asked for consent to use the questionnaire data and records from the application form to students older than 18 and to the parents of students younger than 18.

Table 7.2: Application calendar and procedure for the treatment group and the control group.

Group	Pre	Post	Follow-up
Experimental	Voluntary online questionnaire at the end of the application form	Paper and pencil questionnaire applied just at the end of the courses	Voluntary online questionnaire sent to students' e-mails
Control	Voluntary online questionnaire at the end of the application form	Voluntary online questionnaire sent to students' e-mails	Voluntary online questionnaire sent to students' e-mails
Measurement time	October 2014	January 2015	April - May 2015

7.4 Data Analysis

In order to address the multilevel data structure and to adjust for selection bias we used multilevel linear regression models and linear regression models. Outcome, control and treatment variables are described in Table 7.4. Using list-wise deletion implied a too large loss of information (Table 7.3) because there was an important percentage of missing data in the outcome variables (see Table 7.4). Therefore, we chose to do multiple imputation which is a recommended method for data missing at random (Peugh & Enders, 2004). Multiple imputation of the database with all the sample implied too much added noise (Table 7.3), so we decided to use the approach proposed by (Von Hippel, 2007) of multiple imputation of the data base with all the population and then deletion of observations where the outcome is missing. This decision was a good compromise to use all the available information and avoid an excess of noise (Table 7.3).

Table 7.3: Sample sizes and drawbacks for the missing data strategies.

	List-wise deletion	Multiple Imputation	Multiple Imputation and then deletion
Post	N=168	N=994	N=773
Follow-up	N=168	N=994	N=314
Strategy drawbacks	Loss of power for loosing observations. Biased estimates because data is not missing completely at random (Peugh & Enders, 2004).	Loss of power for added noise in imputed outcome variables. (in post there are 221 individuals with imputed outcome variables, and in Follow 680).	

The analyses were performed with software R (R Core Team, 2016). Reliability estimation and confirmatory factor analysis were done with the packages Psych (Revelle, 2016) and lavaan (Rosseel, 2012), respectively. Multiple imputation was done with package Mice (Buuren & Groothuis-Oudshoorn, 2011) and (Zhao & Schafer, 2016) and multilevel modeling was done with package Nlme (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2017).

Table 7.4: Description of variables and percentages of response.

Variable	Role	% Missing	Description
grades	control	0	GPA for 2013 and first semester of 2014.
level	control	0	School level (between 8th and 12th).
gender	control	0	1 if student is a boy.
Mother education	control	24	Mother education in 5 levels: incomplete secondary education (12-), complete secondary education (12), incomplete tertiary education (12+), complete tertiary education (12++), complete tertiary education with graduate education (12+++).
Self-efficacy pre	control	25	Self-efficacy in science 2-3 months before the courses.
Development pre	control	26	Development 2-3 months before the courses.
Justification pre	control	26	Justification 2-3 months before the courses.
Certainty and Source pre	control	26	Certainty and Source 2-3 months before the courses.
Self-efficacy post	outcome	20	Self-efficacy in science at the end of the courses.
Development post	outcome	20	Development at the end of the courses.
Justification post	outcome	21	Justification at the end of the courses.
Certainty and Source post	outcome	21	Certainty and Source at the end of the courses.
Self-efficacy follow	outcome	67	Self-efficacy in science 3-4 months after the courses.
Development follow	outcome	67	Development 3-4 months after the courses.
Justification follow	outcome	67	Justification 3-4 months after the courses.
Certainty and Source follow	numeric	67	Certainty and Source 3-4 months after the courses.
Discipline	treatment	0	Course discipline in 3 levels: Control group, Life course and Non-life course
Lab	treatment	0	1 if the student course had Laboratory work. For students in the control group the value is 0.
ABC	treatment	0	Average belief climate, where a higher value in this variable means that the average of the students in the class have higher school grades, science self-efficacy, development and justification beliefs in the pre-questionnaire. For students in the control group the value is 0.

Chapter 8

Results

8.1 Psychometric analysis

8.1.1 Confirmatory Factor analysis and reliability estimation of student beliefs

In order to validate the questionnaire, in each measurement time we estimated a confirmatory factor analysis for categorical ordered variables to verify the measurement structure. We looked for a model that fits for the three times measures. First we tried the theoretical measurement model with five correlated factors (science Self-efficacy, Certainty, Source, Development and Justification). This model was not satisfactory because the Certainty and Source dimensions had a very high correlation in the pre, post and follow-up measure (0.97, 0.93 and 0.95 respectively). Therefore, we fit a model with 4 factors, where Certainty and Source items were loaded in the same factor (Certainty and source). This model fits acceptably in the three measurement times (see Table 8.1) and it was established as the measurement model. The Cronbach's alphas for the dimensions defined by the measurement model range between 0.73-0.89 for the pre measures, between 0.70-0.86 for the post measures, and between 0.87-0.96 for the follow-up measures.

Table 8.1: Robust fit indexes for the measurement model with four correlated factors (f1: Science Self-efficacy, f2: Certainty and source, f3: Development and f4: Justification).

Time	N	χ^2	d.f.	$\frac{\chi^2}{\text{d.}f.}$	CFI	RMSEA	SRMR
Pre	695	1431.182	458	3.12	0.94	0.055	0.073
Post	773	1728.755	458	3.77	0.92	0.060	0.077
Follow-up	314	1756.898	458	3.84	0.94	0.095	0.122

8.1.2 Compositional variables to describe class climate

To model compositional effects, for each course we estimated course-level aggregated variables as the mean of the class. In these variables, we assigned the value 0 to the control group. These variables were highly correlated and produced collinearity problems, so we summarized them with a principal component analysis. The scree-plot and eigenvalues larger than one suggested extracting two components (the loading matrix is presented in Table 8.2). Component 1 is called average beliefs climate (ABC) and it is calculated primarily with the course aggregated variables: science self-efficacy, development and justification. Component 2 is primarily the aggregated variable Certainty and source. In the multilevel regression models we used as aggregated variables Component 1 and 2 instead of the aggregated original variables, eliminating collinearity problems. However, we did not include Component 2 in the final models because it was not significant in any of them.

Table 8.2: Principal component analysis for the aggregated beliefs variables. The analysis was done over N=49 courses.

Course-level variables	Component 1 (ABC)	Component 2
Aggregated self-efficacy pre	0.84	-0.04
Aggregated justification pre	0.92	0.12
Aggregated development pre	0.77	-0.24
Aggregated certainty and source pre	0.11	0.98
% of Variance	53.61	25.90

8.2 Descriptive analysis

In Table 8.3, means and standard deviations of beliefs variables for each group (Life, Non-Life and Control) show that in the pre-beliefs questionnaire, students from Life courses have higher self-efficacy and more sophisticated beliefs than the Non-Life and the control group students. This is evidence of a self-selection effect. In the post and follow measures there is the same pattern.

Table 8.3: Means and standard deviations of students beliefs in Life courses, Non-life courses and in the control group for each measurement time.

	Pre			Post			Follow-up		
Variable	Life	Non Life	Control	Life	Non Life	Control	Life	Non Life	Control
Science Selfefficacy	4,58 (0,47)	4,44 (0,59)	4,55 (0,52)	4,51 (0,45)	4,4 (0,51)	4,41 (0,76)	4,55 (0,62)	4,28 (0,99)	4,14 (1,24)
Development	4,33 (0,45)	4,28 (0,52)	4,28 (0,49)	4,41 (0,4)	4,35 (0,46)	4,34 (0,65)	4,36 (0,61)	4,16 (0,95)	4,03 (1,17)
Justification	4,56 (0,4)	4,5 (0,5)	4,56 (0,45)	4,59 (0,32)	4,5 (0,39)	4,53 (0,64)	4,52 (0,56)	4,27 (0,95)	4,15 (1,21)
Certainty and Source	2,67 (0,63)	2,69 (0,65)	2,75 (0,63)	2,61 (0,6)	2,65 (0,65)	2,58 (0,64)	2,41 (0,61)	2,35 (0,63)	2,35 (0,82)
N	216	297	182	322	385	66	116	157	41

8.3 Modeling course effects

In order to assess if students belonging to a same course are more similar between them after the courses, we calculated the intra-class correlations for each belief in the pre, post and follow-up measure (Figure 8.1). The results show that for all the variables and time measures the intra-class correlations are very low (the biggest was 4.79%). However, the intra-class correlations in the post measure are systematically higher than in the pre-measure for all the variables, and decreased for the follow-up measure, except for Certainty and Source.

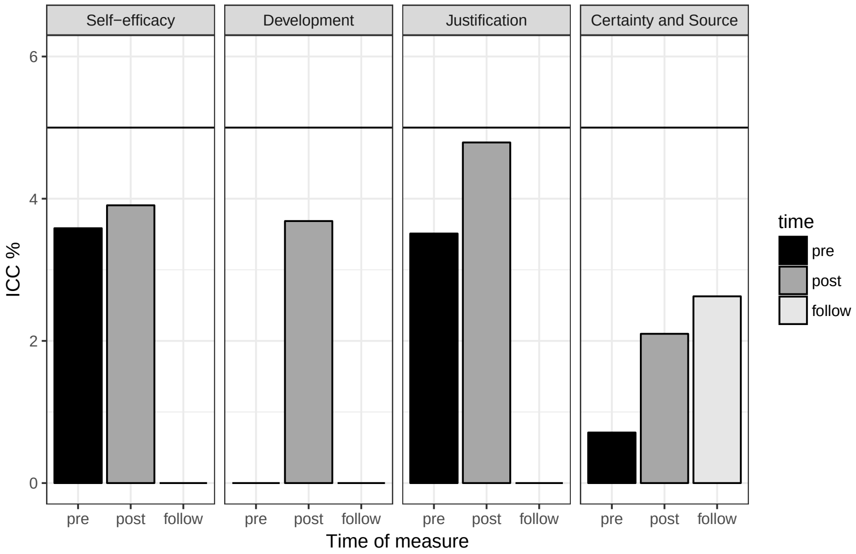


Figure 8.1: The intra-class correlations were estimated only with enrolled students. The sample sizes for the pre, post and follow-up measurement were N=513, N=707 and N=273 respectively.

8.4 Effect of academic climate at the end of the courses

In order to determine the effect of the different courses on science epistemic and self-efficacy beliefs at the end of the courses, we estimated the models from Table 8.4 for each post belief dimension. The models M2 and M3 allow us to estimate the effect of treatment variables on the outcomes variables controlling for selection bias through the inclusion of control variables. The variables representing the different treatments or climates are all at the course level. Results of the modeling of the post-beliefs as outcome variables are presented in Table 8.5. First, it can be seen from the comparison between model M1 with the models M2 and M3 that all the course level variance is explained (Table 8.5). Second, regarding the effect of the treatment variables on self-efficacy there is a significant and positive effect of assisting to a course with Laboratory Work. There are no significant effects of the discipline nor of the compositional variable ABC. In Development there are no significant effect of any course level variable. In Justification, with the model M3, we found significant and positive effects from being in a course with Laboratory Work. In Certainty and Source none of the treatment variables have a significant effect. As a summary, being in a Life or Non-life course did not produced significant effects, in any of the post beliefs variables. Laboratory work had significant effects in self-efficacy and justification post. There are no significant compositional effects from the aggregated variable ABC.

Finally, regarding the effect of the confounding variables, for each belief the most influential control variable is the same belief measured before the courses. There is an unexpected negative effect of previous Justification on post Self-efficacy. Being a boy has a significant positive effect on Self-efficacy post and in Certainty and Source post. Grades have negative significant effects on Certainty and Source post. Mother education is not significant for any variable.

Table 8.4: Estimated multilevel regression models where y_{ij} is the outcome variable for student i belonging to course j (the control group is modeled as a course); x_{ij}^1 to x_{ij}^k are k control variables; $Life_j$, $NonLife_j$,$ABAC_j$,Lab_j are course-level treatment variables and u_j and μ_{ij} are random effects.

Model	Equations	Description
M1	$y_{ij} = \beta + \mu_j + \varepsilon_{ij}$	Empty model
M2	$y_{ij} = \beta + \beta_1 x_{ij}^1 + \cdots + \beta_k x_{ij}^k + Life_j + NonLife_j + ABC_j + \mu_j + \varepsilon_{ij}$	Model with all control and treatment variables, except type of instruction (laboratory)
M3	$y_{ij} = \beta + \beta_1 x_{ij}^1 + \cdots + \beta_k x_{ij}^k + Life_j + NonLife_j + ABC_j + Lab_j + \mu_j + \varepsilon_{ij}$	Model with all the control and treatment variables

Table 8.5: Multilevel regression models to determine the effect of academic climates on science epistemic and self-efficacy beliefs at the post measure. The analyses were done over imputed data for models M2 and M3. N=773, number of groups is N=50 (49 courses plus the control group). * p-valor<0.1, ** p-valor<0.05.

	Science Self-efficacy post			Development post			Justification post			Certainty and Source post		
Parameter	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Intercept	0,01	-0,18	-0,18	0,01	-0,08	-0,08	0	-0,01	-0,02	0	-0,04	-0,04
Student-level variables												
Self-efficacy pre		0,44**	0,44**		-0,03	-0,03		0,02	0,01		-0,03	-0,03
Development pre		0,03	0,04		0,43**	0,43**		0,04	0,04		-0,08*	-0,08*
Justification pre		-0,18**	-0,18**		-0,08	-0,08		0,19**	0,19**		-0,01	-0,01
Certainty and Source pre		-0,01	-0,01		-0,05	-0,05		0,08**	0,08*		0,58**	0,58**
Gender (boy)		0,14*	0,15**		0,03	0,03		0,03	0,03		0,14**	0,14**
Grades		0,02	0,01		0,06*	0,06*		0,05	0,04		-0,1**	-0,1**
Level		0,02	0,02		0,04	0,04		0	0		-0,03	-0,03
Mother education (12)		0,11	0,11		0,04	0,04		-0,01	-0,01		-0,11	-0,11
Mother education (12+)		0,03	0,02		0,12	0,12		-0,06	-0,07		-0,3*	-0,3*
Mother education (12++)		0,13	0,12		0	0		-0,04	-0,04		-0,25*	-0,25*
Mother education (12+++)		0,05	0,05		-0,11	-0,11		-0,07	-0,07		-0,08	-0,08
Course-level variables												
Discipline (Life)		0,17	-0,06		0,19	0,15		0,18	0		0,1	0,2
Discipline (Non-life)		0	-0,09		0,08	0,07		-0,02	-0,09		0,1	0,14
ABC		0,03	0,04		0,03	0,03		0,06	0,06		0,06	0,06
Laboratory Work			0,25**			0,04			0,19*			-0,1
Random Effects												
Residual Variance mean	0,975	0,838	0,831	0,975	0,811	0,811	0,973	0,912	0,908	0,983	0,565	0,566
Residual Variance P5	0,975	0,811	0,805	0,975	0,786	0,786	0,973	0,896	0,893	0,983	0,539	0,54
Residual Variance P95	0,975	0,865	0,858	0,975	0,836	0,836	0,973	0,927	0,922	0,983	0,601	0,602
Intercept Variance mean	0,024	0,002	0	0,024	0,003	0,003	0,026	0,001	0	0,016	0,027	0,024
Intercept Variance P5	0,024	0	0	0,024	0	0	0,026	0	0	0,016	0,016	0,013
Intercept Variance P95	0,024	0,007	0,001	0,024	0,009	0,009	0,026	0,004	0	0,016	0,042	0,038
Fit Indexes												
Deviance mean	2190	2059	2051	2190	2034	2034	2190	2123	2119	2192	1778	1776
Deviance P5	2190	2033	2026	2190	2012	2011	2190	2110	2106	2192	1743	1742
Deviance P95	2190	2083	2075	2190	2058	2058	2190	2136	2132	2192	1830	1829
AIC mean	2196	2093	2087	2196	2068	2070	2196	2157	2155	2198	1812	1812
AIC P5	2196	2067	2062	2196	2046	2047	2196	2144	2142	2198	1777	1778
AIC P95	2196	2117	2111	2196	2092	2094	2196	2170	2168	2198	1864	1865
BIC mean	2210	2172	2171	2210	2147	2154	2210	2236	2239	2212	1891	1896
BIC P5	2210	2146	2145	2210	2125	2131	2210	2223	2226	2212	1856	1862
BIC P95	2210	2196	2195	2210	2171	2177	2210	2249	2251	2212	1943	1948

8.5 Effect of academic climates some months after the courses

To measure the effect of the treatment variables on the beliefs some months after the courses (follow-up) we used the same model specifications used in the post beliefs (see Table 8.4). However, we fitted linear regression models instead of multilevel models because the intra-class correlations for the follow-up measure were close to zero. The results are presented in Table 8.6. Regarding the effect of the treatment variables: on Self-efficacy there are significant effects of assisting to a Life course, that remains after include the effect of laboratory work; on Justification, we found significant and positive effects from being in a Life course, nevertheless when we add the variable Laboratory Work, the effect becomes non significant. In Certainty and Source, both disciplines have a significant effect on follow-up. In the case of Life sciences, it become non-significant after include the Laboratory Work variable.

As a summary, there were significant effects of assisting to a Life course on follow-up Self-efficacy, Justification and Certainty and Source. For Justification and Certainty and Source, the effects become non-significant when we add the Laboratory Work variable, but they remain of an important effect size. Being in a Non-Life course produces significant effects on Certainty and Source. Laboratory Work did not produce significant effects on any of the beliefs variables at this time measure.

There are not significant compositional effects from the average beliefs variable ABC. Regarding the effect of control variables, as before, for each belief the most influential control variable was the same belief measured before the courses. In comparison with the models done over post beliefs, the coefficients are smaller. In addition, for the academic and socio-demographic variables, we see that being a boy have a significant positive effect on follow-up Self-efficacy and Development. Grades have significant effects on Development. Mother education was not significant for any variable.

Table 8.6: Linear regression models to determine the effect of academic climates on science epistemic and self-efficacy beliefs at the follow-up measure. The analyses were done over imputed data for models M2 and M3. N=314, number of groups is N=50 (49 courses plus the control group). * p-valor<0.1, ** p-valor<0.05.

	Science Self-efficacy follow-up		Development follow-up		Justification follow-up		Certainty and Source follow-up	
Parameter	M2	M3	M2	M3	M2	M3	M2	M3
Intercept	-0,3	-0,3	-0,39	-0,39	-0,37	-0,38	-0,17	-0,17
Student-level variables								
Self-efficacy pre	0,16*	0,16*	-0,05	-0,05	-0,06	-0,06	-0,09	-0,09
Development pre	0	0	0,14*	0,14*	0	0	-0,08	-0,08
Self-efficacy pre	-0,02	-0,02	0,1	0,09	0,22**	0,22**	0,08	0,08
Certainty and Source pre	-0,08	-0,08	-0,03	-0,04	-0,05	-0,05	0,49**	0,49**
Gender (boy)	0,25**	0,25**	0,24**	0,24**	0,21*	0,21*	0,03	0,03
Grades	0,08	0,08	0,12**	0,12**	0,07	0,07	-0,08	-0,08
Level	-0,08	-0,08	-0,05	-0,05	-0,09	-0,09	-0,04	-0,04
Mother education (12)	0,16	0,15	0,25	0,25	0,23	0,23	0,02	0,02
Mother education (12+)	-0,07	-0,07	0,15	0,15	0,08	0,08	-0,28	-0,28
Mother education (12++)	0,01	0,02	0,24	0,24	0,11	0,11	-0,22	-0,22
Mother education (12+++)	-0,07	-0,07	0,08	0,08	0,07	0,07	-0,26	-0,26
Course-level variables								
Discipline (Life)	0,47**	0,55**	0,36*	0,31	0,46**	0,43*	0,35**	0,38*
Discipline (Non-life)	0,07	0,09	0,03	0,02	0,04	0,03	0,35**	0,36**
ABC	-0,09	-0,09	-0,09	-0,09	-0,09	-0,09	0,06	0,06
Laboratory Work		-0,08		0,04		0,02		-0,04
Description of R^2								
R^2	0,092	0,093	0,094	0,094	0,088	0,088	0,33	0,331
95% confidence interval (low)	0,037	0,038	0,038	0,038	0,033	0,033	0,233	0,233
95% confidence interval (high)	0,164	0,165	0,168	0,168	0,162	0,162	0,428	0,429

Chapter 9

Discussion and Conclusions

Regarding the impact of Life and Non-Life courses in science Self-efficacy beliefs, we see that there is no effect of the discipline in the post measure, but there is a positive effect of Life courses in the follow-up measure. This implies that in Life courses there was a splash down effect (Stake & Mares, 2005). However, in Non-life courses this effect is not present. Studies about the splashdown effect do not allow us to explain why Non-life courses did not produce a follow-up effect on Self-efficacy, but our results coincide with Larose et al. (2006) findings, where in the group of students with increasing science self-efficacy trajectories there were proportionally more students from the Biology technological program than in the other programs. One possible explanation is that attending a course about a Non-Life object does not increase science Self-efficacy but increases more specific self-efficacy measures, for example mathematics Self-efficacy or engineering Self-efficacy. Another possibility is that in mathematics, physics and similar exact sciences, to learn a content or skill it is very important to have mastered the previous knowledge. This may produce that recovering from a misconception or lack of understanding is harder, and for a student it is easier to get lost. This could explain why Non-Life courses did not change science Self-efficacy beliefs. A last hypothesis comes from the fact that in Non-Life courses there is a tendency to assign lower grades than in Life courses, specially in Mathematics courses (we analyzed the mean of the grades assigned in each course). This could produce negative mastery experiences, which are one of the most relevant sources of science Self-efficacy (Britner & Pajares, 2006).
The effect of laboratory experience on science Self-efficacy was significant only at the post measure. A possible interpretation of this is that experiments provide meaningful vicarious experiences and different types of assessment procedures where it is easier for a student to have a mastery experience. The effect was not present in follow-up Self-efficacy. This is in line with results from Itzek-Greulich et al. (2017) who found that Science Center Outreach Labs (SCOLs) had significant effects on self-competence in a post-test, but not in follow-up, 6 weeks after intervention.

Regarding the effect of Life and Non-life courses on epistemic beliefs, results show that the courses about Life objects have significant positive effects on Justification

at follow-up. The effects became non significant when we added Laboratory Work. Only for post Justification, Laboratory Work has a significant and positive effect.
In Justification, to see effects only from Life courses and from courses having Laboratory Work is reasonable, because in Life sciences the role of experiments to generate knowledge is more explicit than, for example, in mathematics or theoretical physics, where the argumentation is more important. In fact, in the majority of Life courses, there was Laboratory Work while in the Non-life courses, only about half of them had Laboratory Work. These courses were not designed to change epistemic beliefs. The fact that the simple exposition to Life sciences and laboratory experiences changed Justification is an important finding, more if we consider that in the case of Life courses this change remains several months after the intervention. These effects can be interpreted as socialization (Trautwein & Lüdtke, 2008) or enculturation effects (Muis & Duffy, 2013) because we controlled for the selection bias.
We expected effects on Development also, but we only found such effects in Life courses at the 0.1 alpha level. Maybe the courses were too short, or it is necessary to have explicit discussions about the changing nature of science in order to see a significant effect on this variable.
The dimension of Certainty and Source, with items stating that there is a right answer in science and that the scientific authorities are the source of the truth, increased in the follow-up measure for both disciplines. We think that this result can be explained because basic scientific courses in schools and universities do not question these beliefs explicitly, and probably only in advanced courses or post-graduate programs there is a possibility of reflecting about these beliefs.

In summary, through the testing of the effect of Life and Non-life courses and Laboratory work, this study sheds light on the question of whether science courses could produce more sophisticated science epistemic beliefs. We found that Non-life courses produce less sophisticated beliefs in Certainty and Source. However, the ones with laboratory activities produce more sophisticated beliefs on Justification. In Non-life subjects, an epistemic sophistication could be seen in other epistemic dimensions. For example, the social value of science could be more evident in engineering courses or the logical connection of arguments as a source of knowledge should be a relevant dimension in mathematics courses.
In addition, we looked for evidence of peer effects from two sources: the evolution of the intra-class correlation along the three time measures and the effect of the aggregated variables. The intra-class correlations showed that students were more similar in their beliefs at the end of the courses in all beliefs variables. Nevertheless, the intra-class correlations are null in the follow up measure. Regarding the effect of the aggregated variable ABC, for all the models it was non significant. Also the size of the effects of the aggregated variable were very small in comparison to the effect of the type of discipline and Laboratory Work.

This Summer School Program study, devised to see the effect of academic climates, is a great methodological opportunity because it provided the possibility of estimating intra-class correlation and compositional effects. The main limitation of this study is that, even though we control for several variables, there can be uncontrolled selection bias. Another limitation is that there is response bias and informed consent bias, specially in the data obtained trough questionnaires sent to the mail of the students. In addition, the questionnaire was not completely suitable for the Chilean population, because Certainty and Source items loaded in the same factor. This could have been an effect of the instrument adaptation or an effect of the culture or group of students (Buehl, 2008). Future work should describe the experience of belonging to a Life or Non-life course and of having Laboratory Work with qualitative techniques. Also this study could be replicated measuring epistemic dimensions more specific to Non-Life courses.

Part III

School effects in academic trajectories of Chilean students

Chapter 10

Introduction and Literature Review

10.1 Introduction

The research on school effects has received increased interest worldwide. This can be seen in the development of the field School Effectiveness, which focus is to identify the factors within schools and the educational system which can affect students academic and social development (Reynolds et al., 2014). The analysis of school effects can be useful to allocate resources and programs and to understand how the educational systems work. This is useful information to make decisions about the design of the systems.

In the Chilean context, there is a special interest regarding the effect of schools managed by private entities in comparison to schools managed by public entities. The Chilean educational system can be seen as an extreme case of the application of market oriented policies (Bellei, 2008; Valenzuela et al., 2014). It has a mixed structure, where a big portion of the schools is directed by private entities. There are three types of schools. The first type are public schools, which are managed and funded by the state. The second type are voucher schools which are the ones funded by the state but managed by privates like entrepreneurs and foundations. Finally, the third type are private schools which are managed and funded by privates (Bellei, 2008). Only 44.8% of the schools are public schools, 49.7% of schools are voucher schools, and 5% are private schools (Ministerio de Educación, 2014) . In comparison with participant countries in PISA 2009 this is a very high percentage of schools managed by private entities (OECD, 2012).
Also, there is a clear trend where the number of students and schools in public education are decreasing and in private subsidized education are increasing (Paredes & Pinto, 2009).

Market oriented system reforms are supported with the popular idea that private schools are more effective than public schools (Bellei, 2008). Several studies comparing private and public education have been made with Chilean data. They present mixed conclusions and use different methodological strategies, which make it hard to bring conclusions (Bellei, 2008). The most important limitation in all of these studies is selection bias (Bellei, 2008). In fact, to ascertain that there are really school effects it is necessary to control by several factors such as: socio-economical and academic segregation, peer effects, selection of students by schools and selection of school by families (Bellei, 2008).

Selection bias is a general problem regarding the identification of school effects, because students are not assigned randomly to their schools (Raudenbush & Willms, 1995). Therefore, there are many methodological issues to solve. A typical strategy to address this problem is controlling by confounding variables, which can be done using different statistical methods, for example regression analysis or propensity score analysis.
Nevertheless, there is another important issue to address the identification of school effects, that is studying students' trajectories. To the best of our knowledge, the only study that has compared students' trajectories from voucher and public schools is the thesis from (Ortega Ferrand, 2015). After controlling for students and school variables, she found that students from voucher and public schools make the same progress during primary, which is a finding that cannot be done modeling only final achievement measures (Ortega Ferrand, 2015, pp. 202).
In fact, studying trajectories improves the estimation process because it accounts for the measurement error, allows us to study the shape of the change and the rate of the growth (Bressoux, 2010). Guldemond and Bosker (2009) stated that studying trajectories increased the likelihood of detecting school effects because it is a more reliable measure of student progress. In the field of school effectiveness there is a call to do more research regarding trajectories of the same students over time because it is possible to understand the processes and characteristics producing stability and change in schools (Reynolds et al., 2014).
In summary, it is important to know about the characteristics of student trajectories, this is an information about progress, much richer than having only one data point.

The research presented in this part of the thesis aims to model long term trajectories in mathematics and reading scores from 4th to 10th grade. The sample is formed by Chilean students that attended public and voucher schools. We focus on students in public and voucher schools because those schools are the base of the system (they correspond to approximately 95% of all Chilean schools), are part of policy and academic debates and represent a more homogeneous population. The first goal is to describe the characteristics of achievement trajectories of Chilean students for different groups. The second goal is to compare the effect of attending voucher and public schools in students' trajectories in four steps: without controlling by confounding variables, controlling using linear adjustment, controlling using propensity score matching and controlling using both methods simultaneously. This extends the study of Ortega Ferrand (2015) using a different time span that includes secondary education, considering a census-based sample and focusing specifically in the selection bias problem between voucher and public schools.

In summary, the research objectives are:

1. The first objective is descriptive and it consider two specific objectives. The first specific objective is to determine what are the school effects on students' trajectories in mathematics and reading scores. The second specific objective is to characterize these trajectories comparing trajectories of different groups of students.
2. The second objective is to determine what is the effect of public schools in comparison with voucher schools on the students' trajectories in mathematics and reading scores. In particular we want to explore how the estimated effects change controlling for confounding variables using linear adjustment, propensity score matching and both methods simultaneously.

10.2 Context and Literature Review

10.2.1 School effects in students long-term achievement trajectories

School effects is a subject which has received a lot of attention because it may help to find strategies for the improvement of schools and educational systems. The main foci in this area are efficiency and equity (Bressoux, 2008). In general terms, school effects would be the contribution that a school gives to a student all other things being equal (Bressoux, 2008).
Considering the variance explained in achievement outcomes, school effects usually are small. In fact, Bressoux (2008) reports studies that estimate that, on average, school effects explain 8% and 13% of the variance of students' acquisitions and Guldemond and Bosker (2009) say that they explain about 10% to 30% of the variance. Nevertheless, this variance can differ heavily between countries. For example, in France the variance explained by schools is small, because schools are similar and principals do not have so much freedom in decision-making in comparison, for example, with Anglo-Saxons systems where the variance is higher (Bressoux, 2008). In the case of Chile, the variance explained by schools in mathematics and reading achievement is very high (see for example Manzi et al. (2014)). Some reasons to explain the huge variance of school effects in Chile are the high level of socioeconomic segregation between schools (Valenzuela et al., 2014) and the high freedom that school principals from private and vouchers schools have to make decisions in human and financial resources, curriculum, student policy and improvement (Weinstein & Muñoz, 2014).

Usually, the studies in school effects make use of multilevel models with the structure of equation (10.1) (Longford, 2012).

$$y_{ij} = \beta_0 + X_{ij}\vec{\beta} + \underbrace{u_j}_{\text{school effect}} + \epsilon_{ij}, \tag{10.1}$$

In this equation, y_{ij} is a measure of achievement of student i in school j, usually measured at the end of the school year. This achievement is explained with control variables $X = (x^1, \ldots, x^k)$ and the school effect u_j. The school effect can be modeled as a fixed or random effect. The goal in these studies is to identify school effects u_j controlling by confounders variables, as previous achievement and socioeconomic status, which are included in the vector X_{ij}. This is possible if the model is well specified and all the statistical assumptions hold, nevertheless researchers agree that these conditions are never completely met (Everson, 2017). When u_j is modeled as a random effect in equation (10.1), then it is possible to model the school effects with school level variables $X' = (x'^1, \ldots, x')$ as in equation (10.2).

$$u_j = X_j'\vec{\alpha} + \mu_j \tag{10.2}$$

In these models, the school effects are modeled over one measure of the variable y. Previous measures of y can be included as covariates in X. For example, when y is a reading-test score taken at the end of the year, it is possible to include in the control variables X a measure of the same reading test taken at the beginning of the year.

In these studies, students are the level one and schools are the level two or three, depending if the model considers class effects or not. It is important to note that equation (10.1) is a basic example, usually the studies deepen different aspects of the modeling according to methodological issues or to different research questions. There are several studies about school effects with Chilean data, for example Muñoz-Chereau and Thomas (2016) and Troncoso, Pampaka, and Olsen (2016) analyzed the importance of including class and municipalities random effects in the models. Another example is Manzi et al. (2014) that estimated value-added models with a focus in the endogeneity problem and how to correct it.

In contrast, few studies measure school effects in student progress and more research is needed in this area (Reynolds et al., 2014) . In this case, repeated measures of the outcome variable y are modeled as outcomes, and the modeling of change has to be explicitly described. In addition, a new level is added to the model, because repeated measures are considered as nested in the student. For example, in equation (10.3), the change is modeled as a linear function of time. This is different to the case of equation (10.1) that can use longitudinal data, but only the final measure is modeled and the other previous measures are used as covariates.

$$y_{tij} = \beta_0 + \underbrace{\varepsilon_{ij} + \beta_{1ij}t}_{\text{Modelling the change}} + X_{tij}\vec{\beta} + \overbrace{\mu_j}^{\text{school effect}} + \epsilon_{tij} \tag{10.3}$$

Guldemond and Bosker (2009) propose that multilevel growth models can better distinguish school effects because having several measures allow us to account better for the measurement error. Also, these models allows us to question about the school effect on the shape and the rate of the growth and control for time changing variables (Bressoux, 2010).

Nevertheless, modeling school effects in students' trajectories poses several research issues that influence the measured school effects. For instance, if there are more than three measures by person, it is possible to model different types of non-linear curves. Also, the time scale influences the type of school effect that is estimated and the research question that is finally answered (Anumendem, De Fraine, Onghena, & Van Damme, 2013).

Another very important issue is considering students' mobility between schools. Modeling school effects on student progress across one school year is easier because students usually do not change school in the middle of the year. However, if we measure progress across several years the possibility of changing school has to be addressed. This possibility can be modeled using three strategies:

1. Select only students that did not change school
2. Multilevel growth model with cross-classified school effects
3. Multilevel growth model with multiple membership school effects

The first strategy implies selecting only students that did not change schools and use a strictly hierarchical structure. The second strategy is to model a cross-classified structure, where for each year there is a different school clustering or a multiple-membership structure, where we consider that one student can belong to more than one school. To illustrate the difference between the strategy with cross-classified school effects and multiple membership models, we consider the present study where we have data for 4th, 8th and 10th grade. In this design, it is very plausible to have student changing schools. In fact, if we consider the students that have data for the three years, 75.2 % of them changed school at least once. The model for the first strategy is represented in equation (10.4). Its advantage is that it is simple, but can produce an important sample loss, an example is Guldemond and Bosker (2009).

$$y_{tij} = \beta_0 + \beta_{1ij} t + \underbrace{\mu_j}_{\text{school effects}} + \varepsilon_{ij} + \epsilon_{tij} \tag{10.4}$$

The second strategy implies modeling school effects with a cross-classified structure. In this case, we distinguish the effect of the same school in 4th, 8th and 10th grade (see equation (10.5)). It is like they are three different schools. This makes sense because the schools change, and it is important to see the effect of the school in different grades. Also, it allows us to model any time changing variable at the school level more easily, because for each year there is a specific random intercept at the school level which can be explained by the variables measured that year. An example of variables at the school level that change are compositional variables.

$$y_{ti(jkl)} = \beta_0 + \beta_{1i(jkl)} t + \underbrace{\mu_j^{4th} + \mu_k^{8th} + \mu_l^{10th}}_{\text{school effects}} + \varepsilon_{i(jkl)} + \epsilon_{ti(jkl)} \tag{10.5}$$

The third strategy considers the use of multiple membership models. In these models, for each student different weights are defined. The weights represent the multiple membership pattern. For example, if a student has been one year in School X and two years in School Y their weights would be $\frac{1}{3}$ for School Y, $\frac{2}{3}$ for School X and 0 for the rest of the schools. An example of a multiple membership model is described in equation (10.6). In this model, a student can receive the effect of different schools but there is only one estimated effect for each school. This can be very useful because it is simpler and considers the school as a stable institution. Nevertheless, there is a loss of precision because the model cannot distinguish the temporal order of the schools, for example these two schools combinations cannot be distinguished:

- 4th grade in School X, 8th grade in School X and 10th grade in School Y
- 4th grade in School Y, 8th grade in School X and 10th grade in School X

But being in School Y in 4th grade can be different than being in School Y in 10th grade. In fact, Goldstein and Sammons (1997) and Vanwynsberghe, Vanlaar, Van Damme, and De Fraine (2017) found that the variance at the primary school level was substantially different than the variance at the secondary school level in outcomes measured during secondary education. In both studies cross-classified models were used.

$$y_{ti} = \beta_0 + \beta_{1i}t + \underbrace{\sum_{j=1}^{N_J} \omega_{ij}\mu_j}_{\text{school effects}} + \varepsilon_i + \epsilon_{ti} \tag{10.6}$$

In Chile, at the best of our knowledge, the only study that measures school effects on student trajectories is the PhD thesis of Ortega Ferrand (2015). In this study, the author modeled student, teacher and school random effects in student progress, using an accelerated longitudinal design. She used a cross-classified multiple membership structure to model teachers and students random effects. She combined several sources of information and used data that expanded from 3rd grade to 8th grade, and found that a quadratic growth curve better explained the achievement trajectories in mathematics and language. Her findings replicate some trends from the literature: schools and teacher effects where higher and more sizable than in regular multilevel models; teacher effects where bigger than school effects and school effects in emerging economies are bigger than in post-industrialized countries. She found school compositional effects from achievement variables and SES variables. Nevertheless, these effects where only found in achievement status, not in growth. Also school effects varied across student groups (Ortega Ferrand, 2015, 185-187).
The scarceness of longitudinal studies done over student trajectories in Chile is understandable, because longitudinal studies are very expensive and more complicated to run. Moreover, most of the Chilean studies about school effects are done with data from the national assessment system SIMCE and the possibility of having three SIMCE measures of the same students appeared for the first time in 2013.

10.2.2 Chilean school system

In this subsection, we describe the foundations of the current Chilean school system and the principal social and political changes that have occurred in education until the year 2013. The data analyzed in this study is from students that in 2007 were in 4th grade. This implies that they entered the school system in 2003. Our last measure corresponds to the same students in 10th grade during the year 2013. Therefore, the goal is to understand the system and the specific dynamics that were present between 2003 and 2013.

In order to understand the Chilean context, first it is necessary to define the different types of schools that composed the system during this period:

- Public schools: Schools that are managed and funded by the state.
- Voucher schools: Schools that are managed by private entities, but receive funding from the state in form of a voucher for each student. They can be self-declared for profit or not for profit, this is a choice of the school owner. In these schools we can distinguish two types:
  - No family-fee: The schools that do not charge families with a fee. In this group there are schools that are only financed with the vouchers from the state and others have extra funding, for example from religious or charity associations.
  - Family fee: Schools that, in addition to the public funding, charge families with a fee.
- Private schools: Schools that are managed and funded by private entities.

The Chilean school system can be described as an example of profound and intense application of neoliberal ideas (Bellei, 2008). The reforms that originated this system started with a law promulgated in 1980, during dictatorship. This law passed the direction of public schools from the Ministry of Education to the municipalities, with the aim of decentralize the system, and created voucher schools (Corvalán & García-Huidobro, 2016). The basic idea behind this law was that voucher schools would receive a defined amount of money for each student that actually attends the school. This would produce that private entities could expand the number of schools and the competition for students would produce better educational quality and let the better schools grow and the worst schools perish (Corvalán & García-Huidobro, 2016). In this context, two components of the system have to be mentioned. The first one is that families can choose schools. This is a fundamental element for having schools competing for students. The second one is that private and voucher schools are allowed to select their students and public schools do not have the right to choose, unless they have less places than applicants (Contreras, Sepúlveda, & Bustos, 2010). This is a difference with other countries which have an important sector of private education but do not allow private schools to choose, for example Sweden, Netherlands and Belgium (Contreras et al., 2010).

All these components made the Chilean system special in comparison with other countries, because few countries had implemented in such a large scale and for many years a system so drastically based on neoliberal policies (Bellei & Cabalin, 2013).

In the following paragraphs we describe the main laws and reforms that defined the specific dynamics of the system within the period under study. After the creation of the voucher system and the end of the dictatorship in 1990, authorities looked for national reconciliation and tried to normalize the functioning of public institutions (Bellei & Vanni, 2015). In this sense, the democratic parties provided regulation and tried to improve the system instead of replace it.
One important change that received the system was in 1994 when voucher schools where allowed to charge a fee to families (Corvalán & García-Huidobro, 2016). This created price discrimination and promoted the development of private education (Bellei & Vanni, 2015).
Another very important development occurred in 1988 when census-based national assessments were implemented. The system is called SIMCE[1] and implements tests aligned with the school curriculum. These assessments have been applied every year (Meckes & Carrasco, 2010). The subjects evaluated more frequently are mathematics, language, sciences and social sciences. Nevertheless, measures about other subjects and regarding other variables have been gradually implemented.
The results of the tests at the school level have been disseminated for school principals and teachers, families, general public and researchers since 1995 (Meckes & Carrasco, 2010). The SIMCE system was conceived as a tool for schools, so no information at the student level is reported (Meckes & Carrasco, 2010). In addition, it is seen as a tool for families to choose schools. The use of SIMCE scores has been a controversial subject. On the one hand, it has been very useful for developing diagnoses of the system, design and evaluate public policies, target resources and define teacher incentives. On the other hand, it has also promoted selection practices from schools.

In addition to the establishment of voucher schools and the SIMCE evaluation system, it is important to consider the reforms and improvement programs that have been installed in Chile. Bellei and Vanni (2015) characterize different periods. They propose that between 1990 and 1995 there was a period of *educational improvement programs* which included several and diverse programs which aimed to change internal schools practices and teachers' work. They conclude that there is evidence of their impact on the schools functioning but their impact in student outcomes is not clear.
Then, the period 1996-2005 is the *reform to educational quality and equity* which describes a major educational reform that considers four major aspects: expansion and strengthening of the school-improvement programs from the past periods; several measures to support teachers; a curriculum reform and, finally, a radical increase in school time. This period also was characterized by an increase of the public investment in education, in part to funding the reform.
Several aspects of the reform to educational quality and equity had impact, but the results in the SIMCE of 2000 and the poor results in international tests produced a vision of failure which leaded to a period between 2000 and 2005 of *impact crisis and*

[1] Sistema de Información y Medición de la Calidad de la Educación.

the reorientation of educational reform. There was a perception that the reform was ineffective and students' results were stagnated. It appeared a loss of confidence in the school system as a whole and a sense of impotence from the government. New attempts were made to guarantee the continuity of the educational reform, adjusting its initial characteristics to be more effective. In this moment, the focus was to increase the results in standardized tests (Bellei & Vanni, 2015).

In this context of a negative view regarding the quality of education, there appeared in 2006 and in 2011 two huge social movements leaded, respectively, by high school students and university students that changed the public education agenda (Bellei & Cabalin, 2013; Bellei, Cabalin, & Orellana, 2014). The movement from 2011 is considered as the most important social movement in Chile after dictatorship (Bellei et al., 2014). These movements raised a demand for free education, the defense of public education, the reject of for-profit educational institutions and the rejection of schools' selection practices (Bellei et al., 2014). This leaded to the period between the year 2006 and 2013 defined by Bellei and Vanni (2015) as *The New Architecture of Chilean education.* In this period, it was recognized that only market dynamics cannot regulate the system to achieve quality and equity. This implied considering not only educational quality as a main issue but also equity and questioning the functioning of the system. This moved the political agenda and several laws and institutions where created. One of the most important is SEP law, which acronym means Preferential School Subsidy that was established in 2008. Its goal is to give extra funding for schools having students with disadvantaged socioeconomic conditions. SEP law defines priority students according to several socioeconomic criteria and gives 60% more funding for each priority student (Raczynski, Muñoz, Weinstein, & Pascual, 2016). Schools having a high percentage of priority students get an extra amount of funding. This law is for voucher and public schools and the extra funding is managed by the school administrator (Raczynski et al., 2016). In order to access to the funding, a school has to commit with several responsibilities as: do not select students, have a plan of school improvement and report to the government and the community the implementation of the improvement plan (Raczynski et al., 2016). The implementation of this law has been done progressively and it was not completely enacted in the year 2013. In fact, Valenzuela, Villarroel, and Villalobos (2013) report that even though the law has a good basis and probably is the most important reform of the last decade, the results in academic achievements have been heterogeneous and there should be some new policies to complement it.

In addition, in the year 2011, new institutions to monitoring the educational quality and regulate the management of resources were created. On the one hand, the Agency for the Quality of Education ACE[2] was created as the institution to ensure the quality of education in Chile. Its duty is to evaluate learning and educational institutions. This includes evaluating schools. The evaluations have consequences because if a school has sustained bad results it is intervened. In particular, the ACE manages the application, analysis and dissemination of the SIMCE results and the international assessments. On the other hand, it was also created the Superintendence of Education[3]. This institution monitors the use of resources from the schools.

[2] Agencia de Calidad de la Educación

[3] Superintendencia de Educación.

They have to inspect that the resources are used according to the law.

It is important to note that many of the reforms have improved the system. This can be seen in the increasing scores in the PISA tests. Nevertheless, there are three important characteristics of the Chilean system related to equity that should be considered. First, the number of students in public schools is decreasing steadily. Second, there is a huge amount of school socioeconomic segregation. Third, there is high stratification between students from private, voucher and public schools, where the students from public schools have the less advantageous conditions.

The contraction of public education has been reported in several research works (e.g. Paredes and Pinto (2009) and Bellei and Cabalin (2013)). To have an idea of what happened in the period considered in this study, the total number of public, voucher and private schools for the years 2004 to 2013 is presented in Figure 10.1. The number of voucher schools has been growing steadily and the number of public schools decreasing.

Regarding the segregation of the school system, the work of Valenzuela et al. (2014) with PISA data, shows that Chile is the country with the highest socioeconomic school segregation from the countries analyzed. Valenzuela et al. (2014) name as key factors to understand this segregation that vouchers school can charge fees to families and that schools can select their students.

The third main characteristic is the stratification between public, voucher and private schools. The fact is that students from public schools have much more disadvantageous socioeconomic conditions than students from voucher or private schools. This has been reported in several research works (e.g. Bellei (2008)) and we report some of the differences for the sample used in this study. In Table 10.1, it is possible to see the immense difference in socioeconomic conditions and in attendance to preschool education between students from public, voucher and private schools. We can see that students from voucher schools have parents with 2 more years of education in average and that the income is 50% larger. The access to computer and internet shows a very marked increasing trend across the years, probably reflecting that the cost of technology has been reduced. But again, students from voucher schools have more access to computers, internet and books. The differences remain the same in the three levels considered.

To summarize, the aim of this section was to understand the Chilean school system between the years 2003 and 2013. We can say that in those years the voucher system was stabilized and expanded and several programs to improve education where installed.

A turning point was in 2006 and 2011 where social movements started a social awareness about how highly stratified and segregated the school system is and the establishment of laws to decrease inequality and augment quality. The most important law enacted during these years was the SEP law. However, the implementation of this law has been slow and with some challenges to produce all the intended changes (Valenzuela et al., 2013). In fact, the years between 2006 and 2013 were identified by Bellei and Vanni (2015) as the period of *The New Architecture of Chilean education*.

As a conclusion, probably the years 2003 and 2013 represent a period where the

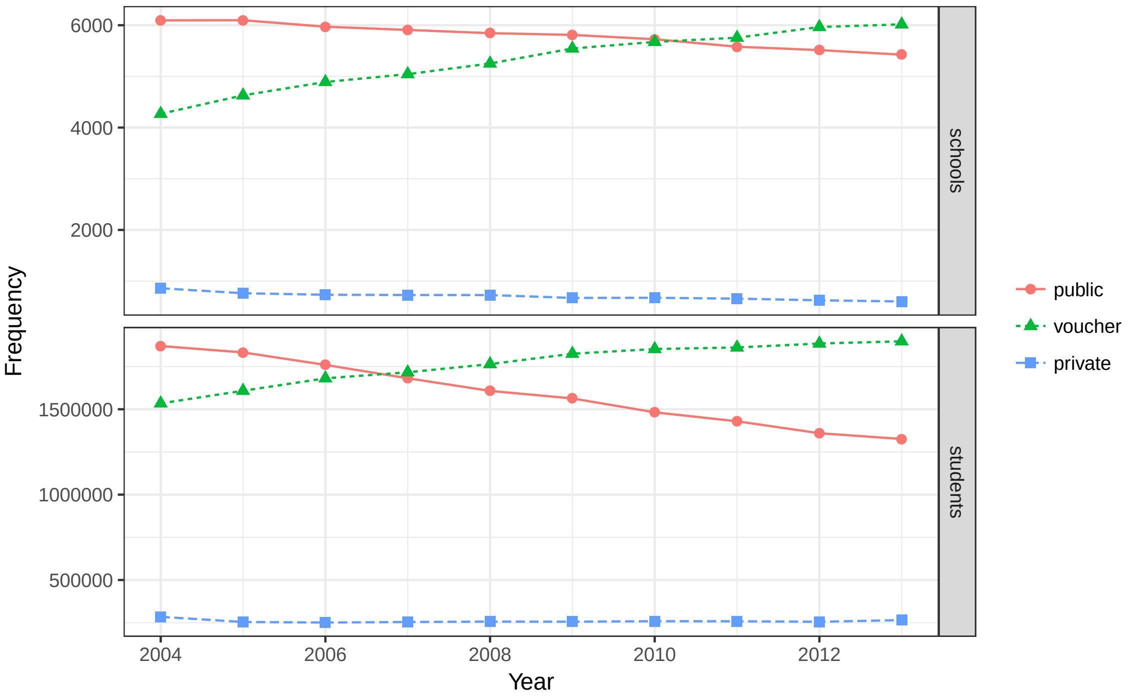


Figure 10.1: Number of public, voucher and private schools from 2004 to 2013

stratification and socio-economic segregation was in a peak with an inflection point from 2006 where several laws and reforms focused on improve the equity and quality of the system. With respect to the research question, we can hypothesize that the selection bias between public and private schools during 2003 an 2013 was very high because the enactment of the laws and reforms is slow.

10.2.3 Comparisons between voucher and public schools in Chile

In Chile, rough comparisons of student achievement show systematically that public schools have lower results than voucher schools. However, these differences cannot be attributed to the efficiency of each type of school, because of the huge disparity in student background characteristics (e.g. Table 10.1). Questions about the efficiency of voucher and private schools have been the focus of several studies and policy debates, mainly because it affects decisions regarding the allocation of resources.
Although, the studies about the effect of public and private education that have been done in Chile show mixed results regarding which type of school leads to better learning gains. Bellei, in his review, found that even studies using the same data set can have different conclusions (Bellei, 2008). In the studies that he reviewed, he identified several issues regarding the methodology to compare public and private schools. He concluded that the most important problem is the high selection bias between these two types of educational alternatives and identified the following confounding factors:

- Different socio-economic composition, this can have an effect at the student level, and also a composite effect at the class, school, or local authority level. In addition, it is important to distinguish between structural or economical variables and cultural or social variables.
- Different academic compositional, this can have an effect at the student level, and also a composite effect at the class or school level.
- School' selection from the part of the families.
- Students' selection from the part of the schools.

Several studies have deepened in each of these factors. Regarding socioeconomic and academic school composition, usually all the studies use socioeconomic variables at the student level as control variables. There are some studies that have deepen the analysis using aggregated variables and which have a focus on peer effects. For example, McEwan (2003) identified peer effects with Chilean data, where the strongest effect was from the class average of mother education, a socioeconomic variable.
Regarding school selection, there are variables that can be used to approximate the amount of school selection. For instance, the SIMCE assessment applies a questionnaire to students' parents or guardians and ask about which were the requirements to enter the school. Contreras et al. (2010) used this information and data regarding the geographical variability of schools to control for school selection of students. Their findings show that the voucher-public gap is very reduced after controlling by selection variables and they argue that an important part of the gap is related to voucher schools selecting good students, and not to the efficiency of voucher schools. This is in line with the findings from Bellei (2008).
Regarding families selecting schools, studies show that aspects that families consider in order to select schools are: distance from the house, quality of the schools and the social composition of the schools. Elacqua, Schneider, and Buckley (2006) studied the process surveying a random sample of parents from first grade students in the Metropolitan Region of Santiago. They asked which schools they would choose for their children and their criteria to select these schools. Then the authors collected objective indicators of each school and compared what the parents declare as important (*stated preferences*) with the real characteristics of the schools (*revealed preferences*). The surveyed parents reported that the academic characteristics were important to select schools. Nevertheless, they chose schools that differed in academic quality but were similar in socioeconomic dimensions. The authors summarize with the following quote "as parents choose schools in Chile, class - not the classroom - may matter more" (Elacqua et al., 2006, p. 578). In addition, the distance to the school is an important variable for school selection. This is supported with studies which survey the families and also with studies that analyze real decisions data (Chumacero, Gómez, & Paredes, 2011).

The studies described previously show the relevance and complexity of the confounding factors highlighted by Bellei (2008). In fact, the literature support that universal vouchers, voucher-school fees, school's selection of students and family's selection of schools increase segregation and stratification (Hanchane & Mostafa, 2012; OECD, 2012; Valenzuela et al., 2014). This implies that in Chile, it is more

difficult to assert that public and voucher schools are comparable and that robust estimation strategies should be used in order to attempt to compare both types of schools. These strategies should consider the selection bias problem and the multilevel nature of the variables that came from the students nested in schools.
New studies have compared public and voucher schools with novel methodological and statistical techniques. For example, Anand, Mizala, and Repetto (2009) compared public schools, free voucher schools and voucher schools that charge a family fee. They selected students with scholarships, enrolled in voucher schools that charge a fee and used propensity score matching in order to have comparable samples. They found that voucher schools that charge a family fee had test-score gains of 0.2 standard deviations over public schools. Nevertheless, the authors state that this identification strategy is limited because it adjust only for observed characteristics.
Another example is Lara, Mizala, and Repetto (2011) that used an strategy based on propensity scores and a changes-in-changes estimation method. They exploited the fact that there are public schools where there is only primary education (until 8th grade) and students have to change schools to continue secondary education. The sample were students that assisted to public schools until 8th grade. They compared students' SIMCE scores in 10th grade from students that stayed in public education with the scores from the students that switched to voucher schools. They found small effects in favor of voucher schools that were about 4% to 6% of one standard deviation.
Manzi et al. (2014) estimated the value added of the schools using panel data with two SIMCE measures. They used the parents educational level as instrumental variables to address endogeneity problems. They found that the effectiveness of public and voucher schools was different according to the use of an instrumental variable.
Finally, Zubizarreta and Keele (2016) compared voucher and public schools using a novel matching strategy that accounts for the clustered nature of the measures and matched students and schools. They compared SIMCE scores and found non significant differences between voucher and public schools and an effect size of 0.027 of a standard deviation, which is very small.

Regarding results from newer studies and including the review by Bellei (2008), we can conclude that in almost all of the studies, after controlling for confounding variables the voucher-public gap heavily decreases and sometimes changes its sign. In the case of Anand et al. (2009) and Lara et al. (2011) the effect is significant but small and in the case of Zubizarreta and Keele (2016) the effect is small and non-significant. However, none of these studies have addressed the effect of public and voucher schools on students' trajectories. The only study that analyzed trajectories is the thesis of Ortega Ferrand (2015). She studied trajectories from 3th grade to 8th grade and found that voucher schools did not have an effect on student achievement growth after controlling for student and school variables. This means that even public schools showed lower achievement levels than voucher schools, their students are making the same progress as the students in voucher schools (Ortega Ferrand, 2015, pp. 143, 202).

In conclusion, studies that compared public and voucher schools show that differences are negligible or slightly favor voucher schools. Nevertheless, comparing the effect of these types of schools on students trajectories is an open problem, which has only been addressed in primary grades by the work of Ortega Ferrand (2015) showing no effects on student growth.

Table 10.1: Mean differences in socio-economic variables and attendance to pre-school between students from public, voucher and private schools in 4th, 8th and 10th grade.

	2007 - 4th grade			2011 - 8th grade			2013 - 10th grade		
	Public	Voucher	Private	Public	Voucher	Private	Public	Voucher	Private
Years of father education	9,78	12,13	16,67	9,75	12,01	16,75	10,03	11,89	16,72
Years of mother education	9,73	12,07	16,17	9,77	12,02	16,24	10,13	11,95	16,22
Average income[a]	196.762	344.421	1.404.291	231.131	400.689	1.736.314	295.707	464.662	1.800.014
% with access to a computer	33	60	97	72	86	100	81	90	100
% with access to internet	10	30	89	47	69	98	57	75	98
Average number of books	20,24	32,29	72,01	29,66	42,19	81,57	34,26	44,87	82,73
% attended to ECE (0-2)[b]	5	8	14	7	9	14	7	9	14
% attended to ECE (2-4)[c]	29	40	80	33	41	80	37	42	80
% attended to Pre-kindergarten	43	56	85	48	60	87	51	61	88
% attended to Kindergarten	53	66	83	65	78	89	76	85	90
Number of students	118.049	115.583	16.586	109.604	115.193	17.895	94.349	118.372	18.013

[a] Chilean pesos

[b] Early childhood education for children with ages between 3 months and 2 years

[c] Early childhood education for children with ages between 2 and 4 years

Chapter 11

Methodology

11.1 Sample

The sample came from the census-based assessment implemented by the System of Measurement of the Quality of Education (SIMCE) (Meckes & Carrasco, 2010). The SIMCE tests aim to measure all the students assisting to regular education in a defined level (Agencia de Calidad de la Educación, 2015, pp. 63). The sample used in this study came from the SIMCE assessments done in 4th, 8th and 10th grade during the years 2007, 2011 and 2013 respectively. It forms a panel and represents Chilean students that where on track between 4th and 10th grade and that attended Public and Voucher schools. The sample was selected according to the following steps:

1. First, for each year we deleted about 7% of observations with invalid identifiers. They correspond to observations with duplicated identifiers or undefined identifiers. It remained a sample with 338.888 students.
2. Second, we selected the students that had valid identifiers for the three years, in total they were 157.814. Possible reasons for having students with missing identifiers can be having missed the test because they did not go to school that day or because they were delayed or overtaken. This selection is indispensable, because if the longitudinal identifier of an student is missing, also its school identifier is missing for that year.
3. Third, we selected students that attended only to public and voucher schools during the three years, which in total were 141.584. The deleted observations correspond to students that for the three measures were in private schools ($N = 12134$) and to students that have switched between the private system and the public and(or) the voucher system ($N = 4096$). The reason for deleting these students from the sample is that private schools are not part of the research objectives. In addition, students that switched between private, voucher and public schools correspond to 2.6% of the total. Thus, it was better delete them from the data base than include them in the analysis.

In summary, the final sample is formed of $N = 141.584$ students. In the Table 11.1, we specified the number of classes, schools, municipalities and regions of Chile for each year.

Level (year)	Students	Classes	Schools	Municipalities	Departments	Regions
4th grade (2007)	141.584	9.735	7.146	342	50	13
8th grade (2011)	141.584	7.971	5.424	341	54	15
10th grade (2013)	141.584	6.388	2.440	325	54	15

Table 11.1: Sample sizes for each level

11.2 Variables

The variables can be classified according to their level of measure (occasion, student, class, school, etc.) and according to their role in the model (outcome, control variable or identification variable). In Table 11.2, we present the level of measure, role in the problem and percentage of missing data for each variable. The percentage of missing data is measured according to the level, for example if it is at the student level is the percentage of students with missing values, if it is at the school level is the percentage of schools.

All these variables came from the SIMCE assessments of 2007, 2011 and 2013. For each year, we used the data bases at the student level, parent level and school level. In total, 9 databases were merged in order to have a unique data base. In the following subsections, we describe the different types of variables.

11.2.1 Time variable

The scale of the $time$ variable is presented in Table 11.3. The aim of the study is to measure the effect of the treatment on the trajectories. We can only model linear trajectories because we have three measures. A linear trajectory can be characterized with a slope and a intercept. We defined $time = 0$ for the measures taken in 10th grade. Therefore, the intercepts of the growth curves correspond to the students' test scores at 10th grade. The values for $time$ for the measures at 4th and 8th grade were assigned in order to have variance equal to one. We restricted the variance of the $time$ variable to one because with a larger variance we had estimation problems in the multilevel models.

Variable	Level	Role	% Missing
Time	occasion	time	0
Mother education	occasion	control	23
Father education	occasion	control	26
Number of books	occasion	control	22
Home Income	occasion	control	23
Computer	occasion	control	23
Internet	occasion	control	27
Parents expectations	occasion	control	25
Normalized score reading	occasion	outcome	7
Normalized score mathematics	occasion	outcome	7
Trajdep	student	treatment	0
Gender	student	control	0
Kindergarten	student	control	48
Pre-kindergarten	student	control	40
ECE (0-2)	student	control	33
ECE (2-4)	student	control	27
mrun	student	id	0
SES index 4th	school 4th	treatment	16
SES Selection index 4th	school 4th	treatment	16
Academic Selection index 4th	school 4th	treatment	16
Preschool attendance index 4th	school 4th	treatment	16
rbd 4th	school 4th	id	0
SES index 8th	school 8th	treatment	1
SES Selection index 8th	school 8th	treatment	1
Academic Selection index 8th	school 8th	treatment	1
Preschool attendance index 8th	school 8th	treatment	1
rbd 8th	school 8th	id	0
SES index 10th	school 10th	treatment	3
SES Selection index 10th	school 10th	treatment	3
Academic Selection index 10th	school 10th	treatment	3
Preschool attendance index 10th	school 10th	treatment	3
rbd 10th	school 10th	id	0

Table 11.2: Description of variables used in the multilevel models and in propensity score analysis

Grade	Time coding
4th	-2.405
8th	-0.802
10th	0

Table 11.3: Time coding

11.2.2 Treatment variable

The treatment variable *trajdep* defines the possible treatments. It is a categorical variable, where the categories are denoted with three letters indicating the belonging to the public or voucher system in each of the measured years. For example, the category *ppp* corresponds to students that attended public schools in 4th, 8th and 10th grade. The category *vpp* corresponds to students that in 4th grade attended a voucher school and in 8th and 10th grade attended a public school. In total there are $2^3 = 8$ possible treatments. The distribution of the sample in the different treatments is presented in Table 11.4. It is important to clarify that this variable does not vary at the school level, but at the student level because students can change schools.

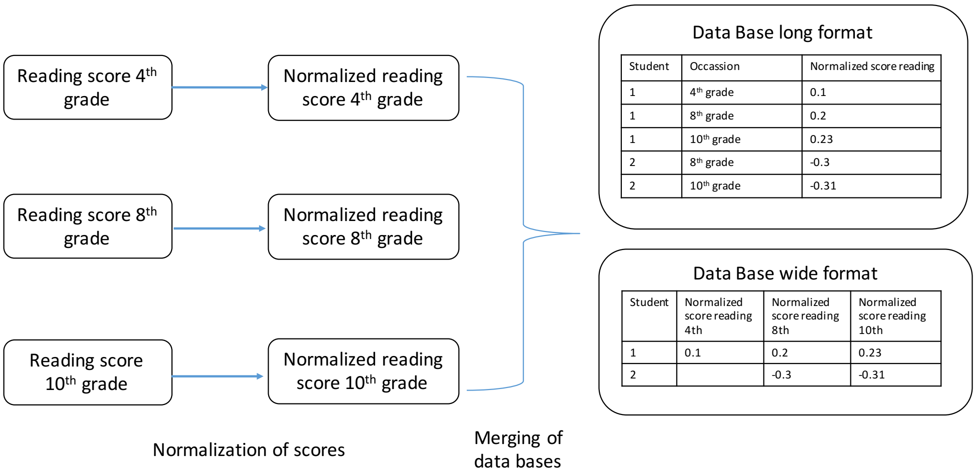
Treatment	N	%
ppp	39.699	28.04
ppv	3.909	2.76
pvp	1.943	1.37
pvv	10.041	7.09
vpp	18.608	13.14
vpv	2.282	1.61
vvp	7.827	5.53
vvv	57.275	40.45
Total	141.584	100

Table 11.4: Sample sizes for each treatment

11.2.3 Outcome variables: Student achievement measures

The outcome variables *Normalized score mathematics* and *Normalized score reading* are normalized scores in mathematics and reading from the national standardized tests implemented by SIMCE. The tests have the aim to evaluate the student learning on different subjects. For each subject, they are designed to evaluate the contents and abilities from the national curriculum (Agencia de Calidad de la Educación, 2015, pp. 3). They are applied close to the end of the school year and several levels are evaluated each year.

The SIMCE test scores are calibrated trough item response theory to compare the same level across different years. They are not designed to be compared longitudinally between different levels, that means it is not correct to model trajectories in the test scores given by SIMCE. This is why, for each year, we normalized the SIMCE test scores before merging the data bases. The process is represented in Figure 11.1. This implies that we forced the distributions of the scores to be normal in our sample and that the average trajectory of all the sample is constant and equal to 0.



Data Base long format

Student	Occassion	Normalized score reading
1	4th grade	0.1
1	8th grade	0.2
1	10th grade	0.23
2	8th grade	-0.3
2	10th grade	-0.31

Data Base wide format

Student	Normalized score reading 4th	Normalized score reading 8th	Normalized score reading 10th
1	0.1	0.2	0.23
2		-0.3	-0.31

Figure 11.1: Normalization of the SIMCE test scores

11.2.4 Control variables: Student socioeconomic and expectatives measures

The control variables corresponding to *mother education*, *father education*, *number of books*, *computer*, *internet*, *income*, *parents expectations* and *attendance to pre-school* came from the parents questionnaire. This questionnaire is sent with the student to their parents or guardians to be answered during the days of application of the SIMCE test (Agencia de Calidad de la Educación, 2015, pp. 59). Its design changes every year and we only considered the variables that were measured the years 2007, 2011 and 2013.

The variables *mother education*, *father education*, *income* and *parents expectations* were measured with alternatives, but before the analysis they were transformed to numeric variables according to the conversion tables from Agencia de Calidad de la Educación (n.d.). The conversion tables are reproduced in Tables 11.5 and 11.6. We did a similar re-coding for the variable *number of books in the house*, which is detailed in the Table 11.7. After this procedure, all the variables were used in the analysis as numeric variables. In addition, we divided the *income* variable by 100.000. This re-scaling was necessary because the original scale in Chilean pesos has a variance too large. Using the scale in Chilean pesos produced estimation problems for the multiple imputation and multilevel models.

The variables *computer*, *internet*, *Kindergarten*, *Pre-kindergarten*, *ECE 0-2*, *ECE 2-4* are dummy variables. In the case of *computer* and *internet*, they value 1 if the student has a computer or internet connection at his or her home and 0 if not. The variables *Kindergarten*, *Pre-kindergarten*, *ECE 0-2*, *ECE 2-4* have the value 1 if the student has attended to the corresponding level of preschool education and 0 if not.

Educational level	Imputed years of education
No studies	0
1th grade	1
2th grade	2
3th grade	3
4th grade	4
5th grade	5
6th grade	6
7th grade	7
8th grade	8
9th grade	9
10th grade	10
11th grade	11
12th grade academic track	12
12th or 13th grade vocational track	12
Incomplete short-cycle tertiary education	14
complete short-cycle tertiary education	16
Incomplete tertiary education (university)	15
complete tertiary education (university)	17
Master degree	19
Doctoral degree	22
Does not know or do not remember	It is not converted

Table 11.5: Conversion table for educational level, retrieved from (Agencia de Calidad de la Educación, n.d.)

11.2.5 Treatment variation: School aggregated variables

In Chile, school socioeconomic segregation is very strong, which implies that schools' composition can vary heavily. Therefore, it is very important to consider aggregate variables at the school level defined from the student-level variables.

The student level variables used to define school aggregate variables are socioeconomic variables, attendance to pre-school and variables about school selection. The socioeconomic variables and attendance to pre-school where already described and are listed in Table 11.2.
The variables about school selection came from the parent questionnaire. They correspond to questions regarding what did the school require to accept the student. For example, questions asking if the school applied a test, asked for a parent interview, required an income certificate, etc. Each question was coded as a dummy variable, so the mean is the fraction of the students in the school that were asked for the specific requirement.

To develop the school-level indexes, we calculated the school average for each variable, creating school-level aggregate variables. We denote these variables with

Income interval (Chilean pesos)	Imputed income (Chilean pesos)	Re-scaling
Under 100.000	50.000	0,5
Between 100.000 and 200.000	150.000	1,5
Between 200.001 and 300.000	250.000	2,5
Between 300.001 and 400.000	350.000	3,5
Between 400.001 and 500.000	450.000	4,5
Between 500.001 and 600.000	550.000	5,5
Between 600.001 and 800.000	700.000	7
Between 800.001 and 1.000.000	900.000	9
Between 1.000.001 and 1.200.000	1.100.000	1,1
Between 1.200.001 and 1.400.000	1.300.000	1,3
Between 1.400.001 and 1.600.000	1.500.000	1,5
Between 1.600.001 and 1.800.000	1.700.000	1,7
Between 1.800.001 and 2.000.000	1.900.000	1,8
Between 2.000.001 and 2.200.000	2.100.000	1,9
Over 2.200.000	2.300.000	2,3

Table 11.6: Conversion table for the monthly income imputed to the student's home, retrieved from (Agencia de Calidad de la Educación, n.d.). In addition, we precise the re-scaling done to the variable in order to have with lower variance.

Number of books interval	Number of books imputed to the student's home
No books	0
Less than 10	5
Between 10 and 50	30
Between 51 and 100	75
More than 100	120

Table 11.7: Conversion table for number of books at home.

the same name than at the student level, but we add a ^{sc} as superscript. In order to give the same importance to each variable we standardized each of them. To develop the indexes we did different principal component analysis, which results are described in the following list.

1. **School Socioeconomic Index**
   For each year, we made a principal component analysis with the variables *Mother education*^{sc}, *Father education*^{sc}, *Number of books*^{sc}, *Income*^{sc}, *Computer*^{sc}, *Internet*^{sc}. In all the analysis the results suggested one factor, where all the variables have high loadings. Therefore, we defined the School Socioeconomic Index as the standardized average of these variables (Table 11.8).
2. **School Socioeconomic Selection Index and School Academic Selection Index**
   For each year, we made a principal component analysis with the variables *Preschool certificate*^{sc}, *Grades certificate*^{sc}, *Test*^{sc}, *Marriage certificate*^{sc}, *Church-*

related certificate^{sc}, *Income Certificate*^{sc}, *Game sesion*^{sc} and *Parent interview*^{sc}. We tested Varimax and Promax rotations, in order to get more interpretable solutions. The results of the years 2007 and 2011 suggested extracting two components. In the first component the variables with high loadings where *Marriage certificate*^{sc}, *Church-related certificate*^{sc},*Income Certificate*^{sc} and *Game sesion*^{sc}. These are all variables related with social selection. In the second component the variables with high loadings where *Preschool certificate*^{sc}, *Grades certificate*^{sc} and *Test*^{sc} and can be interpreted as academic selection variables. In these solutions, *Parent interview*^{sc} had double loading, and after eliminating it, the structure was clearer. In 2013, the structure remains unclear because *Test*^{sc} with *Income Certificate*^{sc} load on the same factor. In spite of this, with the aim to have the same indexes for the three years, the indexes were defined considering the solutions from 2007 and 2011 (Table 11.8).

3. **Preschool attendance index**
For each year, we made a principal component analysis with the variables *Kindergarten*^{sc}, *Pre-kindergarten*^{sc}, *ECE 0-2*^{sc}, *ECE 2-4*^{sc}. All the analysis suggested to extract one component, excepting the one done with data from the year 2007. In this analysis, *Kindergarten*^{sc} had a very low loading (0.158). With the aim to have the same indexes for the three years, the index was defined considering the solutions from 2011 and 2013 (Table 11.8).

School Index	Variables	Interpretation
SES	*Mother education*^{sc} *Father education*^{sc} *Number of books*^{sc} *Home income*^{sc} *Computer*^{sc} *Internet*^{sc}	Schools with higher values in this index are composed with students with better socioeconomic conditions
SES Selection	*Marriage certificate*^{sc} *Church-related certificate*^{sc} *Income Certificate*^{sc} *Game session*^{sc}	Schools with higher values in this index asked more frequently for socioeconomic related certificates to accept a student
Academic Selection	*Preschool certificate*^{sc} *Grades certificate*^{sc} *Test*^{sc}	Schools with higher values in this index asked more frequently for academic related certificates or test to accept a student
Preschool attendance	*Kindergarten*^{sc} *Pre-kindergarten*^{sc} *ECE 0-2*^{sc} *ECE 2-4*^{sc}	Schools with higher values in this index are composed with students with more years of pre-school education

Table 11.8: School indexes definition

11.3 Data Analysis

In this study, the structure of the data is multilevel and the research objectives concerns school effects on students' trajectories. The most suitable models for this situation are multilevel growth models, which allow to model trajectories and data with a hierarchical structure. Also, they are flexible enough for modeling students' mobility between schools through cross-classified structures.

Two methodological problems appeared in this research: selection bias and missing data. The main problem of comparing public and voucher schools in Chile is selection bias (Bellei, 2008), and probably during the years under study the school segregation and stratification was in its peak, increasing the selection bias problem. This is why, we will address selection bias trough linear adjusting for the control variables and using propensity score matching. A summary of the data analysis strategies are presented in Figure 11.2.

The missing data problem is going to be addressed through multiple imputation techniques. The imputation models have to preserve the relevant characteristics of the data (Enders et al., 2016). In the case of our data, the hierarchical nature is a key aspect of the data structure. Measurement of the intra-class correlations of most of the variables using the grouping of schools in 4th, 8th and 10th grades were very high. This is why imputation techniques specific for multilevel variables where used. The description of the multiple imputation procedure is presented in appendix B.1.

The data can be formatted as a *person-level data set* or wide format where each row is a student. Also it can be formatted as a *period-level data set* or long format where each row is one occasion of measure of the student (Singer & Willett, 2003, pp. 17). Multiple imputation and propensity score matching was done over the *person-level data set*. Multilevel models were estimated over the *period-level data set*.

All the estimation processes were done using the software R (R Core Team, 2016). Propensity score matching was done with the Matching package (Sekhon, 2011) . For multilevel models we used the package lme4 (Bates, 2010) and for multiple imputation the packages mice (Buuren & Groothuis-Oudshoorn, 2011) and pan (Zhao & Schafer, 2016). Multiple imputation procedures and the estimation of multilevel models with cross-classified random effects was very demanding in terms of computational resources. This is why we used the super-computing infrastructure of the National Laboratory for High Performance Computing NLHPC (ECM-02).

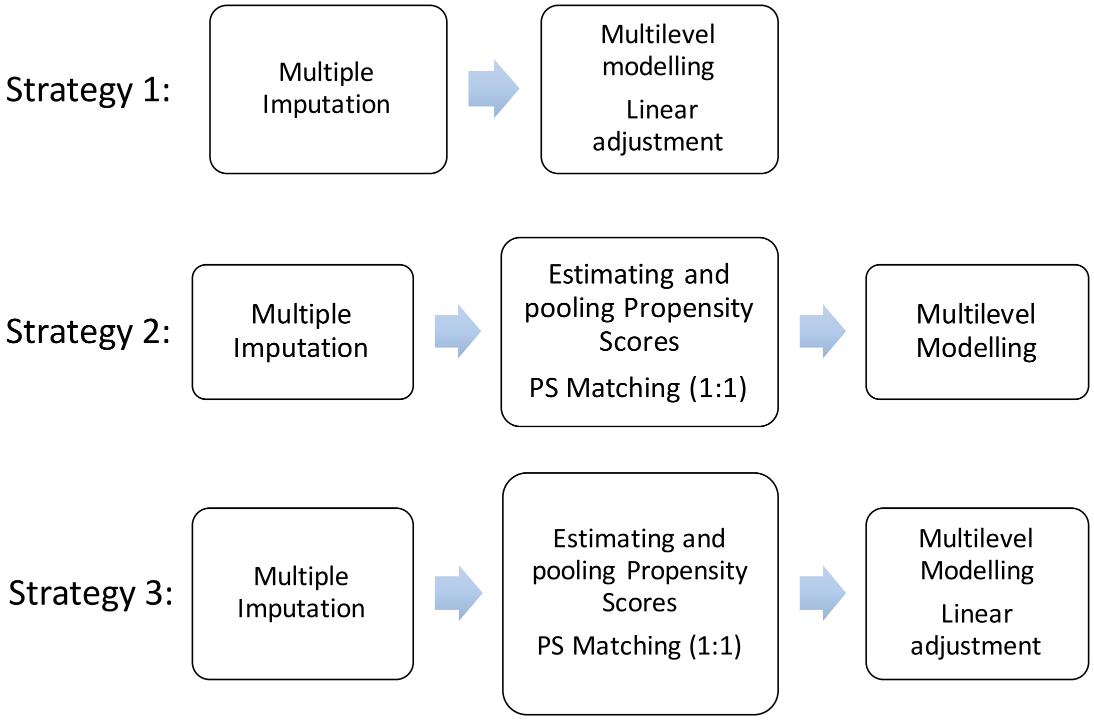


Figure 11.2: Statistical strategies to address selection bias and missing data.

Chapter 12

Results

In the first section of this chapter, we present a descriptive analysis of the variance distribution and the trajectories in mathematics and reading scores for different groups.
In the second section we present several estimations of the effect of the public system and the voucher system on trajectories in mathematics and reading scores. The section starts with raw comparisons of both groups. Then, we present the estimated effects of public education adjusting by confounding variables trough linear adjustment, propensity score matching and using both methods simultaneously.

In all the analyses presented in this chapter the nested nature of the data was modeled with multilevel models and missing data was handled with multiple imputation.

12.1 School effects and characteristics of achievement trajectories of Chilean students

In this section we present the results regarding the first research objective. This objective implies doing descriptive analyses and it consider two specific objectives. The first specific objective is to determine what are the school effects on students' trajectories in mathematics and reading. The second specific objective is to characterize the achievement trajectories in mathematics and reading of Chilean students.

12.1.1 School effects on students' trajectories

In this subsection we understand school effects as the variance explained by the random intercepts at the school level. To interpret these results, it is important to consider that the duration of primary and secondary education. In Chile, primary education starts at 1st grade and ends at 8th grade and secondary education starts at 9th grade and ends at 12th grade. Primary and secondary education are compulsory by law.

We present mean unconditional models to understand how the variance is distributed in mathematics and reading normalized scores (Singer & Willett, 2003, pp. 92-101). Tables 12.1 and 12.2 present the results for mathematics and reading respectively. Model 1.a is the simplest model and only specifies random intercepts at the student level. Models 1.b to 1.d define, respectively, random intercepts for the school at 4th, 8th and 10th grade. Model 1.e includes cross-classified random intercepts at the school level in 8th and 10th grade. Finally, Model 1.f specifies cross-classified random intercepts at the school level in 4th, 8th and 10th grade. The models' exact definitions are detailed in appendix B.2.

In all the models, the variance intra-individual or residual is lower than the variance inter-individuals, but remains very important. In the models for reading scores it represents about 35% of the total variance and for the mathematics scores it represents about 30% of the total variance.
The variance inter-individuals, or at the student level, changes if we add random intercepts at the school level. In Model 1.a, where there are not school random intercepts, the variance at the student level is about 64% and 70% for reading and mathematics scores respectively. If school effects from the primary schools (4th and 8th grade) and the secondary school (10th grade) are added (Models 1.e and 1.f) the student level variance descends to about 40 % in math and 38% in reading.

An important question is what is the size of the school level variances. In Models 1.b to 1.f we included school random intercepts for the school at 4th grade, 8th grade and 10th grade. Regarding the variance from the random intercepts for the different school clustering, it can be seen that it is important to define random effects at the school level, because all the models are considerable better than Model 1.a.
Models 1.b, 1.c and 1.d define school effects from the school at 4th, 8th grade and 10th grade respectively. In all the models the percentage of variance explained at the school level is important, ranging from 13% to 23% for reading scores and from 19% trough 31% for mathematics scores. Also, the percentage of variance explained from the school in 10th grade is bigger than the variance from the school in 8th and 4th grade.
Nevertheless, when the models include school effects simultaneously from the school at 4th grade, 8th grade and 10th grade (Model 1.f) the variance from the school at 4th grade vanishes. This is probably related with identification issues between the school in 4th grade and the school in 8th grade. In Chile, primary education ends at 8th grade, and several students change school to attend secondary education. Usually students do not change school during primary, in fact 69.7% of the sample had the same school between 4th and 8th grade. However, only 33.3% of the sample had the same school between 8th grade and 10th grade. The identification issue appears also in the amount of computer time used to estimate the models. When we added school random intercepts for 4th grade and 8th grade simultaneously, the computer time increased enormously.

In summary, considering the almost null variance from the school effects in 4th grade when school effects at the 8th and 10th grade are added and the identification problems, for the conditional models we will model only random intercepts from the school in 8th grade and in 10th grade.

Table 12.1: Variances of the random effects from the unconditional mean models for mathematics. The parameters are the mean trough 20 imputations

	Mathematics					
Parameters	1.a	1.b	1.c	1.d	1.e	1.f
Fixed effects						
Intercept	0 (0,00)	-0,15 (0,01)	-0,16 (0,01)	-0,07 (0,01)	-0,11 (0,01)	-0,11 (0,01)
Random intercepts						
Residual	0,31	0,31	0,31	0,31	0,31	0,31
Student	0,69	0,5	0,46	0,41	0,37	0,37
School 4th		0,19				0,02
School 8th			0,21		0,05	0,03
School 10th				0,32	0,24	0,24
Fit Indexes						
Deviance	993007	964524	955515	937151	932983	932142
df	424749	424748	424748	424748	424747	424746
AIC	993013	964532	955523	937159	932993	932154
BIC	993046	964576	955567	937203	933047	932219

Table 12.2: Variances of the random effects from the unconditional mean models for reading. The parameters are the mean trough 20 imputations

	Reading					
Parameters	1.a	1.b	1.c	1.d	1.e	1.f
Fixed effects						
Intercept	0 (0,00)	-0,09 (0,01)	-0,11 (0,01)	-0,04 (0,01)	-0,06 (0,01)	-0,07 (0,01)
Random intercepts						
Residual	0,36	0,36	0,36	0,36	0,36	0,36
Student	0,64	0,51	0,48	0,43	0,41	0,4
School 4th		0,13				0,01
School 8th			0,15		0,03	0,03
School 10th				0,24	0,2	0,2
Fit Indexes						
Deviance	1030107	1012073	1005941	990315	987987	987624
df	424749	424748	424748	424748	424747	424746
AIC	1030113	1012081	1005949	990323	987997	987636
BIC	1030146	1012124	1005993	990367	988052	987702

The second step for model specification was testing unconditional growth models, where we added random slopes (Singer & Willett, 2003, pp. 92-101). We tested random slopes at the student and school level. The exact definitions of the models are detailed in appendix B.2. In all the unconditional growth models we defined random intercepts and random slopes at the student level and cross-classified random intercepts at the school level at 8th and 10th grade. Model 2.a defines random slopes only at the student level, Models 2.b and 2.c include random slopes at the school level in 8th grade and in 10th grade respectively. Finally, Model 2.d includes random slopes at the student level and at the school level in 8th and 10th grade. How the time variable is coded in a way that $time = 0$ in 10th grade, the intercepts correspond to the scores at 10th grade. The results for reading and mathematics are presented in Tables 12.4 and 12.3.

When we add random slopes at the school level, the variances of the slopes at the student level decrease. The slopes' variances are small, but it is important to note that the size of these variances depends on how the time is coded. In fact, the difference in deviance from Model 2.a with Models 2.b, 2.c and 2.d shows that including random slopes at the school level is relevant.

In summary, from the analyses of Models 1.a to 1.f and 2.a to 2.d, the school effects on students' trajectories are significant.
With respect to the random intercepts, the most important school effects are from the school at 10th grade. The effect of primary schools on the intercepts are important. However, when the effect of secondary school is included, they decrease substantially. Regarding random slopes, it is relevant to model random slopes at the student level and at the school level in 8th and 10th grade. From these analyses, we decided to model the random effect structure from equation (12.1). We will use this structure in the following sections for the models that include explanatory variables.

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)}. \end{aligned} \tag{12.1}$$

Table 12.3: Fixed effects and variances of the random effects from the unconditional growth models for mathematics. The parameters are the mean trough 20 imputations

	Mathematics			
Parameters	2.a	2.b	2.c	2.d
Fixed Effects				
Intercept	-0,106 (0,011)	-0,111 (0,011)	-0,111 (0,013)	-0,107 (0,013)
Time	0 (0,001)	-0,005 (0,002)	-0,005 (0,003)	0,001 (0,003)
Random Effects				
Residual	0,246	0,246	0,246	0,246
Student	0,379	0,37	0,368	0,365
Student-time	0,04	0,028	0,027	0,022
Student time-intercepts	0,013	0,003	0,002	-0,002
School 8th	0,048	0,063	0,048	0,047
School 8th time		0,011		0,007
School 8th time-intercepts		0,013		0,002
School 10th	0,262	0,258	0,336	0,34
School 10th time			0,014	0,013
School 10th time-intercepts			0,05	0,053
Fit Indexes				
Deviance	925845	918094	913552	910656
df	424744	424742	424742	424740
AIC	925861	918114	913572	910680
BIC	925949	918224	913681	910811

12.1.2 Students trajectories in different groups

In this subsection, we aim to describe the achievement trajectories in mathematics and reading for different groups of Chilean students. In order to do this, we will test the following model:

$$\begin{aligned} &y_{ti(jk)} = \beta_0 + \beta_{02}x_{ti(jk)} + \beta_{1i(jk)}\text{time} + \beta_{12}\text{time} * x_{ti(jk)} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ &\beta_{1i(jk)} = \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned} \tag{12.2}$$

The results of models with the form of equation (12.2) allow us to describe how are the trajectories for different groups of students determined by the variable x. Nevertheless, they do not allow to do causal inferences regarding the effect of x on the trajectories, unless very strong suppositions are true. In this sense, these findings are only descriptive for the population of Chilean students attending to voucher and public schools.

To summarize the results, for different variables x, we estimated the model 12.2. We present graphics of the predicted trajectories for different groups defined by x. In the case that x is a categorical variable that varies only at the student level, like gender or attendance to preschool education, the groups are the ones defined

Table 12.4: Fixed effects and variances of the random effects from the unconditional growth models for reading. The parameters are the mean trough 20 imputations

	Reading			
Parameters	2.a	2.b	2.c	2.d
Fixed Effects				
Intercept	-0,064 (0,01)	-0,059 (0,01)	-0,058 (0,011)	-0,054 (0,011)
Time	0 (0,001)	0,004 (0,002)	0,005 (0,002)	0,01 (0,003)
Random Effects				
Residual	0,316	0,316	0,316	0,316
Student	0,425	0,419	0,415	0,414
Student-time	0,027	0,02	0,018	0,014
Student time-intercepts	0,017	0,01	0,007	0,005
School 8th	0,034	0,039	0,034	0,032
School 8th time		0,007		0,005
School 8th time-intercepts		0,006		0,001
School 10th	0,205	0,205	0,244	0,249
School 10th time			0,009	0,009
School 10th time-intercepts			0,025	0,029
Fit Indexes				
Deviance	985936	982546	979250	977841
df	424744	424742	424742	424740
AIC	985952	982566	979270	977865
BIC	986039	982676	979380	977996

by x. In the case that x is a continuous variable, like mother education or number of books, we plotted the trajectories at the first, second and third quartile of the variable to understand it relation with the outcome. In the case that x is a dummy variable and vary over time, the process is analogous.

The results for each model are presented in appendix B.3. We present a graphical summary of the models in Figures 12.1 and 12.2. Figures 12.1a, 12.1b, 12.1c and 12.1d compare the trajectories of students according if they attended or not to preschool education. We can see that students' trajectories in reading and mathematics scores show no differences between students that attended or not to ECE-02. Nevertheless, the trajectories are different for students that assisted to higher levels of pre-school educations. The largest differences are for the variable Kindergarten.

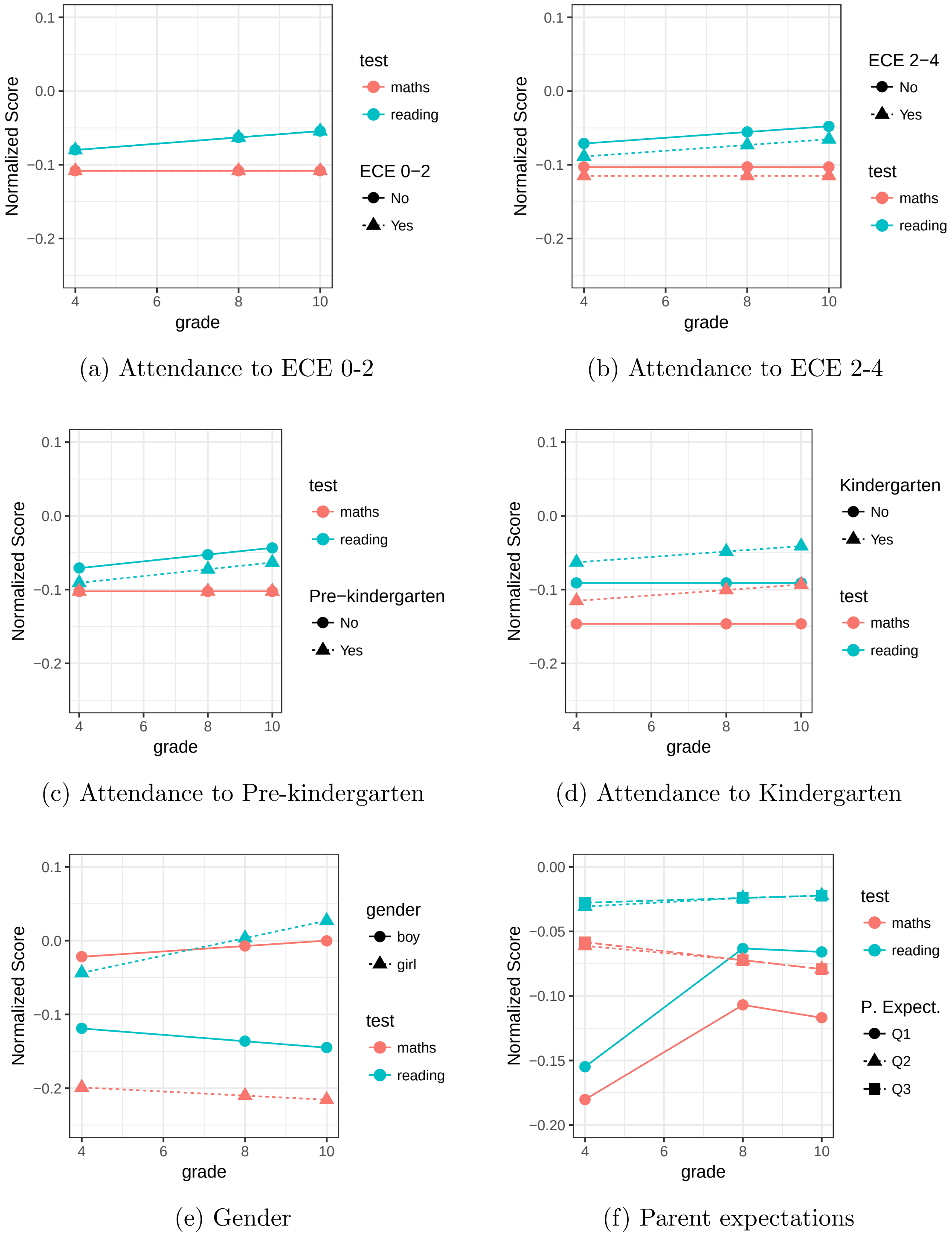
In Figure 12.1e we can see the predicted trajectories by gender. The figure show advantage for boys in mathematics and for girls in reading. The gap is bigger in mathematics than in reading and the differences are in intercepts and slopes. In mathematics, boys have a larger an positive growth rate and in reading girls have a larger and positive growth rate. The differences increase over time.

Regarding parents' expectations, Figure 12.1f presents the trajectories for the first, second and third quartile of the parents expectation distribution. We can see important differences between the first quartile with the second and third quartile.

The trajectories for different groups defined by the socioeconomic indicators and access to different resources are presented in Figure 12.2. Students' trajectories in mathematics and reading scores show large differences between the groups defined by mother education and father education. The largest differences are between the first quartile and the second quartile. This reflects the structure of inequalities in Chile.

Regarding the number of books, in 4th grade the biggest differences are between the first and second quartile and in 10th grade the biggest differences are between the second and third quartiles. The variables that show the largest differences between the trajectories are number of years of parents' education and number of books.

Having access to a computer produce differences on the trajectories at 4th grade, but not in 8th and 10th. This is because during 8th grade and 10th grade the quartiles do not variate for having a computer. In addition, there are no significant differences. Regarding access to internet, there are differences between the quartiles, but they are not important for the predicted trajectories. These trajectories are probably explained by the increasing access to technological resources during the last decade.



(a) Attendance to ECE 0-2

(b) Attendance to ECE 2-4

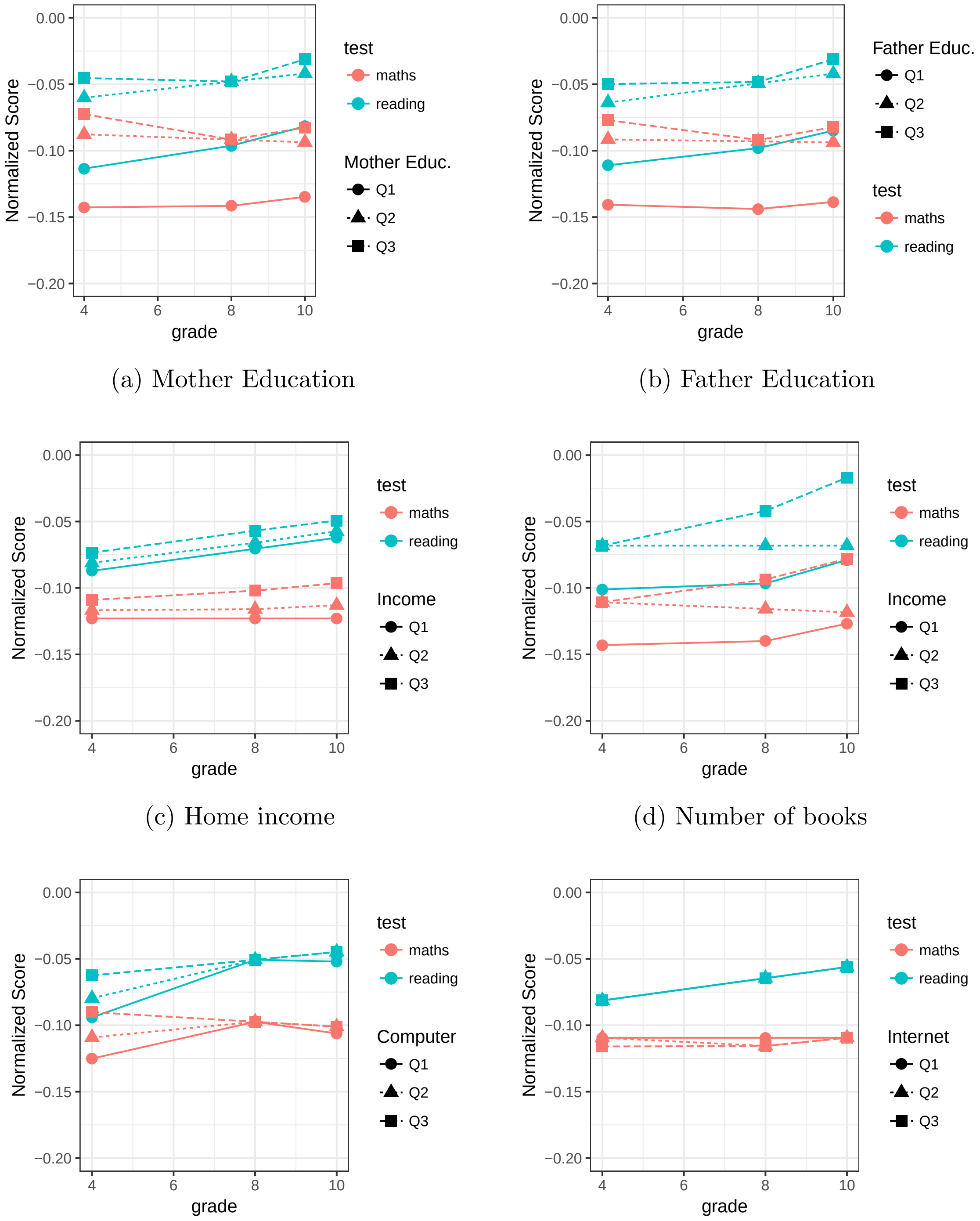
(c) Attendance to Pre-kindergarten

(d) Attendance to Kindergarten

(e) Gender

(f) Parent expectations

Figure 12.1: Predicted trajectories for students with different characteristics



(a) Mother Education

(b) Father Education

(c) Home income

(d) Number of books

(e) Access to computer

(f) Access to internet

Figure 12.2: Predicted trajectories for students with different characteristics

12.2 Comparison of public and vouchers education in students trajectories

In this section we present the results regarding the second research objective. We will measure the effect of public schools in comparison with voucher schools on students' trajectories in mathematics and reading scores. We will compare the estimated effects controlling for confounding variables using linear adjustment, propensity score matching and both methods simultaneously.
The results will be estimated over the subset of students that where enrolled in public schools for the three measurement points (ppp, N=39.699) or in voucher schools for the three measurement points (vvv, N=57.275). We decided to compare only this two groups because the sample sizes for patterns of students that switched from voucher to public schools, and vice versa, are small. Therefore, we could not guarantee to have enough common support. How the trajectories covers a time span of 5 years (from 4th grade to 10th grade), we are measuring the effect of being in the public school system in comparison with the voucher school system.

12.2.1 Background characteristics of the treatment groups

In this subsection, we present the characteristics of the two treatment groups in different variables. The aim is understanding the amount of selection bias. In addition, we will describe the treatments characteristics presenting descriptive statistics of school aggregated variables for students in public and voucher schools. Table 12.5 presents the means and standard deviations for each control variable between the students in public and in voucher education. These are pooled estimates across the 20 imputations. In addition, the standardized mean differences are presented. Ellis (2010, pp. 41) gives as a guideline that standardized differences of size 0.2 are small, of size 0.4 are medium and of size 0.8 or more are large. We do not present the statistical significance of the mean differences because all the differences where significant, which is natural because the sample sizes are very large. To see the distribution of these variables for students in voucher and public schools see Figures B.1, B.2 and B.3 in appendix B.4.

It is clear from Table 12.5 that in the complete sample there are differences between students attending public and voucher schools. Regarding the differences in access to learning resources such as internet, computer and books, they are from size medium to large and slightly decrease across time, specially the access to a computer.
The differences regarding socioeconomic conditions such as parents' education and income are large and stable for the three time measures. The only variable where the standardized difference is negligible is the percentage of girls. Nevertheless, in all the other variables the differences are sizable and negative. Regarding the attendance to pre-school, the standardized mean differences are medium to small. In summary there is a high amount of selection bias, that remains stable during primary and

secondary education.

Table 12.5: Descriptive statistics and standardized mean differences (SMD) for background variables. We compare students attending public and voucher education. Means are the pooled estimates from the 20 imputed data bases.

	Complete sample			Matched sample		
Variable	Public Mean (SD) 39699	Voucher Mean (SD) 57257	SMD	Public Mean (SD) 22342	Voucher Mean (SD) 22342	SMD
Computer 4th	0,32 (0,47)	0,66 (0,48)	-0,73	0,47 (0,49)	0,42 (0,49)	0,11
Computer 8th	0,75 (0,42)	0,9 (0,31)	-0,42	0,85 (0,36)	0,81 (0,39)	0,11
Computer 10th	0,81 (0,37)	0,93 (0,28)	-0,36	0,88 (0,32)	0,85 (0,34)	0,07
Internet 4th	0,09 (0,31)	0,35 (0,46)	-0,63	0,16 (0,36)	0,12 (0,33)	0,08
Internet 8th	0,46 (0,49)	0,74 (0,44)	-0,59	0,6 (0,48)	0,54 (0,49)	0,12
Internet 10th	0,53 (0,48)	0,79 (0,42)	-0,59	0,65 (0,46)	0,62 (0,47)	0,08
Father Education 4th	9,67 (3,32)	12,56 (3,26)	-0,88	11 (3)	10,5 (3,01)	0,15
Father Education 8th	9,64 (3,36)	12,52 (3,34)	-0,86	10,96 (3,06)	10,41 (3,08)	0,16
Father Education 10th	9,61 (3,39)	12,5 (3,4)	-0,85	10,92 (3,13)	10,39 (3,14)	0,16
Mother Education 4th	9,69 (3,27)	12,57 (3,13)	-0,9	11,1 (2,9)	10,57 (2,93)	0,17
Mother Education 8th	9,71 (3,31)	12,56 (3,19)	-0,88	11,09 (2,94)	10,52 (2,98)	0,18
Mother Education 10th	9,71 (3,32)	12,56 (3,22)	-0,87	11,06 (3)	10,54 (3,04)	0,16
Number of books 4th	19,76 (26,36)	35,06 (33)	-0,5	26,27 (28,92)	23,79 (27,59)	0,08
Number of books 8th	29,74 (28,96)	45,46 (35,36)	-0,48	36,43 (31,36)	33,74 (29,85)	0,08
Number of books 10th	32,15 (30,63)	48,67 (36,3)	-0,48	38,93 (32,5)	36,29 (31,19)	0,08
Income 4th	1,88 (1,78)	3,73 (3,08)	-0,7	2,38 (1,92)	2,16 (1,7)	0,08
Income 8th	2,23 (2)	4,47 (3,96)	-0,68	2,8 (2,25)	2,49 (1,83)	0,09
Income 10th	2,73 (2,5)	5,21 (4,45)	-0,66	3,37 (2,7)	3,09 (2,34)	0,08
Parents expectations 4th	14,47 (2,79)	16,41 (2,19)	-0,79	15,53 (2,41)	15,29 (2,51)	0,1
Parents expectations 8th	15,37 (2,3)	16,72 (1,75)	-0,68	15,95 (2,05)	16 (2)	-0,02
Parents expectations 10th	15,77 (1,99)	16,94 (1,55)	-0,67	16,24 (1,78)	16,34 (1,71)	-0,06
Gender (fraction of girls)	0,51 (0,5)	0,53 (0,5)	-0,03	0,51 (0,5)	0,51 (0,5)	0
ECE (0-2)	0,31 (0,45)	0,43 (0,48)	-0,26	0,36 (0,47)	0,33 (0,46)	0,05
ECE (2-4)	0,66 (0,44)	0,82 (0,39)	-0,37	0,73 (0,42)	0,74 (0,42)	-0,03
Pre-kinder	0,47 (0,49)	0,63 (0,49)	-0,33	0,53 (0,49)	0,51 (0,49)	0,04
Kinder	0,05 (0,21)	0,09 (0,29)	-0,17	0,06 (0,24)	0,05 (0,23)	0,03

Regarding the differences between treatments, Table 12.6 illustrates that, for the complete sample, the composition and selection practices of schools generate large and huge differences.

Students from voucher schools have peers with very different social composition. In fact, the effect sizes for these differences are $1,49$ and $1,57$, which is about twice the size of a large effect. Regarding the Attendance to preschool indexes, the differences are also large.

Standardized differences regarding academic selection and the socioeconomic selection indexes are between $0,63$ and $1,66$.

These results are a reflex of the system, where vouchers schools are allowed to select students but public schools are not, except in the case they have more applicants than places. To see the distribution of these variables for voucher and public schools see Figures B.4 and B.5 in appendix B.5.

Table 12.6: Descriptive statistics and standardized mean differences (SMD) for school indexes. We compare students attending public and voucher education. Means are the pooled estimates from the 20 imputed data bases.

	Complete sample			Matched sample		
Variable	Public	Voucher		Public	Voucher	
	Mean (SD)	Mean (SD)	SMD	Mean (SD)	Mean (SD)	SMD
	39699	57257		22342	22342	
SES Index 8th	-0,31 (0,54)	0,64 (0,64)	-1,57	-0,09 (0,5)	0,23 (0,6)	-0,52
SES Index 10th	-0,72 (0,55)	0,18 (0,64)	-1,49	-0,52 (0,52)	-0,21 (0,63)	-0,5
Preschool attendance Index 8th	-0,15 (0,75)	0,33 (0,64)	-0,7	0,02 (0,65)	0,1 (0,65)	-0,12
Preschool attendance Index 8th	-0,54 (0,78)	0,08 (0,67)	-0,88	-0,34 (0,67)	-0,18 (0,69)	-0,23
SES selection index 8th	-0,31 (0,28)	0,68 (1,45)	-0,89	-0,3 (0,26)	0,35 (1,18)	-0,59
SES selection index 10th	-0,53 (0,2)	0,28 (1,1)	-0,95	-0,52 (0,2)	0,03 (0,9)	-0,64
Academic selection index 8th	-0,38 (0,65)	0,87 (0,82)	-1,66	-0,22 (0,66)	0,56 (0,81)	-1,03
Academic selection index 10th	-0,17 (0,9)	0,37 (0,84)	-0,63	-0,07 (0,91)	0,17 (0,85)	-0,28

12.2.2 Modeling strategies to estimate the effect of public education

In order to estimate the effect of public education, it is necessary to define which confounding variables will be used to adjust for selection bias. We tried to consider all the types of variables that are described by Bellei (2008). A first group are socio-economical variables, that include structural or economical variables and cultural and social variables. These variables are considered at the occasion and student level (see Table 12.5) and at the school level trough the school indexes regarding SES school composition and preschool education school composition (see Table 12.6). The school selection is measured with school indexes about academic and socioeconomic selection (see Table 11.8).
In summary, we considered all the type of variables described by Bellei (2008), excepting a measure of academic school composition. Tables 12.5 and 12.6 supports that it is relevant to include these variables.

For the modeling, we distinguish between the student level variables and the school level variables. We consider the school-level variables as treatment variations because varied at the school-level and depend also on school level decisions.
Nevertheless, considering the segregation and stratification of the Chilean system and the work of Hanchane and Mostafa (2012), without school level aggregated variables probably the treatment variable is endogenous, or analogously the ignorability assumption is not plausible.

Another decision is considering the control variables only measured in the 1st time point (4th grade) or in the three time points (4th, 8th and 10th grade). Considering the variables only from 4th grade is reasonable because they are previous to the treatment. Nevertheless, considering the variables at the three time points also is reasonable because it permits to compare students that are similar in the background characteristics during all the treatment. We choose the last option, because we think that these background characteristics should be similar between groups during all the trajectory and probably they are not affected by treatment assignment. Nevertheless, we tested the models with only the variables from 4th grade and the results were similar.

Regarding the modeling strategy to adjust for selection bias, we consider three strategies: linear adjustment for confounding variables, propensity score matching and both methods simultaneously (where we used the matched data and in addition we added the confounding variables as covariates). All these options where implemented in conjunction with multilevel models, to be able to measure the treatment effect on the random slopes and random intercepts and also being able to include school aggregated variables. Models with the following structure where estimated:

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + X_{ti(jk)}\vec{\beta}_1 + \beta_{intercepts}Z_{i(jk)} + \beta_{1i(jk)}\text{time} + X_{ti(jk)}\vec{\beta}_{11} * \text{time} \\ &\quad + u_j^{8th} + u_k^{10th} + \eta_{i(jk)} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_2 + \beta_{slopes}Z_{i(jk)} + X'_j\vec{\beta}_2^{8th} + X''_k\vec{\beta}_2^{10th} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \\ u_j^{8th} &= X'_j\vec{\beta}_3 + \mu_j^{8th} \\ u_k^{10th} &= X''_k\vec{\beta}_4 + \mu_k^{10th} \end{aligned} \tag{12.3}$$

The outcome variable $y_{ti(jk)}$ is the normalized test score at time t of student i which in 8th grade attended school j and in 10th grade attended school k. How the time is coded in a way that the measure in 10th grade was assigned the value 0, the intercepts $\eta_{i(jk)}$ correspond to students scores at 10th grade. The students' slopes are denoted by $\beta_{1i(jk)}$. In the vector X there are the control variables at the occasion and student level. In the vectors X' and X'' there are the variables at the school level for the school at 8th and 10th grade respectively. The coefficients $\beta_{intercepts}$ and β_{slopes} are the estimated average treatment effect for public schools on the intercepts and the slopes respectively. We adjusted by selection bias with propensity score matching and with linear adjusting, including the linear effect of the confounding variables through the vector X. We included the effect of X in the intercepts and slopes.

The tested models are described in Table 12.7. We tested models with and without school aggregated variables (X', X''). Raudenbush and Willms (1995) define two types of school effects. Type A effects consider the effect of the school without distinguish if it came from the school context (for example being in a privileged neighborhood or having students with good previous achievement) or from good school practices (for example an effective staff). Type A effects are important for parents, which want that their kids have good results regardless if that came from the school practice or context. Type B effects are important for policy makers (Raudenbush & Willms, 1995). In our study, when we control only by student level variables, our estimation of the public school effect correspond to a Type A effect (Models 4.b, 5.a and 5.c). When we control by student level variables and school compositional effects and selection practices, our estimation of the public school effects correspond to a Type B effect (Models 5.b and 5.d).

Table 12.7: Models for treatment effect estimation.

Model	Adjusting of control variables	Inclusion of school characteristics
4.a	None	No
4.b	Linear adjusting	No
4.c	Linear adjusting	Yes
5.a	Propensity score matching	No
5.b	Propensity score matching	Yes
5.c	Linear adjusting and Propensity score matching	No
5.d	Linear adjusting and Propensity score matching	Yes

In subsection 12.2.3, we describe the details of the implementation of the propensity score matching. Then, in subsection 12.2.4 we present graphic summaries of the results of the models from Table 12.7. For each model, all the estimated parameters are presented in appendix B.6.

12.2.3 Application of propensity score matching

To estimate the propensity scores, first we fitted a logistic regression that predicted for each student the probability of being in public or voucher education as in the following equation:

$$\mathbb{P}(z_i = 1|X_i) = \frac{\exp(\beta_0 + X\vec{\beta_1})}{1 + \exp(\beta_0 + X\vec{\beta_1})} \tag{12.4}$$

To fit the model we used the data base in wide format, where each row corresponds to one student. We used as covariates in X all the control variables at the occasion and the student level described in Table 12.5. In addition, we add all the possible interactions between two variables and all the quadratic terms for the numeric variables. We used non parsimonious models because the focus for propensity score modeling is achieve balance more than parsimony.
For each of the 20 imputed data sets, we estimated the propensity score model and extracted the estimated propensity scores. Then, for each observation, we averaged the 20 propensity scores getting only one propensity score for each individual. With these estimated propensity scores, we matched students from public education with students from voucher education. The matching method was nearest neighborhood matching with a caliper equal to 0.2 times the standard deviation of the logit of the propensity scores (caliper=0.26). This caliper was chosen according to recommendations from Austin (2011). We choose nearest neighborhood matching because it was fast, which was an important property considering the sample size. Also because of the sample size, the results should be robust to the matching algorithm.

To evaluate covariate balance, we compared the standardized mean difference (SMD) between students in public education with students in voucher education in the complete and the matched sample (Pattanayak, 2015). The standardized mean differences before and after matching are detailed in Table 12.5. In all the variables, after matching the standardized mean differences were lower than 0.2 and we considered that the achieved balance was acceptable. We did not consider the significance of the mean differences because is not recommended, as it depend on sample sizes that varies for different matching strategies (Pattanayak, 2015).

After checking balance it is important to verify if there is common support between the groups (Thoemmes & Kim, 2011). Also, this illustrate how comparable are the groups. We plotted in Figure 12.3 the estimated propensity score for each group. We can see that there is a large amount of students in both extremes of the propensity score distribution which do not have matches. This explain the decrease of sample size in the matched data. The original sample has 39699 students from public schools and 57257 students from private schools. The matched sample has 22342 in each group, which implies a major reduction. Considering the high amount of selection bias, loosing so many units is not surprising. This also implies that without matching we are comparing several units without similar counterparts and that after matching the studied population changed. In Table 12.5 it is possible to see that the characteristics of the original sample and the sample after matching differ.

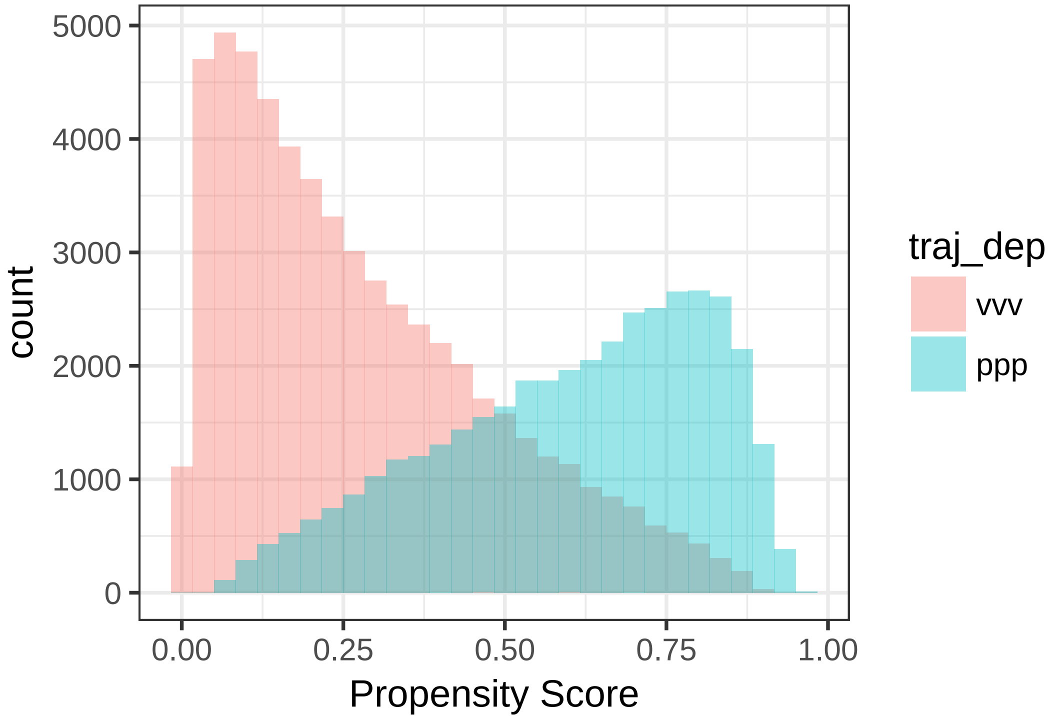


Figure 12.3: Common support of the estimated propensity scores

Finally, to estimate the average treatment effect we used multilevel models to account for the multilevel structure of the data. This allows us to model the effect on the intercepts and the slopes and to include variables regarding schools' characteristics. The results are described in the following section.

12.2.4 Effect of public education: combination of multilevel models, linear adjusting and propensity score matching

The results of the models for estimating the effect of public education on students trajectories are detailed in appendix B.6 (Tables B.6, B.7, B.8, B.9). Graphical summaries of the estimated effects of public education in intercepts and slopes are presented in Figure 12.4. For each model, the predicted trajectories for public and voucher education in math and reading are depicted in Figures 12.5 and 12.6.

We recall that model 4.a is a baseline model where no control for confounding variables was added. In Model 4.b there is linear adjusting for time varying and student level variables and Model 4.c includes, beside the control variables, school level variables.
Models 5.a, 5.b, 5.c and 5.d are done over the matched data. In the models 5.a and 5.b, the treatment effect is estimated without linear adjusting for control variables variables. Nevertheless, in Model 5.b school level variables are included.
Model 5.c and 5.d are doubly robust because, in addition of using the matched data, there is linear adjustment for the confounding variables. In Model 5.d also school level variables are included.

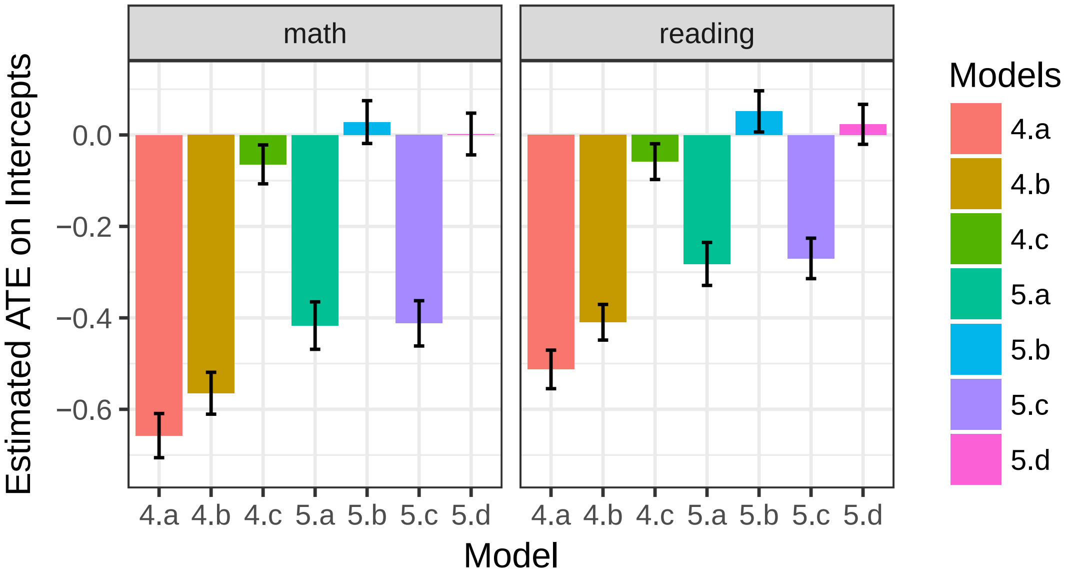
The estimated average treatment effects are presented in Figure 12.4. We can see that in the baseline Model 4.a, the differences between public and voucher schools are very important, showing lower intercepts and negative growth for student from public schools. The differences are bigger for mathematics than for reading.
When control variables at the occasion and student level are included trough linear adjustment (Model 4.b), matching (Model 5.a) or both (Model 5.c), the effect of public schools is less negative for the intercepts and equal or slightly more negative for the slopes. If the control is done trough propensity score matching, the differences between public and voucher schools are smaller.

With models 4.b, 4.a and 5.c we are measuring a Type A effect of public education (Raudenbush & Willms, 1995).
Nevertheless, if we want to measure Type B effects, that is the effect of the practices of the schools, these models are not correct. When we do not include school level characteristics the treatment variable is endogenous, because of the system stratification and segregation (see Hanchane and Mostafa (2012)). This is reflected by the data, where the most dramatic change in the estimated effects is when we add the school level aggregate variables (Models 4.c, 5.b and 5.d). These variables refers to preschool attendance, socioeconomic school composition and academic and socioeconomic selection practices. When these variables are added, the estimated treatment effects on the intercepts became very small. In addition, for the strategies with

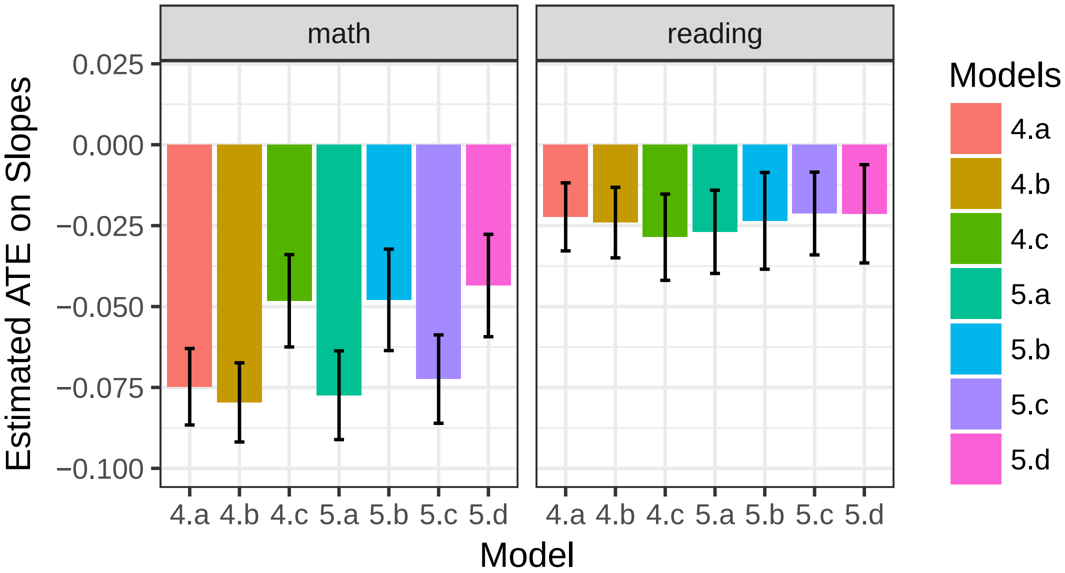
matching, the estimated effects become non significant. The same pattern is present for mathematics and reading. On the other hand, for the effects on the slopes, when the school characteristics are added they decrease notoriously for mathematics but not for reading.

In the schools aggregated variables, the SES index has by far the biggest effect on the trajectories, showing that socioeconomic composition has a strong effect on student trajectories (see Tables B.6, B.7, B.8, B.9 in appendix B.6).

To see how the estimated treatment effects are transformed in predicted trajectories, we plotted the predicted trajectories for the different models in Figures 12.5 and 12.6. We can see that, without controlling, voucher schools have better trajectories, but the differences start from the beginning. In mathematics the difference on slopes between public and voucher schools is visible.
The differences between public and voucher schools get smaller after controlling for confounding variables at the occasion and student level trough linear adjustment (Model 4.b), propensity score matching (Model 5.a) and both methods (Model 5.c). In addition, after controlling for confounding variables the difference in slopes is visible for mathematics but nor for reading.
After adding the school level variables, there are no relevant differences between the trajectories (Models 4.c, 5.b and 5.d). In fact, the differences in slopes are significant, but visually it is clear that they are not relevant. In conclusion, there are visible differences in the trajectories from students in public schools and voucher schools when controlling only by background student variables. Nevertheless, these differences appear to be explained completely by school compositional effects and school selection practices.

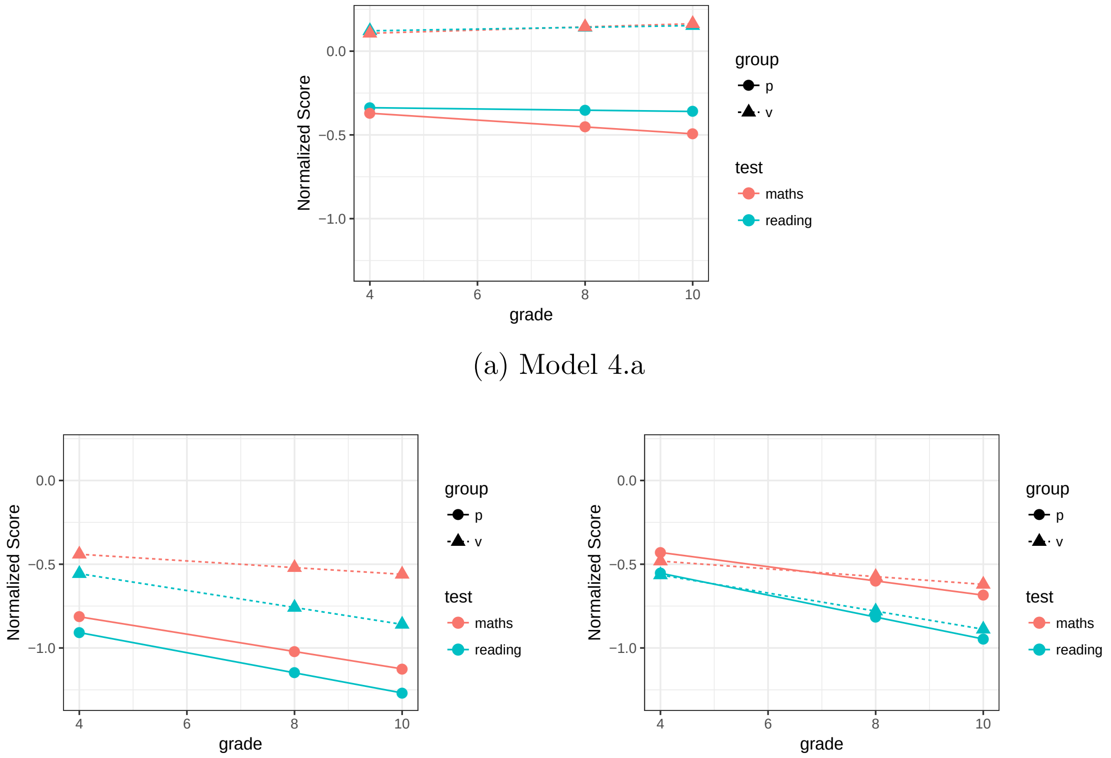


(a) Estimated ATE on intercepts. Bars represent 95% confidence intervals.



(b) Estimated ATE on slopes. Bars represent 95% confidence intervals.

Figure 12.4: Estimated ATE on reading and mathematics scores using different models.

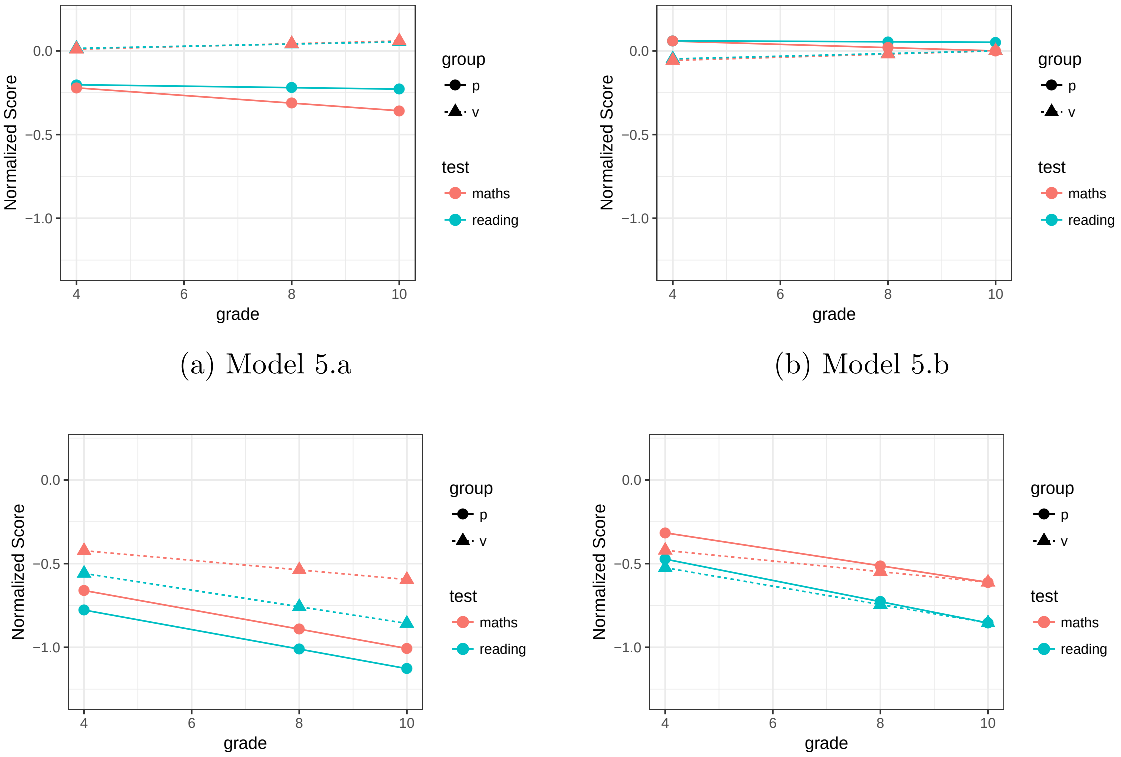


(a) Model 4.a

(b) Model 4.b

(c) Model 4.c

Figure 12.5: Predicted trajectories for students with different characteristics



(a) Model 5.a

(b) Model 5.b

(c) Model 5.c

(d) Model 5.d

Figure 12.6: Predicted trajectories for students with different characteristics

Chapter 13

Discussion and Conclusions

In this chapter we present a discussion about the results regarding school effects in trajectories of Chilean students. We modeled three repeated measures of mathematics and reading tests scores taken in 4th, 8th and 10th grade. In Chile, primary education starts in 1st grade and ends in 8th grade and secondary education starts in 9th grade and ends in 12th grade. The 5 years time span considered in this study covers students from the middle of primary to the middle of secondary.
The first objective was to describe how the variance of the test scores were distributed among the intra-student level, the inter-student level and the school level. We modeled linear trajectories for each student, thus, each student trajectory can be characterized with an intercept and a slope. The time coding was defined in a way that the intercept correspond to the student score in 10th grade. In particular, we wanted to understand the size of the school-level variances on the intercepts and slopes of students' growth in mathematics and reading scores. Also, we wanted to describe the average trajectories for different groups of students.
The second objective was to measure the effect of being in the public school system and in the voucher school system in students' trajectories. For this, we controlled for confounding variables using linear adjusting, propensity score matching and doubly robust estimation.

Regarding the first research objective, the results from the unconditional mean models and the unconditional growth models show important differences in student trajectories among schools. We tested several unconditional mean models. First, we modeled only random intercepts for the student level (Model 1.a), this model had considerable worst fitting than the models that included school effects. Then, we model independently random intercepts from the school at 4th, 8th and 10th grade (Models 1.b, 1.c and 1.d). These models show large and increasing variances for each school level. The variances at the school level at 4th, 8th and 10th were respectively for mathematics 19%, 21% and 32%; for reading they were respectively 13%, 15% and 24%. However, when we modeled cross-classified school effects, the variances explained by the primary schools dramatically dropped (Models 1.e and 1.f). The final unconditional mean model only included the school effects at 8th

and 10th grade. For mathematics the percentage of explained variance was 5% for the schools at 8th grade and 24% for the schools at 10th grade. For reading, it was 3% for the schools at 8th grade and 20% for the schools at 10th grade (Model 1.e, Tables 12.1 and 12.2).
After, we tested unconditional growth models which include random slopes. The variances of the random slopes between schools in primary and secondary were significant. This shows that there are differences between primary and secondary schools on the rate of the students' growth. These results coincide with the study from Ortega Ferrand (2015, pp. 118) done over trajectories during primary. For mathematics and reading, the slope variance of the schools at 8th grade was about half of the slope variance of the schools at 10th grade.
For the random intercepts, the drop of the school-level variances for the school at 4th and 8th grade can be attributed to an identification issue. In Chile, primary ends at 8th grade and most of the students do not change school during primary years. Therefore, we cannot distinguish the effects from the schools at 4th and 8th grade. However, this does not explain that in the model with school effects at 8th and 10th grade the variance of the school at 8th grade dropped to 5% in mathematics and to 3% in reading (Model 1.e). These results show that we cannot measure the effect of primary school in an outcome at secondary without including the effect of the secondary school because they change dramatically. However, despite the decrease in values after including secondary school effects, the primary school variances remain relevant and show a long term effect of primary schools.
Goldstein and Sammons (1997) and Vanwynsberghe et al. (2017) modeled cross-classified school effects from primary schools and secondary schools in outcomes measured in secondary education. Goldstein and Sammons (1997) found larger variation from primary schools than for secondary schools. We found the opposite pattern for the variation in intercepts and slopes. Regarding the intercepts, our results are in line with Vanwynsberghe et al. (2017), and even if the variances for the primary school level are small, they are relevant.
In line with trends of the literature, the school variances of slopes and intercepts are larger for mathematics than for reading. Also the school variances of intercepts are large. In Chile there are several reasons that produce high school level variances. First, there is an enormous school socioeconomic segregation (Valenzuela et al., 2014). Second, principals from voucher schools have autonomy in administrative and pedagogical matters (Weinstein & Muñoz, 2014).

Regarding the characteristics of the trajectories, we plotted the predicted trajectories for different groups of students according to attendance to preschool, gender and socioeconomic variables.There are no differences in the trajectories regarding the attendance to ECE (0-2)[1]. Nevertheless, there are differences regarding the attendance to ECE (2-4)[2], Pre-kindergarten and Kindergarten. Students that attended to pre-school education have better trajectories. Attendance to Kindergarten shows larger differences than the other levels. Students that attended Kindergarten have higher average and growth rate in mathematics and reading.

[1] Early childhood education for children with ages between 3 months and 2 years

[2] Early childhood education for children with ages between 2 and 4 years

The variables which show the largest differences between the groups are gender, parents expectations, parents education, and number of books at home. The gender gap is large for mathematics and reading trajectories. Already in 4th grade boys have advantage in mathematics and girls in reading. The gap increase for both measures and in 10th grade they are very important. These results coincide with the TERCE study showing that the gender gap increase across time in Chile (Gelber, Treviño, Inostroza, & others, 2016). The differences in trajectories for parents' expectations, mother education and father education are large when we compare the first quartile with the second quartile. These differences reflect the structures of inequalities in Chile, where there are larger differences between the first and second quartiles than between the second and third quartiles. Regarding parents' expectations, higher education in Chile is very expensive, which implies that parents expectations should be very correlated with socioeconomic conditions. This fact should explain in part the observed differences.

Students with access to computer and internet have different trajectories than the ones without access, but the differences are not relevant in 10th grade. These are good news, probably explained by the greater access to technological resources (see for example Table 10.1).

These results do not show a causal effect of the variables. For example we cannot say that attendance to Kindergarten produce better trajectories between 4th and 10th grade, because probably that effect is confounded with unmeasured variables. Nevertheless, they illustrate the inequalities of the system, in particular the gender and socioeconomic differences.

The second research objective was to measure the effect of public schools and voucher schools in students' trajectories in mathematics and reading. We considered a time span from 4th to 10th grade (5 years) and we compared trajectories of students that were in the public system or in the voucher system for the three measures. Therefore, we are measuring an effect of the public school system over the voucher school system for students maintained in one of the two system for 5 years.

First, students from public and voucher schools vary a lot on background variables. We found absolute standardized differences from 0,36 to 0,88 for socioeconomic related variables and from 0,17 to 0,37 in attendance to pre-school education. Second, the most dramatic differences are between school aggregated variables, where the absolute standardized differences range between 0,63 and 1,66. Considering usual criteria to interpret standardized differences, several of these values can be considered large and illustrate the high level of stratification of the Chilean educational system.

Before continuing, it is convenient to recall the definition of Type A and Type B school effects Raudenbush and Willms (1995). Type A effects consider the effect of the school without distinguish if it came from the school context (for example being in a privileged neighborhood or having students with good previous achievement) or from good school practices (for example an effective staff). Type A effects are important for parents, which want that their children to have good results regardless if that came from the school practice or context. Type B effects are important for policy makers (Raudenbush & Willms, 1995). In our study, when we control only by student level variables, our estimation of the public school effect correspond to a Type A effect (Models 4.b, 5.a and 5.c). When we control by student level variables, socioeconomic school composition and school selection practices, our estimation of the public school effects correspond to a Type B effect (Models 5.b and 5.d).

We first analyze the effect of public education on the intercepts of the trajectories, that correspond to the test scores in 10th grade and then we analyze the effect on students slopes. Without controlling for background variables, the estimated effect of public schools is students intercepts is negative and large. The effect is larger for mathematics than for reading. When we control by students background variables, the effect of public schools in the students intercepts became less negative, but of an important size (Models 4.b, 5.a and 5.c , Figure 12.4). These estimates correspond to Type A effects and include the effect of the school context and the school practices (Raudenbush & Willms, 1995). They are important for the families which want effective schools regardless of what explains the differences. Then, when we added the school aggregated variables the effect of public schools dramatically change and became null for mathematics and reading scores (Models 4.c, 5.b and 5.d, Figure 12.4). This pattern occurs for the three estimation methods. These estimates correspond to Type B effects, under the assumption that we do not have endogeneity. They are important for policy makers which want to evaluate school practice.

With respect to the students slopes, the effect of public schools are negative and significant. They do not change systematically when we add students background variables and school aggregated variables. However, the results imply that for each year of education the student normalized scores in mathematics decrease between $0,02$ and $0,03$ and for reading the decrease is about $0,01$. These differences are not relevant, which is clear in the plotted trajectories in Figures 12.5 and 12.6.

If we aim to estimate Type B effects, in the models without school aggregated variables the treatment variable is endogenous, so the estimated treatment effect is not valid (Bellei, 2008; Hanchane & Mostafa, 2012; Valenzuela et al., 2014). This also shows that the use of multilevel models is indispensable because permits to add the school level variables.

Regarding the use of propensity score matching, linear adjustment or both methods combined. The conclusions are the same for the three methods. We could argue that the propensity score analysis did not give much new information. Nevertheless, this analysis evidenced a very large group of students which are not comparable. Figure 12.3 shows summarily that public schools are in charge of educating the students with the most disadvantaged socioeconomic conditions. This is depicted in detail for each control variable in Figures B.1, B.2 and B.3 in appendix B.4.

Therefore, it is not posible to measure the effect of voucher schools in the group of students with the most disadvantaged socioeconomic conditions, because these schools are not educating this population. Also we cannot measure the effect of public schools in the group of students with more advantaged socioeconomic conditions.

When we only control for student background variables the estimated effect of the public schools was more negative for linear adjusting. This can be explained with two reasons. First, it could be an effect of extrapolating results for a large number

of students which are not comparable. Second, it can just be an effect of including only linear effects of the control variables.
In summary, the present study suggest that there are no different effects of the public system versus the voucher system on students trajectories between 4th and 10th grade. These is in line with the conclusions from Ortega Ferrand (2015). In addition, this coincide with several studies that have found no relevant or not significant differences between public and voucher schools using one final outcome measure (Anand et al., 2009; Bellei, 2008; Lara et al., 2011; Zubizarreta & Keele, 2016). We found a huge effect of school socioeconomic composition and selection practices which explain the differences between public and voucher schools.

Future research

Regarding primary and secondary school effects in students trajectories, more complex models that take account simultaneously the multiple membership structure and the cross-classified structure could be explored. This could enable us to understand better the relation between primary and secondary school effects. For example, the cross-classified multiple membership model described by Sun and Pan (2014) could allow modeling the effect of the schools at 4th grade, 8th grade and 10th grade simultaneously. Also the scale of the score is relevant, we modeled normalized scores for each year. This implies modeling the positions of the students in relation with the rest of the population.

Future research should consider schools effects in variables that are not achievement variables, for example the school climate and motivational variables (Reynolds et al., 2014). It is possible to study these issues in Chile with census-based samples because the institution in charge of SIMCE has been including assessments on other indicators of educational quality. For example, the Indicators of Personal and Social Development[3] which include measures of academic self-esteem and motivation, school climate, healthy living habits, civic participation, equity, school attendance, retention and school completeness (Agencia de Calidad de la Educación, 2017).

With regard to future research in the study of voucher and public schools , a limitation of this study is that we did not include school aggregated variables about academic composition. This factor is highlighted by Bellei (2008) as a source of selection bias. Also, this study was based in that the assignment process depend only on observed variables. Other methods which do not have these hypotheses could be used, for example instrumental variables. Future research could also model more levels, for example modeling fixed effects from the municipality level to account for that source of bias. Muñoz-Chereau and Thomas (2016) shows that the municipality level is important in Chile. In our study we did not included the municipality level because it had low variance and it was technically complex.

In addition, we should understand which schools produce better learning gains

[3] Indicadores de Desarrollo Personal y Social

inside the public system. The same for the voucher system. Also, non cognitive measures should be considered. Special attention should be given to public schools, because they are educating the students with more disadvantageous conditions. The segregation of the system is harshly harming these students. The large negative effect of public education was finally explained by the school aggregated variables. If we want to stop impairing these students the mechanism producing this segregation should be eliminated. An attempt to do this is the SEP law, but probably this law is not enough to fix the system. There is a need for reforms that improve the capacities of teachers and directors, reinforce public education, and the inclusion of other social systems that share the caring of the students should be considered (Valenzuela et al., 2013).

General Discussion

Chapter 14

General Discussion

In this chapter we summarize the results and discuss future perspectives. The research for this thesis is framed in the field of educational effectiveness. It contributes to the field with an statistical discussion and two empirical studies.
Part I of this thesis correspond to the statistics part. Its objective was to offer a discussion about how to combine statistical methods to address selection bias and missing data problems in the context of multilevel models. For selection bias, we studied linear adjustment and propensity score matching. For missing data, we studied multiple imputation. All of these methods are remarkably useful in educational research.
Regarding the empirical studies, we present two studies related with the modeling of academic environments in Chilean students beliefs and achievement. The first study investigates the influence of Life and Non-life sciences courses in secondary students' epistemic and self-efficacy beliefs related to sciences. It is an intervention and it has a moderate amount of selection bias. Also the outcome variables are students beliefs. The second study concerns school effects on academic trajectories of Chilean students. This study allows us to better understand the Chilean educational system. We modeled students trajectories in standardized test scores and there is a large amount of selection bias. Both studies are different examples of the area of educational effectiveness research and their differences illustrate different statistical issues that have to be considered.

We presented in Part I the discussion about statistical methods. First, we analyzed how to model simultaneously selection bias and multilevel models. Then, we discussed how to apply multiple imputation for multilevel data. Finally, we proposed strategies to combine the multilevel models, selection bias methods and multiple imputation. Regarding multilevel models and selection bias, we articulated the discussion according to two cases: intra-cluster treatment assignment and inter-cluster treatment assignment. For each case, we analyzed how to linearly adjust for counfounding variables in multilevel models. In particular, we described how the exogeneity hypothesis is extended when there are several random effects. Also we described strategies to solve endogeneity problems using means of first-level variables and fixed effects. Then, we described the strategies for doing propensity score matching in the case of intra-cluster and inter-cluster treatment assignment, considering that the final outcome model was a multilevel model. For the intra-cluster case, we described the strategies defined by Thoemmes and West (2011). Regarding the inter-cluster case we classified and discussed four strategies for matching. This classification is useful to design strategies to control for selection bias in observational studies with clustered data. Regarding multiple imputation and multilevel modeling, we described latest developments and open questions. Finally, regarding combining multilevel models with methods for selection bias an multiple imputation we proposed three strategies.

1. Strategy 1: The first step in this strategy is doing multiple imputation. The second step is using multilevel models and linearly adjusting for selection bias in each imputed data set. Finally, pool the treatment effect estimates.
2. Strategy 2: The first step in this strategy is doing multiple imputation. The second step is, for each imputed data set, estimate the propensity score and the treatment effect with multilevel models for the outcome. Finally, pool the treatment effect estimates.
3. Strategy 3: The first step in this strategy is doing multiple imputation. The second step is, for each imputed data set, estimate the propensity score and pool the propensity scores. This will lead to only one estimated propensity score for unit. The third step is matching the pooled propensity scores, which will produce only one matched set of units. Select the matched units in each imputed data set and estimate the treatment effects with multilevel models. Finally, pool the treatment effect estimates.

We discussed the strengths and weakness of the strategies considering latest simulation studies and practical considerations. In sum, Strategy 1 is relevant because in some cases it is too complex to do propensity score matching, for example for sample size reasons or because there are several treatments. Strategy 2 has prove to lead to more bias reduction than Strategy 3 in simulation studies. But, Strategy 3 has several practical advantages because each imputed data set has the same units and sample sizes after matching. In particular, the balance produced by the propensity score model can be evaluated in the pooled standardized means.

In Part II we studied the effect of summer sciences courses on students beliefs. We compared the effect of Life and Non-life sciences courses with a control group in a post and follow-up measure. This was a quasi experiment with a small sample size of the control group. Also there were several treatments because we compared Life and Non-life courses and courses with and without laboratory work. These aspects made very complex to use propensity score matching. Thus, we adopted Strategy 1 for the statistical modeling.

We distinguish Life and Non-life courses because this classification can allows us to understand some processes in science education. Indeed, we found different results for Life and Non-life courses. Regarding the effects on self-efficacy beliefs, we found a splashdown effect on science self-efficacy for Life courses but not for Non-life courses.

This can be an effect of the teaching cultures of Life disciplines. Nevertheless, it is also possible that if we use a measure of self-efficacy specifically related with Non-life courses we could found effects. For Laboratory work, we found effects on self-efficacy at post but not for follow-up, this is in line with (Itzek-Greulich et al., 2017). With respect to epistemic beliefs, in Justification we found a significant effect of life courses at the follow up measure. This dimension was related with the role of experiments. These results are coherent considering the role that have experiments in life sciences. For development no significant effects were found. The courses are not designed to change epistemic beliefs, and probably these dimensions needs explicit discussion and analysis. For Certainty and Source, Life and Non-life disciplines produced less sophisticated beliefs. Regarding peer effects, we found that the intra-class correlations increased at the post measure. No effect was found from aggregated variables. In conclusion, this study showed positive effects in science self-efficacy beliefs and justification from the life courses and courses with laboratory work. These results are relevant for science education for the influence of these beliefs on students choices, achievement and learning strategies.

In Part III, we studied school effects on students trajectories in reading and mathematics scores from 4th grade to 10th grade. We also measured the effect of the public system on students trajectories. This study is novel considering the large time span between the measures (5 years) and that we used a census-data base. We used Strategies 1 and 3 for the statistical modeling. This allowed us to compare the results using propensity score matching, linear adjustment and both methods simultaneously. We used multilevel growth models with a cross-classified structure to account for school effects at 8th and 10th grade. Regarding the schools variances in students intercepts, we found small but sizable variances at the primary school level and large variances at the secondary school level. The school variances for the random slopes were significant and more evenly distributed between primary and secondary schools. This revealed long term effects of primary schools. Also, we described students trajectories for different groups, illustrating inequalities of the system. On the question of the effect of public and voucher schools, we found large negative and significant effects of public education on the intercepts when we controlled only for student background variables. Nevertheless, when we included school aggregated variables regarding socioeconomic conditions and selection practices, the effects became non significant.

Regarding students slopes, the effects of public education were significant and negative, but not relevant considering the effect sizes. Linear adjustment and propensity score matching individually and combined produced similar conclusions. However the propensity score analysis revealed a large number of students in the public and in the voucher system which are not comparable. The students with most disadvantaged conditions are in public schools.

We found no relevant difference between trajectories of students in the public and the voucher system when we controlled for socioeconomic school composition and selection practices. The conclusions coincide with several cross-sectional studies (Anand et al., 2009; Bellei, 2008; Lara et al., 2011; Zubizarreta & Keele, 2016) and with the longitudinal study done by Ortega Ferrand (2015).

Research perspectives

Regarding statistical research perspectives, it is necessary to clarify when each different inter-cluster option should be used. In observational clustered studies, it can be unclear how to deal with the cluster variables. Considering that linear adjustment combined with matching is a robust strategy at the individual level, we could argue that the same strategy is optimal at the cluster level. Nevertheless, other aspects have to be considered. First, sample sizes at the cluster level are significantly smaller than at the individual level. Also, the treatment is delivered at the cluster level. This implies that discarding clusters could produce loss of treatment heterogeneity that may be the focus of the research. The selection of strategies for controlling for selection bias in inter-cluster studies should be studied in deep.
Regarding the combination of multilevel models, propensity score matching and multiple imputation there are several issues to explore. A first point is to determine if Strategy 3 produces equal bias reduction than Strategy 2 when matching is combined with linear adjustment. This could allows us to benefit of the practical advantages of Strategy 3 and get the same level of bias reduction. Another aspect to be researched it is determining the characteristics of the proposed strategies when, instead of propensity score matching, stratification of the propensity score is used.

With respect to the study of the effect of Life and Non-life courses on epistemic and self-efficacy beliefs, the main suggestions are that the experiences produced by Life and Non-life courses should be described and other epistemic and self-efficacy dimensions should be explored. It is important to understand which teaching approaches and specific epistemic differences can be identified between Life and Non-life courses. Also, Laboratory work showed to be effective for producing more availing self-efficacy and justification beliefs. It is important to note that this is a very authentic laboratory work done in a University. Research regarding laboratory work in the school place should be developed in the Chilean context. Finally, peer effects on epistemic beliefs should be explored in a school or university context where the students are exposed to their peers more time. Intra-class correlations and compositional effects are relevant aspects to consider.

Regarding the study about school effects on students trajectories and the comparison of public and voucher schools, we found that the differences between public and voucher schools were explained by school composition and school' selection practices. Also, we found that the students with most disadvantaged socioeconomic conditions are only educated by public schools. This implies that public schools and the segregation problem should receive special attention, in order to stop the severe level of segregation that is producing inequalities. Comparisons between the public system and the voucher system have been done by several studies, it would be valuable to study inside each system which schools are more efficient.

References

Agencia de Calidad de la Educación. (n.d.). *Metodología de Construcción de Grupos Socioeconómicos. Pruebas SIMCE 2012.* (Tech. Rep.). Santiago, Chile: Agencia de Calidad de la Educación, Gobierno de Chile. Retrieved 2017-07-07, from `http://www.agenciaeducacion.cl/wp-content/uploads/2013/02/Metodologia-de-Construccion-de-Grupos-Socioeconomicos-SIMCE-2012.pdf`

Agencia de Calidad de la Educación. (2015). *Informe Técnico SIMCE 2013* (Tech. Rep.). Santiago, Chile: Agencia de Calidad de la Educación, Gobierno de Chile. Retrieved 2017-07-07, from `http://archivos.agenciaeducacion.cl/documentos-web/InformeTecnicoSimce_2013.pdf`

Agencia de Calidad de la Educación. (2017, June). *Evaluaciones: Indicadores de Desarrollo Personal y Social.* Retrieved 2017-07-07, from `http://www.agenciaeducacion.cl/evaluaciones/indicadores-desarrollo-personal-social/`

Allison, P. D. (2002). *Missing Data.* Thousand Oaks, CA, USA: Sage.

Anand, P., Mizala, A., & Repetto, A. (2009). Using school scholarships to estimate the effect of private education on the academic achievement of low-income students in Chile. *Economics of Education Review*, *28*(3), 370–381. doi: 10.1016/j.econedurev.2008.03.005

Anumendem, N. D., De Fraine, B., Onghena, P., & Van Damme, J. (2013). The impact of coding time on the estimation of school effects. *Quality & Quantity*, *47*(2), 1021–1040. doi: 10.1007/s11135-011-9581-3

Arpino, B. (2010). *Causal Inference for Observational Studies Extended to a Multilevel Setting: The Impact of Fertility on Poverty in Vietnam* (Unpublished doctoral dissertation). Università degli Studi di Firenze, Italy.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780. doi: 10.1016/j.csda.2010.11.008

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150–161. doi: 10.1002/pst.433

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R.* Retrieved 2016-09-30, from `http://lme4.0.r-forge.r-project.org/lMMwR/lrgprt.pdf`

Bellei, C. (2008). The Public-Private School Controversy in Chile. In R. Chakrabarti & P. Peterson (Eds.), *School Choice International: Exploring Public-Private Partnerships* (pp. 165–192). Cambridge, MA, USA: MIT Press.

Bellei, C., & Cabalin, C. (2013). Chilean student movements: Sustained struggle to transform a market-oriented educational system. *Current Issues in Comparative Education*, *15*(2), 108–123. Retrieved 2015-08-07, from `http://files.eric.ed.gov/fulltext/EJ1016193.pdf`

Bellei, C., Cabalin, C., & Orellana, V. (2014). The 2011 Chilean student movement against neoliberal educational policies. *Studies in Higher Education*, *39*(3), 426–440. doi: 10.1080/03075079.2014.896179

Bellei, C., & Vanni, X. (2015). The Evolution of Educational Policy in Chile, 1980-2014. In S. Schwartzman (Ed.), *Education in South America: Education Around the World* (pp. 179–200). London, United Kingdom: Bloomsbury Academic.

Bellio, R., & Gori, E. (2003). Impact evaluation of job training programmes: Selection bias in multilevel models. *Journal of Applied Statistics*, *30*(8), 893–907. doi: 10.1080/0266476032000075976

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, *57*(3), 195–203. doi: 10.1037/h0034701

Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task value, and achievement goals. *Journal of Educational Psychology*, *93*(1), 23–34. doi: 10.1037/0022-0663.93.1.23

Bressoux, P. (2007). L'apport des modèles multiniveaux à la recherche en éducation. *Éducation et didactique*, *1*(2), 73–88. doi: 10.4000/educationdidactique.168

Bressoux, P. (2008). Effet-établissement. In A. Van Zanten (Ed.), *Dictionnaire de l'éducation* (pp. 212–216). Paris, France: Presses universitaires de France.

Bressoux, P. (2010). *Modélisation statistique appliquée aux sciences sociales.* Brussels, Belgium: De Boeck.

Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching*, *43*(5), 485–499. doi: 10.1002/tea.20131

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, *1*(2), 103–124. doi: 10.1191/147108201128113

Buehl, M. M. (2003). *At the crossroads of epistemology and motivation: Modeling the relations between students' domain-specific epistemological beliefs, achievement motivation, and task performance* (Unpublished doctoral dissertation). University of Maryland, USA.

Buehl, M. M. (2008). Assessing the multidimensionality of students' epistemic beliefs across diverse cultures. In M. S. Khine (Ed.), *Knowing, knowledge and beliefs* (pp. 65–112). Springer. doi: 10.1007/978-1-4020-6596-5_4

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*(3). doi: 10.18637/jss.v045.i03

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications.* New York, NY, USA: Cambridge University Press.

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application.* Chichester, United Kingdom: Wiley.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, Context, and Endogeneity in School and Teacher Comparisons. *Journal of Educational and Behavioral Statistics*, *39*(5), 333–367. doi: 10.3102/1076998614547576

Chai, C. S., Deng, F., Wong, B. K. S., & Qian, Y. (2010). South China education majors' epistemological beliefs and their conceptions of the nature of science. *The Asia-Pacific Education Researcher*, *19*, 111–125. doi: 10.3860/taper.v19i1.1512

Chen, J. A., & Pajares, F. (2010). Implicit theories of ability of Grade 6 science students: Relation to epistemological beliefs and academic motivation and achievement in science. *Contemporary Educational Psychology*, *35*(1), 75–87. doi: 10.1016/j.cedpsych.2009.10.003

Cheslock, J. J., & Rios-Aguilar, C. (2011). Multilevel Analysis in Higher Education Research: A Multidisciplinary Approach. In J. C. Smart & M. B. Paulsen (Eds.), *Higher Education: Handbook of Theory and Research* (Vol. 26, pp. 85–123). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-0702-3_3

Chumacero, R. A., Gómez, D., & Paredes, R. D. (2011). I would walk 500 miles (if it paid): Vouchers and school choice in Chile. *Economics of Education Review*, *30*(5), 1103–1114. doi: 10.1016/j.econedurev.2011.05.015

Clark, M. H. (2015). Propensity Score Adjustment Methods. In W. Pan & H. Bai (Eds.), *Propensity Score Analysis. Fundamentals and developments* (pp. 115–140). New York, NY, USA: Guilford Press.

Cofré, H., Camacho, J., Galaz, A., Jiménez, J., Santibáñez, D., & Vergara, C. (2010). La educación científica en Chile: debilidades de la enseñanza y futuros desafíos de la educación de profesores de ciencia. *Estudios pedagógicos (Valdivia)*, *36*(2), 279–293. doi: /10.4067/S0718-07052010000200016

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, *29*(2), 186–204. doi: 10.1016/j.cedpsych.2004.01.004

Contreras, D., Sepúlveda, P., & Bustos, S. (2010). When Schools Are the Ones that Choose: The Effects of Screening in Chile. *Social Science Quarterly*, *91*(5), 1349–1368. doi: 10.1111/j.1540-6237.2010.00735.x

Corvalán, J., & García-Huidobro, J. E. (2016). Educación y Mercado: El caso Chileno. In J. Corvalán, A. Carrasco, & J. E. García-Huidobro (Eds.), *Mercado Escolar y Oportunidad Educacional. Libertad, Diversidad y Desigualdad* (pp. 17–55). Santiago, Chile: Ediciones UC.

Deng, F., Chen, D.-T., Tsai, C.-C., & Chai, C. S. (2011). Students' Views of the Nature of Science: A Critical Review of Research. *Science Education*, *95*(6), 961–999. doi: 10.1002/sce.20460

Diez, R. (2002). A glossary for multilevel analysis. *Journal of epidemiology and community health*, *56*(8), 588–594. doi: 10.1136/jech.56.8.588

Drechsler, J. (2015). Multiple imputation of multilevel missing data: Rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, *40*(1), 69–95. doi: 10.3102/1076998614563393

Elacqua, G., Schneider, M., & Buckley, J. (2006). School choice in Chile: Is it class or the classroom? *Journal of Policy Analysis and Management*, *25*(3), 577–601. doi: 10.1002/pam.20192

Ellis, P. D. (2010). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results.* New York, NY, USA: Cambridge University Press.

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222–240. doi: 10.1037/met0000063

Everson, K. C. (2017). Value-Added Modeling and Educational Accountability Are We Answering the Real Questions? *Review of Educational Research*, *87*(1), 35-70. doi: 10.3102/0034654316637199

Feller, A., & Gelman, A. (2015). Hierarchical models for causal effects. In R. Scott & S. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource.* (pp. 1–16). doi: 10.1002/9781118900772.etrds0160

Feucht, F. C. (2010). Epistemic climate in elementary classrooms. In L. D. Bendixen & F. C. Feucht (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice* (pp. 55–93). New York, NY, USA: Cambridge University Press. doi: 10.1017/CBO9780511691904.003

Frangakis, C., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, *82*(2), 365–379. doi: 10.1093/biomet/86.2.365

García, M. B., & Mateos, M. (2013). Las cuestiones de dominio intersujeto e intrasujeto en el contenido de las concepciones epistemológicas en docentes universitarios. *Avances en Psicología Latinoamericana*, *31*(3), 586–619. Retrieved 2016-07-11, from `http://www.scielo.org.co/pdf/apl/v31n3/v31n3a11.pdf`

Gelber, D., Treviño, E., Inostroza, P., & others. (2016). *Inequidad de género en los logros de aprendizaje en educación primaria. ¿ Qué nos puede decir TERCE?: resumen ejecutivo* (Tech. Rep.). UNESCO. Retrieved 2017-06-08, from `http://repositorio.minedu.gob.pe/handle/123456789/4355`

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). Chichester, United Kingdom: Wiley.

Goldstein, H., & Sammons, P. (1997). The Influence of Secondary and Junior Schools on Sixteen Year Examination Performance: A Cross-classified Multilevel Analysis. *School Effectiveness and School Improvement*, *8*(2), 219–230. doi: 10.1080/0924345970080203

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of multilevel missing data: An introduction to the R package pan. *SAGE Open*, *6*(4), 1–17. doi: 10.1177/2158244016668220

Guldemond, H., & Bosker, R. J. (2009). School effects on students' progress: a dynamic perspective. *School Effectiveness and School Improvement*, *20*(2), 255–268. doi: 10.1080/09243450902883938

Hanchane, S., & Mostafa, T. (2012). Solving endogeneity problems in multilevel estimation: an example using education production functions. *Journal of Applied Statistics*, *39*(5), 1101–1114. doi: 10.1080/02664763.2011.638705

Hansen, B. B., Rosenbaum, P. R., & Small, D. S. (2014). Clustered Treatment

Assignments and Sensitivity to Unmeasured Biases in Observational Studies. *Journal of the American Statistical Association*, *109*(505), 133–144. doi: 10.1080/01621459.2013.863157

Hill, J. (2004). *Reducing bias in treatment effect estimation in observational studies suffering from missing data. Working paper 04-01.* (Tech. Rep.). Institute for Social and Economic Research and Policy (ISERP), Columbia University,. doi: 10.7916/D8B85G11

Hill, J. (2013). Multilevel Models and Causal Inference. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The SAGE handbook of multilevel modeling* (pp. 201–219). Los Angeles, CA, USA: SAGE.

Hofer, B. K. (2000). Dimensionality and Disciplinary Differences in Personal Epistemology. *Contemporary Educational Psychology*, *25*(4), 378–405. doi: 10.1006/ceps.1999.1026

Hofer, B. K., & Pintrich, P. R. (1997). The Development of Epistemological Theories: Beliefs About Knowledge and Knowing and Their Relation to Learning. *Review of Educational Research*, *67*(1), 88–140. doi: 10.3102/00346543067001088

Hox, J. J. (2010). *Multilevel analysis: techniques and applications* (2th ed.). New York, NY, USA: Routledge, Taylor & Francis.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social and Biomedical Sciences. An Introduction.* New York, NY, USA: Cambridge University Press.

Itzek-Greulich, H., Flunger, B., Vollmer, C., Nagengast, B., Rehm, M., & Trautwein, U. (2017). Effectiveness of lab-work learning environments in and out of school: A cluster randomized study. *Contemporary Educational Psychology*, 98-115. doi: 10.1016/j.cedpsych.2016.09.005

Kim, J.-S., & Frees, E. W. (2006). Omitted Variables in Multilevel Models. *Psychometrika*, *71*(4), 659–690. doi: 10.1007/s11336-005-1283-0

Kim, J.-S., & Swoboda, C. M. (2010). Handling Omitted Variable Bias in Multilevel Models: Model Specification Tests and Robust Estimation. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis.* New York, NY, USA: Routledge.

King, G., & Nielsen, R. (2016). *Why propensity scores should not be used for matching.* Retrieved 2017-06-18, from `https://gking.harvard.edu/files/gking/files/psnot.pdf` (working paper)

Kizilgunes, B., Tekkaya, C., & Sungur, S. (2009). Modeling the Relations Among Students' Epistemological Beliefs, Motivation, Learning Approach, and Achievement. *The Journal of Educational Research*, *102*(4), 243–256. doi: 10.3200/JOER.102.4.243-256

Lang, K. M., & Little, T. D. (2016). Principled Missing Data Treatments. *Prevention Science*. doi: 10.1007/s11121-016-0644-5

Lara, B., Mizala, A., & Repetto, A. (2011). The Effectiveness of Private Voucher Education: Evidence From Structural School Switches. *Educational Evaluation and Policy Analysis*, *33*(2), 119–137. doi: 10.3102/0162373711402990

Larose, S., Ratelle, C. F., Guay, F., Senécal, C., & Harvey, M. (2006). Trajectories of science self-efficacy beliefs during the college transition and academic and vocational adjustment in science and technology programs. *Educational Research and Evaluation*, *12*(4), 373–393. doi: 10.1080/13803610600765836

Lederman, N. G. (2007). Nature of science: Past, present, and future. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 831–879). Routledge.

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. doi: 10.1080/00273171.2014.991018

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265. doi: 10.1126/science.1261375

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., ... Williamson, E. J. (2017). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research*. doi: 10.1177/0962280217713032

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in medicine*, *32*(19), 3373–3387. doi: 10.1002/sim.5786

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2th ed.). Hoboken, NJ, USA: Wiley.

Liu, S.-Y., & Tsai, C.-C. (2008). Differences in the Scientific Epistemological Views of Undergraduate Students. *International Journal of Science Education*, *30*(8), 1055–1073. doi: 10.1080/09500690701338901

Longford, N. T. (2012). A Revision of School Effectiveness Analysis. *Journal of Educational and Behavioral Statistics*, *37*(1), 157–179. doi: 10.3102/1076998610396898

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies. *Psychological Methods*(1), 141–165. doi: 10.1037/met0000096

Manzi, J., San Martín, E., & Van Bellegem, S. (2014). School system evaluation by value added analysis under endogeneity. *Psychometrika*, *79*(1), 130–153. doi: 10.1007/s11336-013-9338-0

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of educational psychology*, *79*(3), 280–295. doi: 10.1037/0022-0663.79.3.280

Mason, L., Boscolo, P., Tornatora, M. C., & Ronconi, L. (2013). Besides knowledge: a cross-sectional study on the relations between epistemic beliefs, achievement goals, self-beliefs, and achievement in science. *Instructional Science*, *41*(1), 49–79. doi: 10.1007/s11251-012-9210-0

Mattei, A. (2009). Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Statistical Methods and Applications*, *18*(2), 257–273. doi: 10.1007/s10260-007-0086-0

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, *32*(19), 3388–3414. doi: 10.1002/sim.5753

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, *9*(4), 403–425. doi: 10.1037/1082-989X.9.4.403

McEwan, P. J. (2003). Peer effects on student achievement: evidence from Chile. *Economics of Education Review*, *22*(2), 131–141. doi: 10.1016/S0272-7757(02)00005-5

Meckes, L., & Carrasco, R. (2010). Two decades of SIMCE: an overview of the National Assessment System in Chile. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 233–248. doi: 10.1080/09695941003696214

Miller, M. C., Montplaisir, L. M., Offerdahl, E. G., Cheng, F.-C., & Ketterling, G. L. (2010). Comparison of views of the nature of science between natural science and nonscience majors. *CBE-Life Sciences Education*, *9*(1), 45–54. doi: 10.1187/cbe.09-05-0029

Ministerio de Educación, R. d. C. (2014). *Estadísticas de la educación 2013.* Retrieved 2017-05-10, from `http://centroestudios.mineduc.cl/index.php?t=96&i=2&cc=2036&tm=2`

Mitra, R., & Reiter, J. P. (2012). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research*, *25*(1), 188–204. doi: 10.1177/0962280212445945

Muis, K. R., Bendixen, L. D., & Haerle, F. C. (2006). Domain-Generality and Domain-Specificity in Personal Epistemology Research: Philosophical and Empirical Reflections in the Development of a Theoretical Framework. *Educational Psychology Review*, *18*(1), 3–54. doi: 10.1007/s10648-006-9003-6

Muis, K. R., & Duffy, M. C. (2013). Epistemic climate and epistemic change: Instruction designed to change students' beliefs and learning strategies and improve achievement. *Journal of Educational Psychology*, *105*(1), 213–225. doi: 10.1037/a0029690

Muñoz-Chereau, B., & Thomas, S. (2016). Educational effectiveness in Chilean secondary education: comparing different 'value added' approaches to evaluate schools. *Assessment in Education: Principles, Policy & Practice*, *23*(1), 26–52. doi: 10.1080/0969594X.2015.1066307

OECD. (2012). *Public and Private Schools.* OECD Publishing. doi: 10.1787/9789264175006-en

Ortega Ferrand, L. C. (2015). *Educational Effectiveness and Inequalities in Chile: A Multilevel Accelerated Longitudinal Study of Primary School Children's Achievement Trajectories* (Unpublished doctoral dissertation). University of Oxford, United Kingdom.

Paredes, R. D., & Pinto, J. I. (2009). ¿El fin de la educación pública en Chile? *Estudios de economía*, *36*(1), 47–66. doi: 10.4067/S0718-52862009000100003

Pattanayak, C. W. (2015). Evaluating Covariate Balance. In W. Pan & H. Bai (Eds.), *Propensity Score Analysis. Fundamentals and developments* (pp. 89–112). New York, NY, USA: Guilford Press.

Penning de Vries, B., & Groenwold, R. (2016). Comments on propensity score matching following multiple imputation. *Statistical Methods in Medical Research*, *25*(6), 3066–3068. doi: 10.1177/0962280216674296

Perry Jr, W. G. (1999). *Forms of Intellectual and Ethical Development in the College Years: A Scheme. Jossey-Bass Higher and Adult Education Series.* San Francisco, CA, USA: Jossey-Bass.

Peugh, J. L., & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*, *74*(4), 525–556. doi: 10.3102/00346543074004525

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2017). nlme: Linear and nonlinear mixed effects models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=nlme` (R package version 3.1-131)

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raczynski, D., Muñoz, G., Weinstein, J., & Pascual, J. (2016). Subvención escolar preferencial (SEP) en Chile: un intento por equilibrar la macro y micro política escolar. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, *11*(2), 164–193. Retrieved 2016-05-20, from `https://revistas.uam.es/index.php/reice/article/view/2902`

Raudenbush, S. W., & Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307-335. doi: 10.2307/1165304

Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from `https://CRAN.R-project.org/package=psych` (R package version 1.6.12)

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): a state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 197–230. doi: 10.1080/09243453.2014.885450

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi: 10.2307/2335942

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688–701. doi: 10.1037/h0037350

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. doi: 10.2307/2335739

Rubin, D. B. (1987). *Multiple Imputation for nonresponse in surveys.* New York, NY: Wiley.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. doi: 10.1002/sim.2739

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology*, *82*(3), 498–504. doi: 10.1037/0022-0663.82.3.498

Schunk, D. H., & Pajares, F. (2009). Self-efficacy theory. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35–53). New York, NY, USA: Routledge.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*(7). doi: 10.18637/jss.v042.i07

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press.

Stake, J. E., & Mares, K. R. (2005). Evaluating the impact of science-enrichment programs on adolescents' science motivation and confidence: The splashdown effect. *Journal of Research in Science Teaching*, *42*(4), 359–375. doi: 10.1002/tea.20052

Stuart, E. A. (2007). Estimating Causal Effects Using School-Level Data Sets. *Educational Researcher*, *36*(4), 187–198. doi: 10.3102/0013189X07303396

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, *25*(1), 1–21. doi: 10.1214/09-STS313

Sun, S., & Pan, W. (2014). A methodological review of statistical methods for handling multilevel non-nested longitudinal data in educational research. *International Journal of Research & Method in Education*, *37*(3), 285–308. doi: 10.1080/1743727X.2014.885012

Thoemmes, F. J., & Kim, E. S. (2011). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*, *46*(1), 90–118. doi: 10.1080/00273171.2011.540475

Thoemmes, F. J., & West, S. G. (2011). The Use of Propensity Scores for Non-randomized Designs With Clustered Data. *Multivariate Behavioral Research*, *46*(3), 514–543. doi: 10.1080/00273171.2011.569395

Thoermer, C., & Sodian, B. (2002). Science undergraduates and graduates epistemologies of science: The notion of interpretive frameworks. *New ideas in Psychology*, *20*(2), 263–283. doi: 10.1016/S0732-118X(02)00009-0

Trautwein, U., & Lüdtke, O. (2008). Global Certainty Beliefs and College Major: How Strong Are Socialization Effects? In M. S. Khine (Ed.), *Knowing, Knowledge and Beliefs* (pp. 241–255). Springer.

Troncoso, P., Pampaka, M., & Olsen, W. (2016). Beyond traditional school value-added models: a multilevel analysis of complex school effects in Chile. *School Effectiveness and School Improvement*, *27*(3), 293-314. doi: 10.1080/09243453.2015.1084010

Tsai, C.-C., Jessie Ho, H. N., Liang, J.-C., & Lin, H.-M. (2011). Scientific epistemic beliefs, conceptions of learning science and self-efficacy of learning science among high school students. *Learning and Instruction*(6), 757–769. doi: 10.1016/j.learninstruc.2011.05.002

Tuohilampi, L., Hannula, M. S., Varas, L., Giaconi, V., Laine, A., Näveri, L., & i Nevado, L. S. (2015). Challenging the western approach to cultural comparisons: young pupils' affective structures regarding mathematics in Finland and Chile. *International Journal of Science and Mathematics Education*, *13*(6), 1625–1648. doi: 10.1007/s10763-014-9562-9

Valenzuela, J. P., Bellei, C., & De los Ríos, D. (2014, March). Socioeconomic school segregation in a market-oriented educational system. The case of Chile. *Journal of Education Policy*, *29*(2), 217–241. doi: 10.1080/02680939.2013.806995

Valenzuela, J. P., Villarroel, G., & Villalobos, C. (2013). Ley de Subvención Escolar Preferencial (SEP): algunos resultados preliminares de su implementación. *Pensamiento Educativo: Revista de Investigación Educacional Latinoamericana*, *50*(2), 113–131. doi: 10.7764/PEL.50.2.2013.17

Van Buuren, S. (2011). Multiple Imputation of Multilevel Data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 173–196). Routledge.

Van Buuren, S. (2012). *Flexible imputation of missing data.* NW, USA: CRC press, Taylor and Francis.

Vanwynsberghe, G., Vanlaar, G., Van Damme, J., & De Fraine, B. (2017). Long-term effects of primary schools on educational positions of students 2 and 4 years after the start of secondary education. *School Effectiveness and School Improvement*, *28*(2), 167–190. doi: 10.1080/09243453.2016.1245667

Von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*(1), 83–117. doi: 10.1111/j.1467-9531.2007.00180.x

Wang, Q. (2015). Propensity Score Matching on Multilevel Data. In W. Pan & H. Bai (Eds.), *Propensity Score Analysis. Fundamentals and Developments.* (pp. 320–247). New York, NY, USA: Guilford Press.

Weinstein, J., & Muñoz, G. (2014). When duties are not enough: principal leadership and public or private school management in Chile. *School Effectiveness and School Improvement*, *25*(4), 651–670. doi: 10.1080/09243453.2013.792850

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2th ed.). Cambridge, MA, USA: The MIT Press.

Zhao, J. H., & Schafer, J. L. (2016). pan: Multiple imputation for multivariate panel or clustered data [Computer software manual]. (R package version 1.4)

Zubizarreta, J. R., & Keele, L. (2016). Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System. *Journal of the American Statistical Association*. doi: 10.1080/01621459.2016.1240683

Appendices

Appendix A

Part II

A.1 Epistemic and Self-efficacy Beliefs questionnaire

The items from the science self-efficacy dimension are presented in Table A.1. The items from the epistemic dimensions are a translation from the questionnaire of Conley et al. (2004). The items' translations are presented in Tables A.2, A.3, A.4, A.5.

Table A.1: Items of Self-efficacy in science

Item English	Item Spanish
I am sure that I can understand the most difficult concepts in sciences	Estoy seguro de que puedo comprender los conceptos más difíciles de ciencias
I can succeed in sciences	Sé que puedo tener éxito en las ciencias
I know I can dominate the abilities that are teached in sciences	Estoy seguro de que puedo dominar las habilidades que se enseñan en ciencias
I am sure that I can learn sciences	Estoy seguro de que puedo aprender ciencias
I think I can do better than now in sciences	Creo que podría hacerlo mejor que hasta ahora en ciencias
I can get good grades in sciences	Puedo sacar buenas notas en ciencias

Table A.2: Items of Certainty: A belief in a right answer

Item English	Item Spanish
All questions in science have one right answer	Todas las preguntas en ciencias tienen una respuesta correcta
The most important part of doing science is coming up with the right answer	Lo más importante de hacer ciencia es encontrar la respuesta correcta
Scientists pretty much know everything about science; there is not much more to know	Los científicos saben prácticamente todo sobre la ciencia, no hay mucho más que conocer
Scientific knowledge is always true	El conocimiento científico siempre es verdadero
Once scientist have a result from an experiment, that is the only answer	Cuando un científico tiene un resultado de un experimento, esa es la única respuesta
Scientist always agree about what is true in science	Los científicos siempre están de acuerdo respecto a que es lo verdadero en ciencias

Table A.3: Items of Justification: Role of experiments and how individuals justify knowledge

Item English	Item Spanish
Ideas about science experiments come from being curious and thinking about how things work	Las ideas sobre los experimentos científicos vienen de ser curioso y de pensar sobre cómo funcionan las cosas
In science there can be more than one way for scientists to test their ideas	En ciencias puede haber más de una manera para que los científicos pongan a prueba sus ideas
One important part of science is doing experiments to come up with new ideas about how things work	Una parte importante de la ciencia es hacer experimentos para obtener nuevas ideas sobre cómo funcionan las cosas
It is good to try experiments more than once to make sure of your findings	Es bueno realizar los experimentos más de una vez para estar seguro de los resultados
Good ideas in science can come from anybody, not just from scientists	Las buenas ideas en ciencias pueden venir de cualquier persona, no solo de los científicos
A good way to know if something is true is to do an experiment	Una buena manera de saber si algo es cierto es hacer un experimento
Good answers are based on evidence from many experiments	Las buenas respuestas estÃ¡n basadas en evidencia de muchos experimentos
Ideas in science can come from your own questions and experiments	Las ideas en ciencias pueden venir de tus propias preguntas y experimentos
It is good to have an idea before you starts an experiment	Es bueno tener una idea antes de empezar un experimento

Table A.4: Items of Source: Knowledge residing in external authorities

Item English	Item Spanish
Everybody has to believe what scientists say	Todos tienen que creer lo que dicen los científicos
In science, you have to believe what science books say about stuff	En ciencias, tienes que creer lo que dicen los libros de ciencias sobre las cosas
Whatever the teacher says in science class is true	Todo lo que dice el profesor en la clase de ciencias es verdad
If you read something in a science book, you can be sure it's true	Si lees algo en un libro de ciencia, puedes estar seguro de que es verdad
Only scientists know for sure what is true in science	Sólo los científicos saben realmente que es verdadero en ciencias

Table A.5: Items of Development: beliefs about science as an evolving and changing subject

Item English	Item Spanish
Some ideas in science today are different than what scientists used to think	Hoy en día algunas ideas en ciencias son distintas a las que los científicos solían tener
The ideas in science books sometimes change	Las ideas en los libros de ciencias a veces cambian
There are some questions that even scientists cannot answer	Hay algunas preguntas que ni si quiera los científicos pueden responder
Ideas in science sometimes change	Las ideas en ciencias algunas veces cambian
New discoveries can change what scientist think is true	Los nuevos descubrimientos pueden cambiar lo que los científicos piensan que es verdad
Sometimes scientists change their minds about what is true in science	A veces los científicos cambian de opinión respecto a lo que es verdadero en ciencias

A.2 Multiple imputation procedure

In this section we present the multiple imputation procedure for the study of part II of this thesis, regarding the effect of academic climates on students' beliefs. There was an important percentage of missing data in the outcome variables (Table 7.4). To develop a missing data procedure to account for it, first we tried to determine which was the missing data mechanism. In order to do that, how there were variables with 100% of response, we compared students with missing data with students with complete data in these variables. Statistical tests showed that the groups differed, for example in the average of school grades. This is evidence suggesting that the data was not missing completely at random. Therefore, it was not appropriate to use list-wise deletion.
In order to deal with missing data, we decided to use multiple imputation, which is a recommended method to deal with data missing at random (Lang & Little, 2016; Peugh & Enders, 2004). We cannot test if data was missing at random, so this is a hypothesis that we will assume. In the missing data pattern, it is important to note that the outcome variables had a high amount of missing values. Doing a regular multiple imputation procedure implied impute the outcomes values for 221 individuals in the post-measure and for 680 individuals in the follow-up measure. This strategy implied a very important loss of power for added noise. To address this problem, we used the approach of multiple imputation and then deletion proposed by Von Hippel (2007). This approach consists in two steps:

1. Impute all the data
2. Delete variables were the outcome missing

This technique is suitable because imputed values in the outcome variable of a linear regression do not contribute to the maximum likelihood (Von Hippel, 2007). As a precision, in this study there were person non-response, when a person did not answer a complete questionnaire and item non-response, when a person answered the questionnaire but leave some items without answer. The percentages of the sample that had missing data for item-non response were small. The main problem was the person-non response, Therefore, we considered the individuals with a missing item as a missing individual, that means we modeled their data as person non-response.

The multiple imputation model was implemented with package mice in the software R (Buuren & Groothuis-Oudshoorn, 2011). This package uses chained equations to model the missing values. In addition, it uses functions from the pan package to deal with multilevel imputation (Grund, Lüdtke, & Robitzsch, 2016).
We used the data of the 994 students, which correspond to the students that answered at least one time the beliefs questionnaire. We used all the variables in Table 7.4 to do the multiple imputation model. Also, we used record variables. These variables were only used in the imputation model and not in the final multilevel models because they were not significant in the later ones.

We specified the imputation method for each variable with missing data according with its type. For post and follow-up beliefs, the imputation method considered the multilevel structure. We added interactions between all the variables with the variable that assigned the type of course discipline (Life, Non-life and Control). The reason for this is that the missing data processes were different for each treatment group. This is very important in the post measure, where questionnaires for Life and Non-life courses were measured in paper-and-pencil format and in the control group the measurement was done online.

We did 100 imputations. For each imputation we did 100 iterations of the algorithm. The procedure took several hours to be finished.
We checked convergence according to Buuren and Groothuis-Oudshoorn (2011) guidelines. All the variables showed good mixing for the Markov Chain procedure, except for follow-up beliefs. At the end of the procedure, we deleted the imputed values. We considered that, in spite of the bad mixing of follow-up variables, the overall imputation procedure was acceptable.
For the analysis where the post-beliefs where the outcomes, we deleted in each of the 100 complete data sets the observations with imputed post beliefs. We did the same for the analysis where the follow-up beliefs where the outcomes.

Appendix B

Part III

B.1 Multiple imputation procedure

In this section we present the multiple imputation procedure for the study of part III of this thesis, regarding school effects on students' trajectories. The percentage of missing data for each variable was moderate (Table 11.2), but using complete case analysis implied an excessive loss of information.

In order to deal with missing data, we decided to use multiple imputation, which is a recommended method to deal with data missing at random (Lang & Little, 2016; Peugh & Enders, 2004). We cannot test if the data was missing at random, so this is a hypothesis that we will assume. The multiple imputation model was implemented with package mice in the software R (Buuren & Groothuis-Oudshoorn, 2011). This package uses chained equations to model the missing values. In addition it uses functions from the pan package to deal with multilevel imputation (Grund et al., 2016).

We used the data base in wide-format in order to have two levels: students and schools. We specified the imputation method for each variable with missing data according with its type. There were three grouping variables, the school at 4th grade, the school at 8th grade and the school at 10th grade. We estimated intra-class correlations for each variable using as grouping variable the school at each grade. Most of the variables had higher intra-class correlations for several groupings. However, it was not feasible with the available softare to define cross-classified multiple imputation models. This is why we decided to model the multilevel structure according to the level defined by the school at the year were the variable was measured.

We generated 20 imputations and for each imputation we run 100 iterations of the algorithm. It is important to note that for estimating 1 imputation with 100 iterations, the algorithm used between 12 and 14 hours. In order to count with the necessary computer capabilities, we used the computer facilities of the National Laboratory NLHPC (ECM-02). These amount of time called for not complicate in excess the models in order to do not increase the estimation time. This is also a reason to explain why we did only 20 imputations. In addition, Goldstein (2011, pp. 306) recommends doing at least 10 imputation for multilevel models.

For each variable with missing data, we tried to define imputation models with the relevant predictors but that were at the same time simple. For longitudinal studies, Van Buuren (2012, pp. 226) proposes that to reduce the number of predictors an option is only use the ones corresponding to the same measurement time. We used this strategy in order to have simple models. The principal predictor variables for each variable with missing data are detailed in Table B.1. In this table, achievement variables refers to normalized scores in reading and mathematics in 4th, 8th and 10th grade. Socioeconomic variables refer to mother education, father education, number of books, home income, computer and internet. Each variable was measured in 4th, 8th and 10th grade. Pre-school variables refers to kindergarten, pre-kindergarten, ECE (0-2) and ECE (2-4). The school level indexes for 4th, 8th and 10th grade were used. The initials in the table mean:

- AV: All the variables, excepting the predicted variable. For example the normalized mathematics score in 8th grade is an achievement variable. The table indicates that it was predicted by all the achievement variables (excepting normalized mathematics score in 8th) all the pre-school variables, gender and trajdep.
- AV-CY: All the variables for the corresponding year. Following the example of the normalized mathematics score in 8th grade, the table implies that only the socioeconomic variables in 8th grade were used to predict it. The same for parents expectations and the school level indexes.
- CY: Corresponding year. It indicates that the multilevel imputation modeling use as clustering variable the school at the corresponding year. For the normalized mathematics score in 8th grade the grouping variable was the school at 8th grade.

We checked convergence according to Buuren and Groothuis-Oudshoorn (2011) guidelines. All the variables showed good mixing for the Markov Chain procedure.

Table B.1: Predictor variables used in the multiple imputation models.

	Predictor variables									
Predicted Variable	1.	2.	3.	4.	5.	6.	7.	8.	Gender and trajdep	Grouping variable
1. Achievement variables	AV	AV-CY	AV-CY	AV	AV-CY	AV-CY	AV-CY	AV-CY	AV	CY
2. Socioeconomic variables	AV	For the variables that measure the same AV For the variables that measure another variable AV-CY	AV	AV	AV-CY				AV	CY
3. Parents' expectations	AV	AV-CY							AV	CY
4. Pre-school variables	AV	Variables from 4th grade.	-	AV				AV-CY	AV	School in 4th grade
5. SES index	AV-CY	AV-CY			AV					CY
6. SES selection index	AV-CY	AV-CY				AV				CY
7. Academic selection index	AV-CY	AV-CY		AV			AV			CY
8. Pre-school attendance index	AV-CY	AV-CY						AV		CY

B.2 Definition of multilevel growth models

Unconditional mean models:

- Model 1.a

$$y_{ti} = \beta_0 + \xi_i + \varepsilon_{ti}$$

- Model 1.b

$$y_{tij} = \beta_0 + \xi_{ij} + \mu_j^{4th} + \varepsilon_{tij}$$

- Model 1.c

$$y_{tij} = \beta_0 + \xi_{ij} + \mu_j^{8th} + \varepsilon_{tij}$$

- Model 1.d

$$y_{tij} = \beta_0 + \xi_{ij} + \mu_j^{10th} + \varepsilon_{tij}$$

- Model 1.e

$$y_{ti(jk)} = \beta_0 + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)}$$

- Model 1.f

$$y_{ti(jkl)} = \beta_0 + \xi_{i(jkl)} + \mu_j^{4th} + \mu_k^{8th} + \mu_l^{10th} + \varepsilon_{ti(jkl)}$$

Unconditional growth models:

- Model 2.a

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \upsilon_{i(jk)} \end{aligned}$$

- Model 2.b

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 2.c

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 2.d

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_k^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

Conditional mean models for trajectory characterization:

- Model 3.a

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{ECE 0-2}_{i(jk)} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \beta_{11}\text{ECE 0-2}_{i(jk)} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.b

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{ECE 2-4}_{i(jk)} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \beta_{11}\text{ECE 2-4}_{i(jk)} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.c

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{Pre kinder}_{i(jk)} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \beta_{11}\text{Pre kinder}_{i(jk)} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.d

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{Kinder}_{i(jk)} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \beta_{11}\text{Kinder}_{i(jk)} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.e

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{gender}_{i(jk)} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \beta_{11}\text{gender}_{i(jk)} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.f

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{Parents expect.}_{ti(jk)} + \beta_{11}\text{Parents expect.}_{ti(jk)} * \text{time} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} \\ &\quad + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.g

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{Mother educ.}_{i(jk)} + \beta_{11}\text{Mother educ.}_{ti(jk)} * \text{time} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} \\ &\quad + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.h

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{Father educ.}_{i(jk)} + \beta_{11}\text{Father educ.}_{ti(jk)} * \text{time} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} \\ &\quad + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.h

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{Income}_{ti(jk)} + \beta_{11}\text{Income}_{ti(jk)} * \text{time} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} \\ &\quad + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.i

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{N books}_{ti(jk)} + \beta_{11}\text{N books}_{ti(jk)} * \text{time} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} \\ &\quad + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.j

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{Computer}_{ti(jk)} + \beta_{11}\text{Computer}_{ti(jk)} * \text{time} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} \\ &\quad + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

- Model 3.k

$$\begin{aligned} y_{ti(jk)} &= \beta_0 + \beta_{01}\text{Internet}_{ti(jk)} + \beta_{11}\text{Internet}_{ti(jk)} * \text{time} + \beta_{1i(jk)}\text{time} + \xi_{i(jk)} \\ &\quad + \mu_j^{8th} + \mu_k^{10th} + \varepsilon_{ti(jk)} \\ \beta_{1i(jk)} &= \beta_{10} + \nu_j^{8th} + \nu_k^{10th} + \upsilon_{i(jk)} \end{aligned}$$

B.3 Results of unconditional multilevel growth models for trajectory description

Table B.2: Estimated parameters for the models 3.a, 3.b, 3.c, 3.d, 3.e and 3.f. in mathematics scores. The parameters are the mean trough 20 imputations, for random effect and fit indexes we present also the 5th and 95th percentile.

	Mathematics					
	3.a	3.b	3.c	3.d	3.e	3.f
Fixed effects						
Intercept	0.006 (0.013)	-0.108 (0.013)	-0.103 (0.013)	-0.102 (0.013)	-0.146 (0.013)	-0.721 (0.021)
time	0.009 (0.003)	0.001 (0.003)	0 (0.003)	0.001 (0.003)	-0.006 (0.004)	-0.073 (0.01)
Gender (girl)	-0.216 (0.005)					
time:Gender (girl)	-0.016 (0.002)					
ECE (0-2)		0.011 (0.009)				
time:ECE (0-2)		0 (0.004)				
ECE (2-4)			-0.012 (0.005)			
time:ECE (2-4)			0.002 (0.002)			
Pre-kinder				-0.009 (0.005)		
time:Pre-kinder				-0.001 (0.002)		
Kinder					0.053 (0.007)	
time:Kinder					0.009 (0.003)	
Parents expectations						0.038 (0.001)
time:Parents expectations						0.004 (0.001)
Random intercepts						
Residual mean	0.246	0.246	0.246	0.246	0.246	0.248
Residual P5	0.246	0.246	0.246	0.246	0.246	0.247
Residual P95	0.247	0.247	0.247	0.247	0.247	0.248
Student mean	0.356	0.365	0.365	0.365	0.365	0.351
Student P5	0.355	0.364	0.364	0.364	0.364	0.35
Student P95	0.357	0.366	0.366	0.366	0.366	0.352
School 2011 mean	0.047	0.047	0.047	0.047	0.047	0.045
School 2011 P5	0.046	0.047	0.047	0.047	0.046	0.044
School 2011 P95	0.048	0.048	0.048	0.048	0.047	0.045
School 2013 mean	0.346	0.34	0.341	0.341	0.337	0.311
School 2013 P5	0.343	0.338	0.338	0.338	0.334	0.309
School 2013 P95	0.347	0.342	0.342	0.342	0.339	0.313
Random slopes						
Student-time mean	0.022	0.022	0.022	0.022	0.022	0.021
Student-time P5	0.021	0.021	0.021	0.021	0.021	0.021
Student-time P95	0.022	0.022	0.022	0.022	0.022	0.021
School 2011-time mean	0.007	0.007	0.007	0.007	0.007	0.007
School 2011-time P5	0.007	0.007	0.007	0.007	0.007	0.007
School 2011-time P95	0.007	0.007	0.007	0.007	0.007	0.007
School 2013-time mean	0.013	0.013	0.013	0.013	0.012	0.012
School 2013-time P5	0.012	0.012	0.012	0.012	0.012	0.012
School 2013-time P95	0.013	0.013	0.013	0.013	0.013	0.013
Fit indexes						
Deviance mean	908123.745	910652.809	910643.656	910649.541	910538.153	907266.444
Deviance P5	907758.745	910297.439	910290.685	910296.814	910196.917	906856.365
Deviance P95	908379.238	910909.507	910892.572	910903.989	910796.216	907604.607

	3.a	3.b	3.c	3.d	3.e	3.f
df	424738	424738	424738	424738	424738	424738
AIC mean	908151.745	910680.809	910671.656	910677.541	910566.153	907294.444
AIC P5	907786.745	910325.439	910318.685	910324.814	910224.917	906884.365
AIC P95	908407.238	910937.507	910920.572	910931.989	910824.216	907632.607
BIC mean	908305.175	910834.238	910825.085	910830.971	910719.583	907447.874
BIC P5	907940.175	910478.868	910472.115	910478.244	910378.346	907037.795
BIC P95	908560.668	911090.936	911074.002	911085.418	910977.646	907786.037

Table B.3: Estimated parameters for the models 3.g, 3.h, 3.i, 3.j, 3.k and 3.l. in mathematics scores. The parameters are the mean trough 20 imputations, for random effect and fit indexes we present also the 5th and 95th percentile.

	Mathematics					
	3.g	3.h	3.i	3.j	3.k	3.l
Fixed effects						
Intercept	-0.243 (0.014)	-0.234 (0.014)	-0.133 (0.013)	-0.149 (0.013)	-0.151 (0.013)	-0.109 (0.013)
time	0.025 (0.005)	0.015 (0.005)	-0.004 (0.003)	-0.003 (0.003)	-0.011 (0.003)	-0.001 (0.003)
Mother education	0.012 (0.001)					
time:Mother education	-0.002 (0)					
Father education		0.012 (0.001)				
time:Father education		-0.001 (0)				
Home income			0.007 (0.001)			
time:Home income			0 (0)			
Number of books				0.001 (0)		
time:Number of books				0 (0)		
Computer					0.05 (0.005)	
time:Computer					0.006 (0.003)	
Internet						0.004 (0.004)
time:Internet						0.008 (0.003)
Random Intercepts						
Residual mean	0.247	0.247	0.246	0.247	0.247	0.246
Residual P5	0.246	0.246	0.246	0.246	0.246	0.246
Residual P95	0.247	0.247	0.247	0.248	0.247	0.247
Student mean	0.363	0.363	0.365	0.362	0.364	0.365
Student P5	0.361	0.362	0.363	0.36	0.362	0.364
Student P95	0.364	0.364	0.366	0.363	0.365	0.366
School 2011 mean	0.045	0.045	0.046	0.046	0.047	0.047
School 2011 P5	0.044	0.044	0.045	0.046	0.046	0.047
School 2011 P95	0.045	0.046	0.047	0.047	0.047	0.048
School 2013 mean	0.324	0.325	0.332	0.331	0.336	0.34
School 2013 P5	0.321	0.322	0.329	0.329	0.334	0.338
School 2013 P95	0.326	0.326	0.333	0.333	0.338	0.341
Random Slopes						
Student-time mean	0.022	0.022	0.022	0.022	0.022	0.022
Student-time P5	0.021	0.021	0.021	0.021	0.021	0.021
Student-time P95	0.022	0.022	0.022	0.022	0.022	0.022
School 2011-time mean	0.007	0.007	0.007	0.007	0.007	0.007
School 2011-time P5	0.007	0.007	0.007	0.007	0.007	0.007
School 2011-time P95	0.007	0.007	0.007	0.007	0.007	0.007
School 2013-time mean	0.013	0.013	0.012	0.013	0.013	0.012
School 2013-time P5	0.013	0.012	0.012	0.012	0.012	0.012
School 2013-time P95	0.013	0.013	0.013	0.013	0.013	0.013
Fit Indexes						
Deviance mean	909761.707	909880.019	910407.21	909762.3	910417.535	910642.901
Deviance P5	909402.711	909477.743	910035.114	909398.803	910057.79	910292.091
Deviance P95	910083.278	910108.005	910652.45	910053.633	910677.641	910900.443
df	424738	424738	424738	424738	424738	424738
AIC mean	909789.707	909908.019	910435.21	909790.3	910445.535	910670.901
AIC P5	909430.711	909505.743	910063.114	909426.803	910085.79	910320.091
AIC P95	910111.278	910136.005	910680.45	910081.633	910705.641	910928.443
BIC mean	909943.137	910061.449	910588.64	909943.729	910598.964	910824.331
BIC P5	909584.141	909659.173	910216.543	909580.232	910239.219	910473.52

	3.g	3.h	3.i	3.j	3.k	3.l
BIC P95	910264.708	910289.435	910833.88	910235.062	910859.07	911081.873

Table B.4: Estimated parameters for the models 3.a, 3.b, 3.c, 3.d, 3.e and 3.f. in reading scores. The parameters are the mean trough 20 imputations, for random effect and fit indexes we present also the 5th and 95th percentile.

	Reading					
	3.a	3.b	3.c	3.d	3.e	3.f
Fixed Effects						
Intercept	-0.145 (0.011)	-0.054 (0.011)	-0.048 (0.011)	-0.043 (0.012)	-0.091 (0.012)	-0.765 (0.02)
time	-0.011 (0.003)	0.011 (0.003)	0.01 (0.003)	0.011 (0.003)	0.004 (0.004)	-0.095 (0.009)
Gender (girl)	0.172 (0.005)					
time:Gender (girl)	0.04 (0.002)					
ECE (0-2)		0.001 (0.01)				
time:ECE (0-2)		-0.003 (0.004)				
ECE (2-4)			-0.018 (0.006)			
time:ECE (2-4)			0.002 (0.002)			
Pre-kinder				-0.02 (0.006)		
time:Pre-kinder				-0.002 (0.003)		
Kinder					0.05 (0.007)	
time:Kinder					0.009 (0.004)	
Parents expectations						0.044 (0.001)
time:Parents expectations						0.006 (0.001)
Random Intercepts						
Residual mean	0.316	0.316	0.316	0.316	0.316	0.317
Residual P5	0.315	0.315	0.315	0.315	0.315	0.317
Residual P95	0.317	0.317	0.317	0.317	0.317	0.318
Student mean	0.408	0.414	0.414	0.414	0.413	0.396
Student P5	0.406	0.412	0.412	0.412	0.412	0.394
Student P95	0.409	0.415	0.415	0.415	0.415	0.397
School 2011 mean	0.031	0.032	0.032	0.032	0.031	0.03
School 2011 P5	0.031	0.031	0.031	0.031	0.031	0.029
School 2011 P95	0.032	0.032	0.032	0.032	0.032	0.03
School 2013 mean	0.244	0.249	0.25	0.25	0.247	0.22
School 2013 P5	0.243	0.248	0.249	0.249	0.246	0.218
School 2013 P95	0.245	0.251	0.251	0.252	0.248	0.222
Random Slopes						
Student-time mean	0.014	0.014	0.014	0.014	0.014	0.014
Student-time P5	0.014	0.014	0.014	0.014	0.014	0.014
Student-time P95	0.014	0.015	0.015	0.015	0.015	0.014
School 2011-time mean	0.005	0.005	0.005	0.005	0.005	0.005
School 2011-time P5	0.005	0.005	0.005	0.005	0.005	0.005
School 2011-time P95	0.005	0.005	0.005	0.005	0.005	0.005
School 2013-time mean	0.008	0.009	0.009	0.009	0.009	0.008
School 2013-time P5	0.008	0.008	0.008	0.009	0.008	0.008
School 2013-time P95	0.009	0.009	0.009	0.009	0.009	0.008
Fit Indexes						
Deviance mean	976494.124	977838.919	977818.424	977818.672	977751.239	974348.497
Deviance P5	976128.381	977496.611	977479.137	977483.413	977412.375	973971.894
Deviance P95	976887.639	978223.776	978206.551	978211.27	978125.697	974745.757
df	424738	424738	424738	424738	424738	424738
AIC mean	976522.124	977866.919	977846.424	977846.672	977779.239	974376.497
AIC P5	976156.381	977524.611	977507.137	977511.413	977440.375	973999.894
AIC P95	976915.639	978251.776	978234.551	978239.27	978153.697	974773.757
BIC mean	976675.553	978020.349	977999.854	978000.102	977932.669	974529.926
BIC P5	976309.811	977678.04	977660.566	977664.842	977593.805	974153.324

	3.a	3.b	3.c	3.d	3.e	3.f
BIC P95	977069.068	978405.205	978387.981	978392.7	978307.127	974927.187

Table B.5: Estimated parameters for the models 3.g, 3.h, 3.i, 3.j, 3.k and 3.l. in reading scores. The parameters are the mean trough 20 imputations, for random effect and fit indexes we present also the 5th and 95th percentile.

	Reading					
	3.g	3.h	3.i	3.j	3.k	3.l
Fixed Effects						
Intercept	-0.187 (0.013)	-0.176 (0.013)	-0.067 (0.011)	-0.108 (0.011)	-0.115 (0.012)	-0.056 (0.011)
time	0.035 (0.005)	0.024 (0.005)	0.012 (0.003)	0.002 (0.003)	-0.009 (0.004)	0.01 (0.003)
Mother education	0.012 (0.001)					
time:Mother education	-0.002 (0)					
Father education		0.011 (0.001)				
time:Father education		-0.001 (0)				
Home income			0.003 (0.001)			
time:Home income			-0.001 (0)			
Number of books				0.001 (0)		
time:Number of books				0 (0)		
Computer					0.07 (0.006)	
time:Computer					0.016 (0.004)	
Internet						0.002 (0.005)
time:Internet						-0.003 (0.003)
Random Intercepts						
Residual mean	0.316	0.316	0.316	0.316	0.316	0.316
Residual P5	0.316	0.316	0.315	0.316	0.315	0.315
Residual P95	0.317	0.317	0.317	0.317	0.317	0.317
Student mean	0.411	0.411	0.413	0.408	0.412	0.414
Student P5	0.41	0.41	0.412	0.407	0.41	0.412
Student P95	0.413	0.413	0.415	0.41	0.413	0.415
School 2011 mean	0.03	0.031	0.031	0.031	0.031	0.032
School 2011 P5	0.03	0.03	0.031	0.03	0.031	0.031
School 2011 P95	0.031	0.031	0.032	0.031	0.032	0.032
School 2013 mean	0.236	0.236	0.246	0.239	0.245	0.249
School 2013 P5	0.235	0.235	0.245	0.238	0.243	0.248
School 2013 P95	0.237	0.238	0.247	0.241	0.246	0.25
Random Slopes						
Student-time mean	0.014	0.014	0.014	0.014	0.014	0.014
Student-time P5	0.014	0.014	0.014	0.014	0.014	0.014
Student-time P95	0.015	0.015	0.015	0.015	0.015	0.015
School 2011-time mean	0.005	0.005	0.005	0.005	0.005	0.005
School 2011-time P5	0.005	0.005	0.005	0.005	0.005	0.005
School 2011-time P95	0.005	0.005	0.005	0.005	0.005	0.005
School 2013-time mean	0.009	0.009	0.009	0.009	0.009	0.009
School 2013-time P5	0.009	0.009	0.008	0.008	0.008	0.009
School 2013-time P95	0.009	0.009	0.009	0.009	0.009	0.009
Fit Indexes						
Deviance mean	977094.791	977218.371	977741.205	976871.312	977550.588	977835.129
Deviance P5	976752.094	976870.695	977399.821	976506.51	977195.303	977490.15
Deviance P95	977462.683	977585.506	978117.957	977251.433	977922.09	978217.863
df	424738	424738	424738	424738	424738	424738
AIC mean	977122.791	977246.371	977769.205	976899.312	977578.588	977863.129
AIC P5	976780.094	976898.695	977427.821	976534.51	977223.303	977518.15
AIC P95	977490.683	977613.506	978145.957	977279.433	977950.09	978245.863
BIC mean	977276.22	977399.801	977922.635	977052.742	977732.018	978016.559
BIC P5	976933.524	977052.125	977581.251	976687.94	977376.733	977671.579

	3.g	3.h	3.i	3.j	3.k	3.l
BIC P95	977644.112	977766.936	978299.387	977432.862	978103.52	978399.293

B.4 Commont support control variables

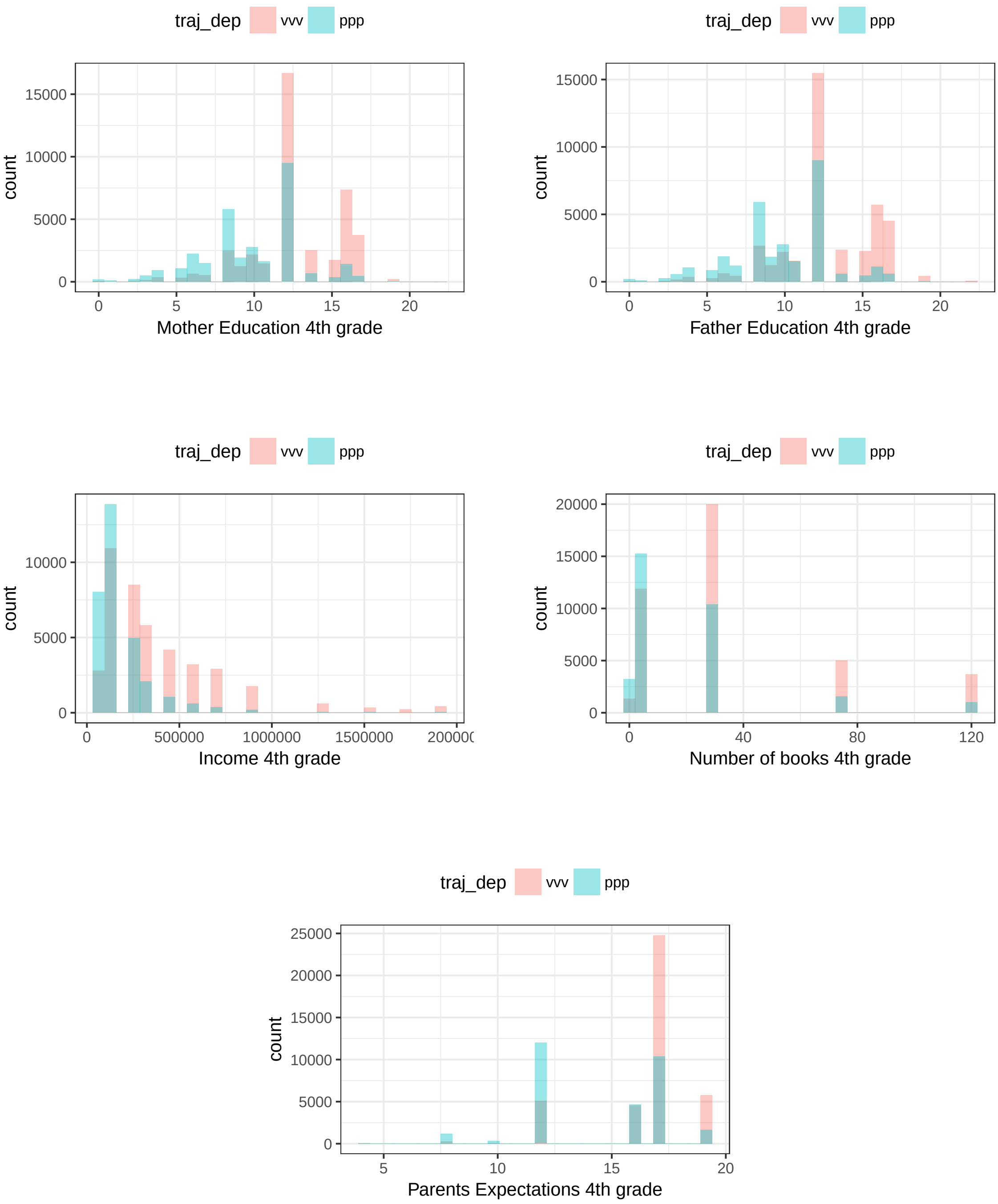


Figure B.1: Histograms for control variables in 4th grade

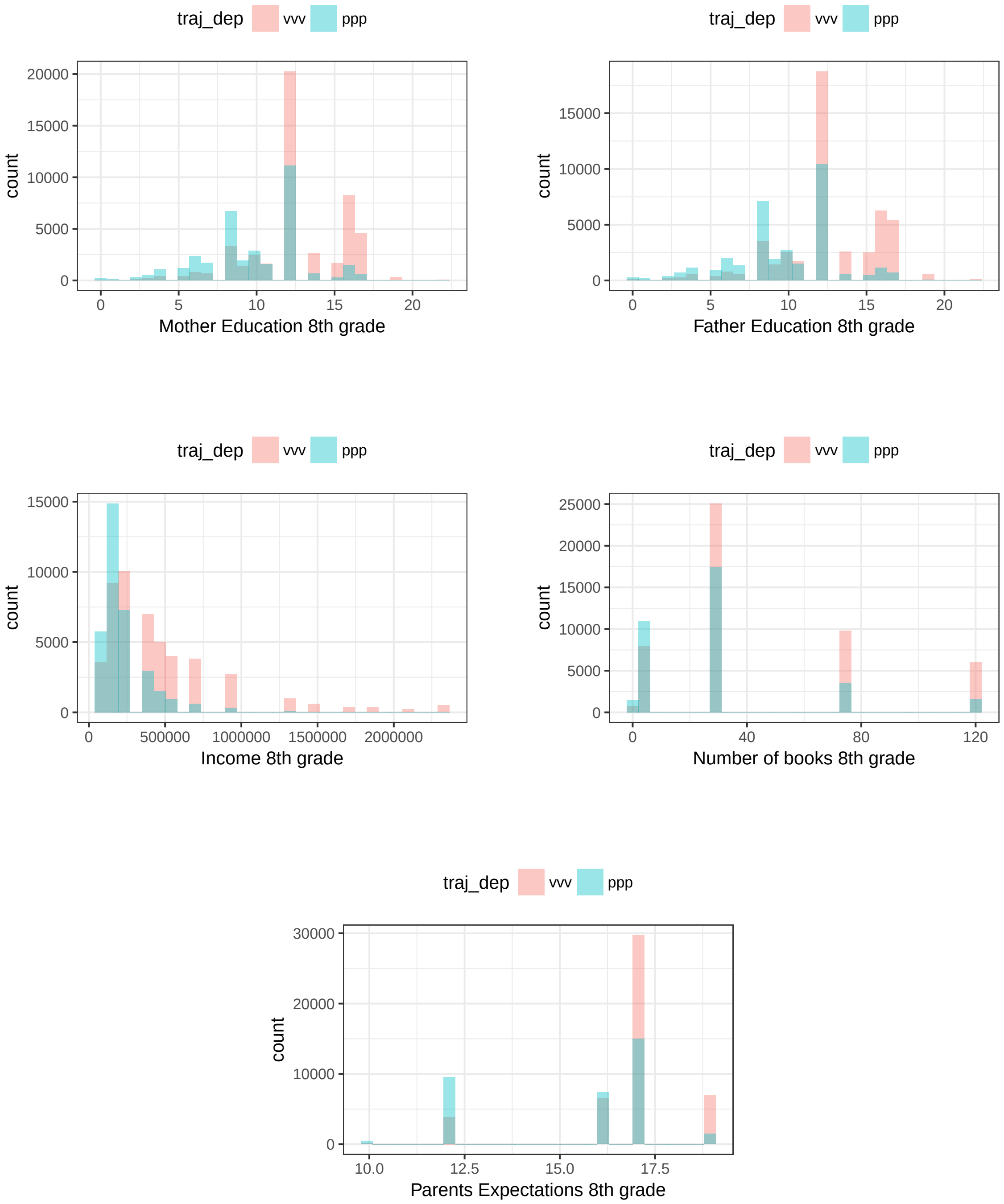


Figure B.2: Histograms for control variables in 8th grade

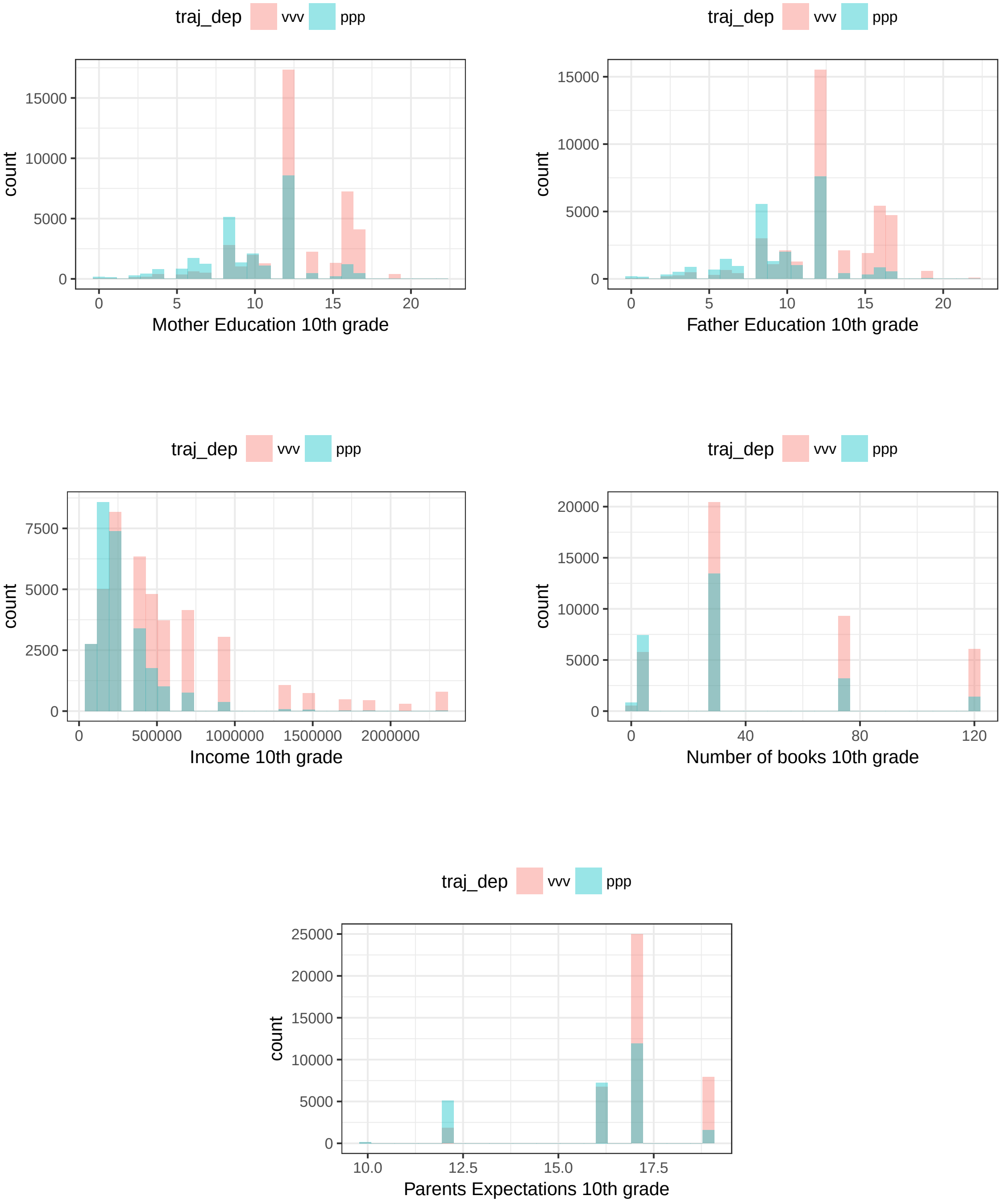


Figure B.3: Histograms for control variables in 10th grade

B.5 Commont support treatment variables

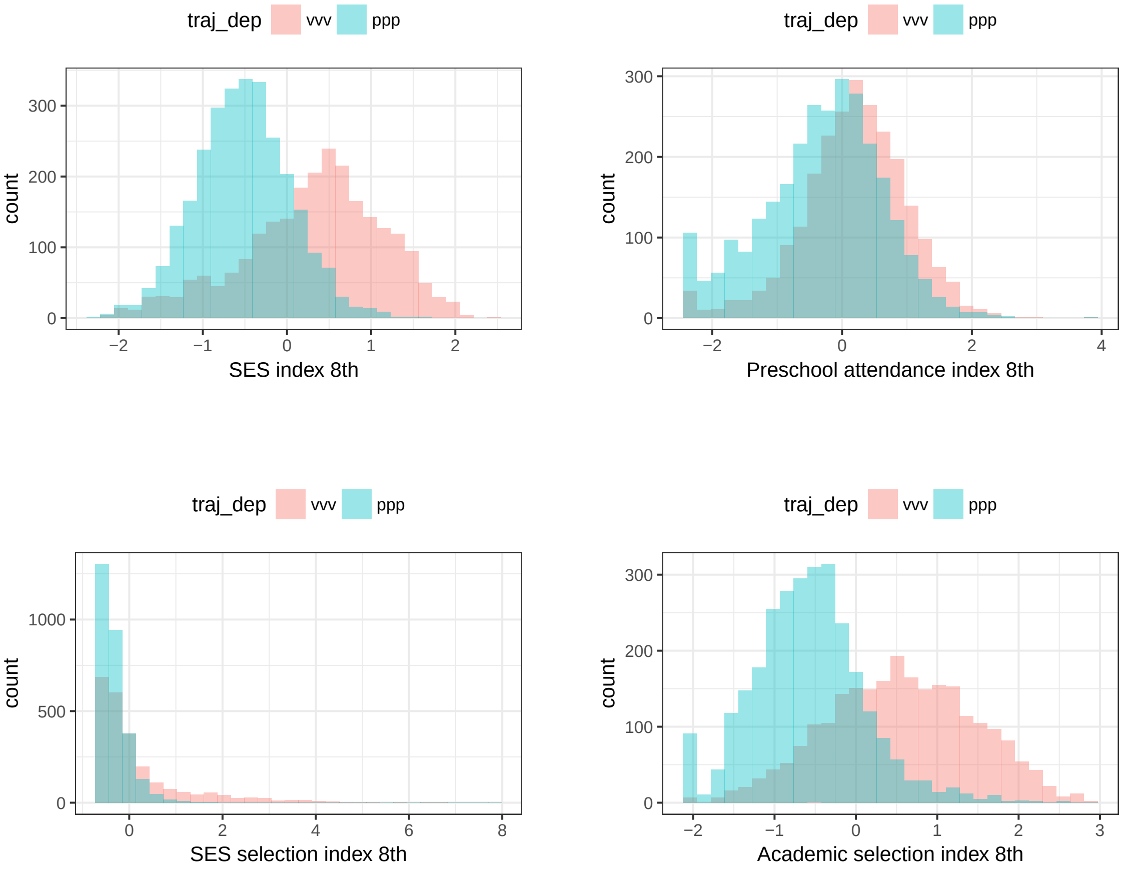


Figure B.4: Histograms for school level variables at 8th grade (treatment variables)

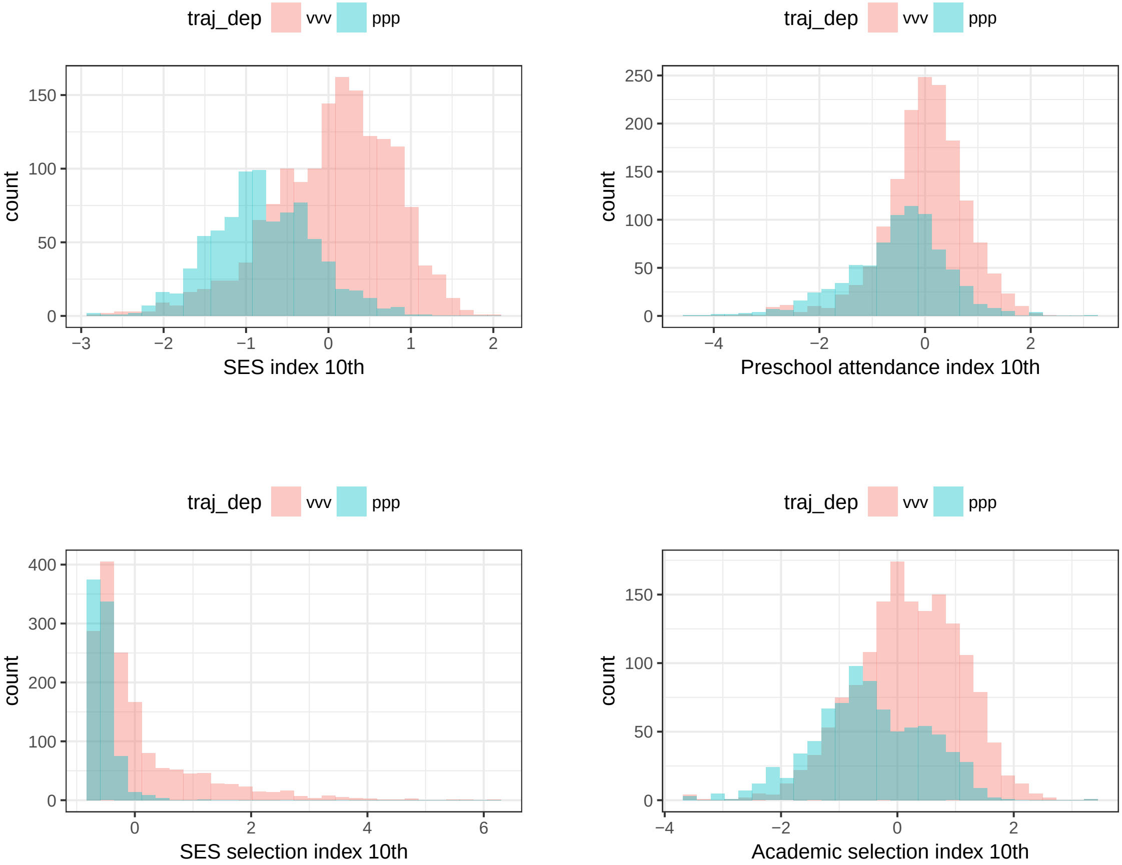


Figure B.5: Histograms for school level variables at 10th grade (treatment variables)

B.6 Results of unconditional multilevel growth models for treatment effect estimation in the complete sample

Table B.6: Estimated parameters for the models 4.a, 4.b, 4.c in mathematics scores. The parameters are the mean trough 20 imputations, for random effect and fit indexes we present also the 5th and 95th percentile.

	Mathematics		
	4.a	4.b	4.c
Fixed Effects			
Intercept	0.164 (0.015)	-0.56 (0.028)	-0.62 (0.027)
time	0.024 (0.004)	-0.05 (0.013)	-0.058 (0.014)
Trajdep	-0.657 (0.025)	-0.565 (0.023)	-0.064 (0.022)
time:Trajdep	-0.075 (0.006)	-0.08 (0.006)	-0.048 (0.007)
ECE (0-2)		0.017 (0.011)	0.017 (0.011)
time:ECE (0-2)		-0.001 (0.005)	-0.001 (0.005)
ECE (2-4)		-0.019 (0.008)	-0.018 (0.008)
time:ECE (2-4)		0.002 (0.004)	0.003 (0.004)
Pre-kinder		-0.031 (0.009)	-0.031 (0.009)
time:Pre-kinder		-0.003 (0.004)	-0.003 (0.004)
Kinder		0.058 (0.009)	0.056 (0.009)
time:Kinder		0.014 (0.004)	0.014 (0.004)
Gender (girl)		-0.22 (0.005)	-0.221 (0.005)
time:Gender (girl)		-0.021 (0.002)	-0.021 (0.002)
Mother education		0.005 (0.001)	0.003 (0.001)
time:Mother education		-0.002 (0)	-0.002 (0)
Father education		0.006 (0.001)	0.005 (0.001)
time:Father education		0 (0)	0.001 (0)
Parents expectations		0.037 (0.001)	0.036 (0.001)
time:Parents expectations		0.004 (0.001)	0.005 (0.001)
Number of books		0.001 (0)	0.001 (0)
time:Number of books		0 (0)	0 (0)
Home income		0.004 (0.001)	0.003 (0.001)
time:Home income		0.002 (0.001)	0.002 (0.001)
Computer		0.046 (0.007)	0.045 (0.007)
time:Computer		0.008 (0.004)	0.009 (0.004)
Internet		-0.03 (0.005)	-0.032 (0.005)
time:Internet		0.006 (0.004)	0.007 (0.004)
SES index 8th			0.084 (0.011)
time:SES index 8th			-0.027 (0.005)
SES index 10th			0.424 (0.017)
time:SES index 10th			0.042 (0.006)
Preschool attendance index 8th			-0.057 (0.007)
time:Preschool attendance index 8th			0 (0.003)
Preschool attendance index 10th			-0.159 (0.013)
time:Preschool attendance index 10th			-0.014 (0.004)
Academic Selection index 8th			0.02 (0.008)
time:Academic Selection index 8th			0.001 (0.003)
Academic Selection index 10th			0.118 (0.011)
time:Academic Selection index 10th			0.018 (0.003)
SES Selection index 8th			0.013 (0.006)

B.6. Results of unconditional multilevel growth models for treatment effect estimation in the complete sample

	4.a	4.b	4.c
time:SES Selection index 8th			-0.006 (0.003)
SES Selection index 10th			0.11 (0.012)
time:SES Selection index 10th			0.022 (0.004)
Random Intercepts			
Residual mean	0.245	0.247	0.246
Residual P5	0.244	0.246	0.246
Residual P95	0.246	0.247	0.247
Student mean	0.377	0.345	0.346
Student P5	0.374	0.344	0.344
Student P95	0.378	0.346	0.347
School 2011 mean	0.04	0.038	0.036
School 2011 P5	0.04	0.037	0.035
School 2011 P95	0.041	0.038	0.036
School 2013 mean	0.279	0.242	0.122
School 2013 P5	0.277	0.24	0.12
School 2013 P95	0.281	0.243	0.123
Random Slopes			
Student-time mean	0.021	0.02	0.02
Student-time P5	0.02	0.02	0.02
Student-time P95	0.021	0.02	0.02
School 2011-time mean	0.008	0.008	0.008
School 2011-time P5	0.008	0.008	0.008
School 2011-time P95	0.008	0.008	0.008
School 2013-time mean	0.011	0.01	0.009
School 2013-time P5	0.01	0.01	0.009
School 2013-time P95	0.011	0.01	0.009
Fit Indexes			
Deviance mean	624520.454	619190.279	617344.632
Deviance P5	624233.826	618858.83	617042.21
Deviance P95	624755.097	619465.472	617618.82
df	290908	290884	290868
AIC mean	624548.454	619266.279	617452.632
AIC P5	624261.826	618934.83	617150.21
AIC P95	624783.097	619541.472	617726.82
BIC mean	624696.586	619668.35	618023.996
BIC P5	624409.958	619336.901	617721.573
BIC P95	624931.228	619943.542	618298.184

Table B.7: Estimated parameters for the models 4.a, 4.b, 4.c in reading scores. The parameters are the mean trough 20 imputations, for random effect and fit indexes we present also the 5th and 95th percentile.

	Reading		
	4.a	4.b	4.c
Fixed Effects			
Intercept	0.153 (0.013)	-0.859 (0.027)	-0.888 (0.026)
time	0.013 (0.003)	-0.126 (0.013)	-0.135 (0.013)
Trajdep	-0.513 (0.021)	-0.41 (0.02)	-0.058 (0.02)
time:Trajdep	-0.022 (0.005)	-0.024 (0.006)	-0.029 (0.007)
ECE (0-2)		0.008 (0.013)	0.008 (0.013)
time:ECE (0-2)		-0.005 (0.005)	-0.005 (0.005)
ECE (2-4)		-0.021 (0.009)	-0.02 (0.009)
time:ECE (2-4)		0.004 (0.004)	0.004 (0.004)
Pre-kinder		-0.037 (0.01)	-0.036 (0.01)
time:Pre-kinder		-0.004 (0.004)	-0.004 (0.004)
Kinder		0.063 (0.009)	0.062 (0.009)
time:Kinder		0.014 (0.005)	0.015 (0.005)
Gender (girl)		0.163 (0.006)	0.162 (0.006)
time:Gender (girl)		0.032 (0.003)	0.032 (0.003)
Mother education		0.005 (0.001)	0.004 (0.001)
time:Mother education		-0.002 (0.001)	-0.001 (0.001)
Father education		0.006 (0.001)	0.006 (0.001)
time:Father education		0.001 (0.001)	0.001 (0.001)
Parents expectations		0.04 (0.001)	0.04 (0.001)
time:Parents expectations		0.006 (0.001)	0.006 (0.001)
Number of books		0.001 (0)	0.001 (0)
time:Number of books		0 (0)	0 (0)
Home income		0 (0.001)	0 (0.001)
time:Home income		0 (0.001)	0 (0.001)
Computer		0.07 (0.008)	0.069 (0.008)
time:Computer		0.028 (0.005)	0.03 (0.005)
Internet		-0.035 (0.007)	-0.036 (0.007)
time:Internet		-0.01 (0.004)	-0.009 (0.004)
SES index 8th			0.037 (0.011)
time:SES index 8th			-0.027 (0.005)
SES index 10th			0.353 (0.016)
time:SES index 10th			0.01 (0.006)
Preschool attendance index 8th			-0.05 (0.007)
time:Preschool attendance index 8th			0.005 (0.003)
Preschool attendance index 10th			-0.158 (0.013)
time:Preschool attendance index 10th			-0.006 (0.004)
Academic Selection index 8th			0.009 (0.008)
time:Academic Selection index 8th			0.002 (0.003)
Academic Selection index 10th			0.081 (0.01)
time:Academic Selection index 10th			0.003 (0.003)
SES Selection index 8th			0.006 (0.006)
time:SES Selection index 8th			-0.008 (0.003)
SES Selection index 10th			0.09 (0.011)
time:SES Selection index 10th			0.012 (0.004)
Random Intercepts			
Residual mean	0.315	0.317	0.317
Residual P5	0.314	0.316	0.316

B.6. Results of unconditional multilevel growth models for treatment effect estimation in the complete sample

	4.a	4.b	4.c
Residual P95	0.316	0.318	0.318
Student mean	0.424	0.394	0.395
Student P5	0.422	0.392	0.392
Student P95	0.426	0.396	0.397
School 2011 mean	0.029	0.026	0.025
School 2011 P5	0.028	0.026	0.024
School 2011 P95	0.03	0.027	0.026
School 2013 mean	0.203	0.164	0.096
School 2013 P5	0.2	0.162	0.094
School 2013 P95	0.205	0.166	0.098
Random Slopes			
Student-time mean	0.013	0.013	0.013
Student-time P5	0.013	0.012	0.012
Student-time P95	0.013	0.013	0.013
School 2011-time mean	0.005	0.005	0.005
School 2011-time P5	0.005	0.005	0.005
School 2011-time P95	0.005	0.005	0.005
School 2013-time mean	0.008	0.008	0.008
School 2013-time P5	0.008	0.007	0.007
School 2013-time P95	0.008	0.008	0.008
Fit Indexes			
Deviance mean	670931.501	666619.467	665177.613
Deviance P5	670687.441	666342.954	664875.297
Deviance P95	671281.981	666985.416	665579.98
df	290908	290884	290868
AIC mean	670959.501	666695.467	665285.613
AIC P5	670715.441	666418.954	664983.297
AIC P95	671309.981	667061.416	665687.98
BIC mean	671107.632	667097.538	665856.977
BIC P5	670863.572	666821.024	665554.66
BIC P95	671458.112	667463.487	666259.343

B.7 Results of unconditional multilevel growth models for treatment effect estimation in the matched sample

Table B.8: Estimated parameters for the models 5.a, 5.b, 5.c and 5.d in mathematics scores. The parameters are the mean trough 20 imputations, for random effect and fit indexes we present also the 5th and 95th percentile.

	Mathematics			
	5.a	5.b	5.c	5.d
Fixed Effects				
Intercept	0.059 (0.016)	-0.01 (0.014)	-0.595 (0.036)	-0.612 (0.035)
time	0.021 (0.004)	0.024 (0.005)	-0.072 (0.017)	-0.079 (0.017)
Trajdep	-0.417 (0.026)	0.028 (0.024)	-0.412 (0.025)	0.002 (0.023)
time:Trajdep	-0.077 (0.007)	-0.048 (0.008)	-0.072 (0.007)	-0.043 (0.008)
ECE (0-2)			0.018 (0.017)	0.017 (0.017)
time:ECE (0-2)			-0.005 (0.008)	-0.005 (0.007)
ECE (2-4)			-0.019 (0.011)	-0.017 (0.011)
time:ECE (2-4)			0.004 (0.005)	0.004 (0.005)
Pre-kinder			-0.035 (0.011)	-0.034 (0.011)
time:Pre-kinder			-0.003 (0.005)	-0.003 (0.005)
Kinder			0.048 (0.013)	0.045 (0.013)
time:Kinder			0.014 (0.005)	0.014 (0.005)
Gender (girl)			-0.231 (0.008)	-0.232 (0.008)
time:Gender (girl)			-0.02 (0.004)	-0.019 (0.004)
Mother education			0.004 (0.001)	0.003 (0.001)
time:Mother education			-0.002 (0.001)	-0.001 (0.001)
Father education			0.005 (0.001)	0.003 (0.001)
time:Father education			-0.001 (0.001)	0 (0.001)
Parents expectations			0.037 (0.002)	0.036 (0.002)
time:Parents expectations			0.006 (0.001)	0.006 (0.001)
Number of books			0.001 (0)	0.001 (0)
time:Number of books			0 (0)	0 (0)
Home income			0.002 (0.001)	0.001 (0.001)
time:Home income			0.001 (0.001)	0.002 (0.001)
Computer			0.061 (0.009)	0.059 (0.009)
time:Computer			0.017 (0.005)	0.018 (0.005)
Internet			-0.041 (0.007)	-0.044 (0.007)
time:Internet			-0.003 (0.006)	-0.002 (0.006)
SES index 8th		0.092 (0.014)		0.066 (0.014)
time:SES index 8th		-0.035 (0.006)		-0.032 (0.006)
SES index 10th		0.438 (0.021)		0.409 (0.02)
time:SES index 10th		0.039 (0.007)		0.036 (0.007)
Preschool attendance index 8th		-0.065 (0.01)		-0.062 (0.01)
time:Preschool attendance index 8th		0.001 (0.004)		0.001 (0.004)
Preschool attendance index 10th		-0.171 (0.016)		-0.161 (0.015)
time:Preschool attendance index 10th		-0.015 (0.005)		-0.014 (0.005)
Academic Selection index 8th		0.041 (0.01)		0.04 (0.01)
time:Academic Selection index 8th		0.006 (0.004)		0.006 (0.004)
Academic Selection index 10th		0.119 (0.012)		0.116 (0.012)
time:Academic Selection index 10th		0.016 (0.004)		0.016 (0.004)
SES Selection index 8th		0.011 (0.009)		0.017 (0.009)

B.7. Results of unconditional multilevel growth models for treatment effect estimation in the matched sample

	5.a	5.b	5.c	5.d
time:SES Selection index 8th		-0.006 (0.004)		-0.005 (0.004)
SES Selection index 10th		0.122 (0.015)		0.118 (0.015)
time:SES Selection index 10th		0.024 (0.005)		0.024 (0.005)
Random Intercepts				
Residual mean	0.245	0.245	0.247	0.247
Residual P5	0.244	0.244	0.246	0.246
Residual P95	0.247	0.247	0.248	0.248
Student mean	0.369	0.369	0.339	0.339
Student P5	0.366	0.366	0.336	0.337
Student P95	0.371	0.37	0.341	0.341
School 2011 mean	0.038	0.036	0.037	0.035
School 2011 P5	0.038	0.035	0.036	0.034
School 2011 P95	0.039	0.036	0.038	0.036
School 2013 mean	0.266	0.133	0.239	0.124
School 2013 P5	0.264	0.131	0.236	0.123
School 2013 P95	0.269	0.135	0.241	0.125
Random Slopes				
Student-time mean	0.019	0.019	0.018	0.018
Student-time P5	0.018	0.018	0.018	0.018
Student-time P95	0.019	0.019	0.019	0.019
School 2011-time mean	0.008	0.008	0.008	0.008
School 2011-time P5	0.008	0.008	0.008	0.008
School 2011-time P95	0.008	0.008	0.008	0.008
School 2013-time mean	0.01	0.009	0.01	0.009
School 2013-time P5	0.01	0.009	0.01	0.009
School 2013-time P95	0.011	0.01	0.01	0.009
Fit Indexes				
Deviance mean	289260.396	287753.024	286896.789	285530.647
Deviance P5	289060.769	287556.115	286693.532	285327.291
Deviance P95	289544.227	288022.578	287107.754	285737.505
df	134038	134022	134014	133998
AIC mean	289288.396	287813.024	286972.789	285638.647
AIC P5	289088.769	287616.115	286769.532	285435.291
AIC P95	289572.227	288082.578	287183.754	285845.505
BIC mean	289425.68	288107.203	287345.416	286168.17
BIC P5	289226.052	287910.295	287142.16	285964.814
BIC P95	289709.511	288376.758	287556.381	286375.028

Table B.9: Estimated parameters for the models 5.a, 5.b, 5.c and 5.d in reading scores. The parameters are the mean trough 20 imputations, for random effect and fit indexes we present also the 5th and 95th percentile.

	Reading			
	5.a	5.b	5.c	5.d
Fixed Effects				
Intercept	0.054 (0.014)	0.013 (0.014)	-0.858 (0.038)	-0.855 (0.037)
time	0.016 (0.004)	0.02 (0.005)	-0.124 (0.017)	-0.137 (0.018)
Trajdep	-0.282 (0.024)	0.051 (0.023)	-0.27 (0.023)	0.023 (0.022)
time:Trajdep	-0.027 (0.007)	-0.024 (0.008)	-0.021 (0.007)	-0.021 (0.008)
ECE (0-2)			-0.005 (0.02)	-0.005 (0.02)
time:ECE (0-2)			-0.013 (0.009)	-0.013 (0.009)
ECE (2-4)			-0.015 (0.013)	-0.011 (0.013)
time:ECE (2-4)			0.008 (0.005)	0.008 (0.005)
Pre-kinder			-0.042 (0.012)	-0.04 (0.012)
time:Pre-kinder			-0.004 (0.006)	-0.004 (0.006)
Kinder			0.058 (0.012)	0.056 (0.012)
time:Kinder			0.014 (0.007)	0.015 (0.007)
Gender (girl)			0.164 (0.009)	0.162 (0.009)
time:Gender (girl)			0.036 (0.004)	0.037 (0.004)
Mother education			0.004 (0.001)	0.003 (0.001)
time:Mother education			-0.001 (0.001)	-0.001 (0.001)
Father education			0.005 (0.001)	0.004 (0.001)
time:Father education			-0.001 (0.001)	0 (0.001)
Parents expectations			0.039 (0.002)	0.039 (0.002)
time:Parents expectations			0.006 (0.001)	0.006 (0.001)
Number of books			0.001 (0)	0.001 (0)
time:Number of books			0 (0)	0 (0)
Home income			-0.002 (0.002)	-0.003 (0.002)
time:Home income			-0.001 (0.001)	0 (0.001)
Computer			0.076 (0.01)	0.075 (0.01)
time:Computer			0.035 (0.006)	0.036 (0.006)
Internet			-0.037 (0.008)	-0.038 (0.008)
time:Internet			-0.013 (0.007)	-0.011 (0.007)
SES index 8th		0.054 (0.015)		0.035 (0.015)
time:SES index 8th		-0.023 (0.006)		-0.02 (0.006)
SES index 10th		0.386 (0.02)		0.342 (0.02)
time:SES index 10th		0.009 (0.007)		0.002 (0.007)
Preschool attendance index 8th		-0.059 (0.01)		-0.051 (0.01)
time:Preschool attendance index 8th		0.005 (0.004)		0.006 (0.004)
Preschool attendance index 10th		-0.183 (0.016)		-0.17 (0.015)
time:Preschool attendance index 10th		-0.008 (0.006)		-0.007 (0.006)
Academic Selection index 8th		0.023 (0.01)		0.023 (0.01)
time:Academic Selection index 8th		0.003 (0.004)		0.003 (0.004)
Academic Selection index 10th		0.082 (0.012)		0.08 (0.012)
time:Academic Selection index 10th		0.001 (0.004)		0.001 (0.004)
SES Selection index 8th		0.012 (0.009)		0.008 (0.009)
time:SES Selection index 8th		-0.009 (0.004)		-0.009 (0.004)
SES Selection index 10th		0.103 (0.015)		0.099 (0.015)
time:SES Selection index 10th		0.015 (0.005)		0.014 (0.005)
Random Intercepts				
Residual mean	0.313	0.314	0.315	0.314
Residual P5	0.312	0.312	0.313	0.313

B.7. Results of unconditional multilevel growth models for treatment effect estimation in the matched sample

	5.a	5.b	5.c	5.d
Residual P95	0.314	0.315	0.316	0.316
Student mean	0.411	0.41	0.383	0.384
Student P5	0.408	0.406	0.38	0.381
Student P95	0.413	0.413	0.386	0.386
School 2011 mean	0.031	0.03	0.029	0.028
School 2011 P5	0.03	0.029	0.028	0.027
School 2011 P95	0.033	0.031	0.03	0.03
School 2013 mean	0.202	0.115	0.172	0.104
School 2013 P5	0.199	0.113	0.169	0.102
School 2013 P95	0.204	0.117	0.175	0.106
Random Slopes				
Student-time mean	0.013	0.012	0.012	0.012
Student-time P5	0.012	0.011	0.011	0.011
Student-time P95	0.013	0.013	0.013	0.012
School 2011-time mean	0.006	0.005	0.006	0.005
School 2011-time P5	0.005	0.005	0.005	0.005
School 2011-time P95	0.006	0.006	0.006	0.005
School 2013-time mean	0.008	0.008	0.008	0.008
School 2013-time P5	0.008	0.008	0.008	0.008
School 2013-time P95	0.008	0.008	0.008	0.008
Fit Indexes				
Deviance mean	309570.808	308391.509	307731.79	306687.23
Deviance P5	309408.977	308224.244	307501.789	306457.871
Deviance P95	309765.159	308573.966	307917.922	306889.304
df	134038	134022	134014	133998
AIC mean	309598.808	308451.509	307807.79	306795.23
AIC P5	309436.977	308284.244	307577.789	306565.871
AIC P95	309793.159	308633.966	307993.922	306997.304
BIC mean	309736.091	308745.689	308180.417	307324.753
BIC P5	309574.261	308578.423	307950.417	307095.394
BIC P95	309930.443	308928.145	308366.55	307526.827