Universitat Pompeu Fabra

Department of Political and Social Sciences

Educational institutions and their effect on inequality

Three Papers on Educational Systems and Inequality of Achievement and Opportunity

Jorge Cimentada

Director de la tesi: Bruno Arpino & Gøsta Esping-Andersen

TESI DOCTORAL UPF 2018

To my mother and my father

This thesis embraces the open science and reproducibility movement and can be completely reproduced following the directions on this website:
github.com/cimentadaj/phd_thesis

Acknowledgments

This PhD would not have been possible without the help of many people. I am fortunate enough to have met my advisers Bruno Arpino and Gøsta Esping-Andersen while completing my Master's degree at Pompeu Fabra University. Bruno has been extremely supportive of my work while allowing me to bother him with endless questions on statistics. You supported many of the projects I wanted to follow only raising concerns that would make them even better. You gave me something that not many advisers give to their PhD students: freedom. I thank you for these three years of unflinching support. I would also like to thank Gøsta Esping-Andersen for giving me the opportunity to work with him throughout these three years. You pushed me towards boundaries I though I could never achieve. Intellectually, I have learned a great deal from your helpful feedback and I thank you for always encouraging me to follow crazy ideas. I never would have been where I am without your support. Thank you both.

I would not have been able to complete this thesis without the help of my friends in the Department of Political and Social Sciences at Pompeu Fabra University. To my fellow PhD candidates João Tavares, Basilio Moreno, My Lan Do Nguyen, Wanli Nie and Luisa Faustini, this PhD would not have been fun without you. To my office mates Luize Ratniece, Daniela Bellani, Lesia Nedoluzhko and Alessandro Di Nalo, I have learned so much from all of you that I cannot summarize it in a paragraph. This thesis goes out to you guys. I would also like to thank Roberta Rutigliano for being there for me. You helped overcome many things while pursuing this goal and I can only hope to give you back half of what you gave me. To Léa Pessin for encouraging me to embark on this PhD when I was in doubt.

My last year in the PhD was spent at the Research and Expertise Center for Survey Methodology (RECSM) thanks to the generosity of Wiebke Weber. Wiebke, I owe you so much for allowing to follow my intellectual endeavors. All the work we developed would not have been possible without your feedback and support. You and the RECSM team showed me that there is still hope to have a rigorous, innovative and fun job while still being in academia. To all other members of the RECSM: Anna De Castellarnau, André Gomes Pirralha, Hannah Schwarz and Diana Zavala-Rojas, I thank you for welcoming me to the team. A special thank you goes out to Oriol Bosch for keeping up with my daily bothering s. Hope we can work together again in the future.

At Pompeu Fabra University I would like to thank Meritxell Canals, Sergio Vich and Maria José Gonzales for answering endless questions about the PhD program. In no particular order of preference, I would like to thank John Palmer, Danilo Serani, Alina Vranceanu, all participants of the RonR meetups, Nandan Rao, Silvia Soriano, Jorge Rodriguez, Sebastian Kluesner, Ilya Kashnitsky, Diederik Boertien and Luis Ortiz. We worked on some projects while I was pursuing my PhD and I am happy to have collaborated with you. You all helped me achieve many of my goals directly or indirectly.

I would like to thank the R community for being so open, encouraging and transparent. Many of the opportunities I have today have been possible because I was able to navigate your waters without being discouraged. I hope many people in the future learn to appreciate the importance of open source software. In particular, I would like to thank the Rstudio team for building amazing software, free to anyone. Without your help I would probably be still searching ways to pay for commercial software.

This whole process would not have been possible without the help of my friend Eduardo Sorribas. I have probably messaged Eduardo over five hundred times in these three years with my programming questions. Not once has Eduardo said he would not explain a concept to me. I owe you much more than just a trip to Denmark, you are truly the best. Thank you.

Andrea Casanova has been an instrumental part of this thesis. She has pushed me in many ways to finish it and gave me all of the support that I was lacking. She never complained (well, maybe once or twice) of me spending my weekends working, but even more, she pushed me to spend countless hours in front of the computer to finish it on time. I thank you very much and I can assure you that this would not have been possible without your constant encouragement.

Para todos mis amig@s de España y Dominicana: Michi, el boro, Antonio, Joey, Luima, Mario, Paula, Gennaro, Gabriela, Brian, Paola, Guillermo, Gaby, Francesc, Bori, Adriana, Marta y Dorian los quiero mucho. Me he curado pilas con utede y epero que eto sea un triunfo de todito. Pa lante.

A mi familia Dominicana de la que ya partí hace 4 años, los quiero mucho. A mi familia Española, siempre han ocupado un espacio especial en mi corazón. A mis hermanos Javier, María y Alejandro, este logro es para ustedes. Aprendí una cosa muy valiosa en estos 3 años: la vida gana sentido caminando y no cuando se llega. Sigan caminando. A mi abuela, me curo demasiado contigo. Te quiero muchisimo.

Esta experiencia no hubiera sido posible sin la ayuda de mis padres. En ningún momento me he sentido solo ya que ellos me han dado el apoyo necesario. Siempre me han ayudado a conseguir lo que quiero y siempre me han dado las herramientas que necesito. En momentos de duda me han ayudado a trascender y en todo momento me he sentido con su apoyo. Estos 3 años son más de ustedes que de nadie más.

Abstract

By placing particular attention to the socio-economic dimension, this thesis explores the role of three pillars of an educational system: decentralization, early education and curricular tracking. The first article focuses on an experiment in Mexico that aimed to increase parental empowerment in the school's decision-making. Results show that increased participation produced an increase in cognitive abilities. However, this was mediated by the SES of the student. The second article brings an international perspective by comparing the evolution of the achievement gap in over 15 years and 30 countries. The findings suggest that tracking and vocational enrollment are important explanatory mechanisms of the cross-country variability in achievement gaps in cognitive abilities. The third article revisits the question of whether early education is associated with adult outcomes but concentrating on a particularly vulnerable population in the United States: GED recipients. Results show that participation in early education is associated with greater odds of graduating high school over attaining a GED.

Resum

Aquesta tesi explora, amb una atenció particular a la dimensió socioeconòmica, el rol dels tres pilars d'un sistema educatiu: descentralització, educació infantil i seguiment curricular. El primer article es centra en un experiment a Mèxic que tenia com a objectiu incrementar la participació dels pares en la presa de decisions de l'escola. Els resultats mostren que una major participació produeix un augment de les habilitats cognitives, tot i que aquest augment està condicionat per la SES de l'estudiant. El segon article aporta una perspectiva internacional tot comparant l'evolució del disparitat educativa en 30 països durant més de 15 anys. Les anàlisis suggereixen que el seguiment i la inscripció en formació professional són els mecanismes més importants que expliquen la variabilitat entre països en l'assoliment de les habilitats cognitives. El tercer article revisa la qüestió sobre si l'educació infantil està associada amb els resultats per a adults, tot concentrantse en una població especialment vulnerable als Estats Units: els estudiants que decideixen examinarse als GED. Els resultats mostren que la participació en l'educació infantil està associada amb majors probabilitats de graduarse a l'institut amb una titulació GED.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The goal of this thesis is to explore the notion that institutional features of an educational system can have different impacts based on the socio-economic status of individuals. In other words, do changes in the educational system at different levels benefit more people than others? Moreover, do these institutional characteristics explain some of the actual gaps between socio-economic groups? The three empirical articles presented here attempt to answer these questions by focusing exclusively on three institutional aspects: decentralization of decision-making, curricular and vocational tracking and early education. Even though these three elements are inherently different and operate at different levels, they are all tightly linked.

Decentralization of educational decision making is one piece of the puzzle. First, it allows regional and local stakeholders to take decisions based on local problems with feedback from higher levels of the administration (Bruns et al., 2011). This is directly linked to the management of the early education system, which is often hindered by the lack of freedom to innovate in pedagogical techniques (Fullan and Watson, 2000). This of course, is only one of the dimensions at which both institutions share common ground. In terms of budget, stringent financial accounts between the central administration and local stakeholders can hinder schools from working appropriately (H. A. Patrinos and Fasih, 2009). In terms of curricular freedom, some regions and localities might have particular cultural practices that clash with the central government's curriculum (Caldwell, 2005). In terms of management, public schools are often lacking behind in technological advances and state-of-the-art materials both due to the lack of income as well as the slow bureaucratic interactions between central units and individual schools (Fullan and Watson, 2000).

Decentralized decision making is an alternative to the previous dilemmas. First, it makes sure that local stakeholders can take decisions that they can be good at. They can address local problems tailored to their specific needs (Bruns et al., 2011). Moreover, they can assess specific situations and make quick and meaningful decisions without the overhead cost of bureaucratic transactions. This process is complemented by an accountability framework (Bruns et al., 2011) that aims to limit decision making to topics that are well suited for the local decision maker. Moreover, it emphasizes the notion of responsibility as decision makers have to report the effectiveness of their solutions.

When this is applied to early education, the link becomes more evident. Early education has been found to be one of the most important stages for child development (Dämmrich and Esping-Andersen, 2017; J. Heckman and Kautz, 2013). The cost of giving a bad education due to low quality and/or slow/bureaucratic decision making is high compared to the relative cost of attempting to make up for it at later stages. Early education should be of the highest levels, with the overarching aim of equalizing students to similar levels and giving them a fair start to formal schooling (J. Heckman, 2006).

On the other hand, curricular tracking is in many cases very much connected to the experiences in early education. Early education can be a definitive deal breaker on who goes ahead in terms of achievement at school. This has been found in numerous studies that compare different populations across different countries (Elango et al., 2015; Bauchmuller et al., 2014; Havnes and Mogstad, 2011). Curricular tracking in many cases starts in early adolescence, around 12 and 13 year olds. This makes the impact of early education particularly worrying because the initial bump that early education can give could be a competitive advantage in joining the high level tracks. In other words, the impact of early education can have an important influence on who joins specific tracks.

The interactions between each of these features and several different outcomes has been studied extensively in the past. Decentralization of decision making is a topic mostly studied by applied policy makers mainly because it is a reform widely implemented in international organizations such as the World Bank and national educational systems.

The work of Harry Patrinos is particularly telling in the impact that this reform has on school outcomes. In their book 'Making Schools: Work New Evidence on Accountability Reforms', Barbara Bruns, Deon Filmer and Patrinos go over a policy reform called School-Based Management that aims to streamline the decision making process all the way down to individual schools, which has been shown to empower schools in the decision-making process. (Bruns et al., 2011).

There is ample evidence showing that this reform has improved school outcomes such as repetition rates, dropout rates as well as student test scores (Bruns et al., 2011). There is even international evidence that links school autonomy to higher human capital in over 30 countries (Hanushek et al., 2006). The reform has been implemented by many different countries such as the United States (Borman et al., 2003), Nicaragua (Di Gropello, 2006), Mexico (Shapiro and Trevino, 2004; Gertler et al., 2012), El Salvador (Jimenez and Sawada, 1999) and Brazil (Barros and Mendonca, 1998). The evidence shows that decentralization can improve many of the outcomes mentioned above but there is some mixed evidence in terms of student test scores (Bruns et al., 2011). Some studies show that even after 10 years after the study, there were no recorded changes in test scores (Barros and Mendonca, 1998). Despite this, Bruns et al. (2011) suggest that there is overwhelming evidence in favor applying this type of reform as it is cost-effective and improves educational outcomes for students.

Despite the size and richness of this literature, there has not been any evidence showing whether the impact of this reform has varying effects based on the socio-economic status of the student. This is important, considering that, for example, parent's of children who come from highly educated families might get more involved in the school's decision making and in the interactions with their teachers. This can prompt them to help their children in school chores or simply pressure the school for improved accountability.

Curricular tracking is another branch of literature which has been documented extensively throughout the years (Van de Werfhorst and Mijs, 2010a). Curricular tracking is the organization of grades into several tracks, which students can access based on their performance/willingness. The international evidence on the topic has shown consistently that tracking seems to be associated with increased inequality (Hanushek et al., 2006; Brunello and Checchi, 2007). This is evident in many different countries as well as after/before tracking reforms (Van de Werfhorst and Mijs, 2010a). The work of Hanushek et al. (2006) is one interesting example. They show that the size of the cognitive achievement gap between socio-economic groups seems to be exacerbated after the year of selection into curricular tracking passes by. This is present in other studies, specially in sociology (Bol and Van de Werfhorst, 2013; Van de Werfhorst and Mijs, 2010b). The review of Van de Werfhorst and Mijs (2010a) documents the recent advances in understanding the phenomena and the evidence suggests that inequality increases because the top are improving their performance while the bottom is stagnated. This can be linked to the quality of instruction in different tracks.

Despite these findings, the vast literature on curricular tracking has not been able to confirm whether the sizable country differences in cognitive achievement gaps are actually explained by curricular and vocational tracking. Moreover, there is scarce evidence linking the evolution of this gap to the features of the tracking system.

Both the literature on decentralization and tracking are missing two pieces to the puzzle which complement each other. The literature on decentralization of education still misses whether this institutional change can have different impacts based on the socio-economic status of the student. Conversely, tracking research has not yet fully explained whether the actual socio-economic gaps can be explained by this institutional feature. In the first case, we do not know whether the institutional change promotes inequality by having varying effects. In the second case, we do not know if the actual inequality between socio-economic groups can be linked to the institutional features. The first two articles of this thesis attempt to study the two previous questions.

The literature on early education is, over the previous two, the most widely debated of all. In particular, the topic has gained ground in the past few years when scholars such as James Heckman have built the case in favor of high quality early education as one of the most important interventions in terms of effectiveness. We know from past research that quality early education is associated with improved performance at school and other skills such as motivation and responsibility (Kautz et al., 2014). This research has evolved up to the point that we now have rigorously randomized controlled trials that confirm the effectiveness of early education in improving cognitive and non-cognitive abilities (Schweinhart and Weikart, 1981; Campbell et al., 2002). This evidence shows in particular that children coming from low socio-economic families tend to benefit the most from these interventions (Reid et al., 2001; Elango et al., 2015). J. Heckman (2006) shows that for every dollar spent in a subsidized early education program there is a return ratio of between 7$ to 12$ U.S. dollars.

One of the main criticisms of these programs is that their impact tends to fade out over time. There is evidence showing that early education can have lasting effects up to 20 and 30 years later (Reynolds et al., 2011; Campbell et al., 2002) on dimensions such as as marital stability, reduced welfare dependency, increased income and higher university graduation rates (J. Heckman, 2006). However, for educational attainment, researchers have narrowly concentrated only on graduating high school (for example, Reynolds et al. (2011)). The third article of the thesis attempts to test whether quality early education is associated with a particularly growing population in the United States: GED recipients. The General Educational Development (GED) test are a group of exams which have been developed to assess whether someone has the necessary skills to attain high school-level knowledge. Students which could not graduate high school, but posses the necessary skills to achieve it can request the examination. This test was developed in the 1970's because the dropout rate in the United States was increasing, and the labor market qualifications required more and more that employees have high school diplomas (J. Heckman and Rubinstein, 2001)

This last chapter relates to the previous two articles as it poses an important question which is still under study: do institutional features have a long lasting impact on educational outcomes? The second article touches on this topic briefly as it tests whether tracking can explain the evolution of the achievement gap but it does so in a limited perspective given that tracking doesn't change over time. On the other hand, the third article complements it adequately given that it focuses exclusively on testing the association of early education on adult-measured outcomes.

To summarize, the first article demonstrates that changes in characteristics of the educational system can have different impacts based on the socio-economic status of the individuals experiencing the changes. The second article shows that much of the actual differences in achievement between children from different socio-economic backgrounds can be explained by features of the educational system on an international scale. The third article highlights the fact that these institutional features can have strong associations to outcomes measured in adulthood, something that can leave a scarring effect if not dealt with properly.

The importance of this problem lays in the fact that some children might fare much worse than others depending on their specific situation. More concretely, children coming from households with violent environments, children receiving little to no education and children coming from families that cannot secure a quality level of education are at a greater disadvantage than other children at improving their future opportunities (J. Heckman and Raut, 2016). Research has shown that a bad education can negatively affect indicators ranging from income, personal relationship, propensity to crime and improved opportunities in adulthood (J. Heckman, 2006). Coped with the fact that receiving quality education is one of the fundamental articles in the Universal Declaration of Human Rights (Assembly, 1948), making sure that all children are taken care of is an important task for researchers and policymakers.

1.1 Structure of the thesis

The second chapter, titled 'How do we improve cognitive abilities? An unequal intervention of the AGE program in Mexico', looks to study the impact of decentralized decision making at the school level with an experiment carried out in Mexico by the World Bank. The schools participating in the treatment group received US$1200 in quarterly payments through the parent's association to increase parental involvement and allow them to invest in school infrastructure and study materials for the students. The theory and current evidence predicts that it will indirectly improve student outcomes such as repetition and dropout rates. In contrast, the *control* schools only received US$600 under the same conditions. The main objective of the chapter is to evaluate whether the treatment improved test scores, something which has had mixed effects in the literature, and whether this improvement is mediated by the socio-economic status of the student.

The results show that the differences between treated and control groups for the first two years are negligible, which was expected as the treatment needs time to fade in, but for the last two years the treated units had higher probabilities of about $\sim$ 13% - 17% of scoring higher marks than the control units. Moreover, predictions show that this effect is strongly moderated by the SES origin of the student where students in the treatment group coming from high SES background had a privileged "boosting" effect relative to the same treated low SES students. These results are equally valid in Mathematics and Natural Sciences, showing robustness in replication.

Titled 'Does curricular tracking explain global SES gaps? an international comparison of the SES achievement gaps from 2000 to 2015', the third chapter of the thesis concentrates on curricular tracking. This article aims to investigate whether there are discernible patterns in the evolution of the cognitive achievement gap from a comparative perspective. Moreover, the article investigates whether the between-country variation can be explained by the degree of curricular and vocational tracking. The results show that there is considerable variation in the way in which the gap is evolving, with the U.S. and Germany closing their achievement gap at a rate of about 50% and 30% in the last 15 years while France is widening at a similar rate. The findings suggest that curricular tracking and vocational enrollment explain about 40% of variance in the achievement gap between countries and show that the relationship is conditioned by a strong interaction. Low curricular tracking is associated with a small achievement gap, whereas high levels of curricular tracking is associated with wide achievement gaps. However, once tracking is coupled with high vocational enrollment this can remedy the potential adverse effects and reduce the gap by over .6 standard deviations.

The fourth chapter, titled 'The long-term relationship between early education and educational trajectories: the case for GED recipients', looks to understand whether quality early education is associated with improved chances of graduating high school over following a GED diploma and the improved chances of graduating a GED over being a dropout. Exploiting the detailed care history of children in more than 17 years of data from the Panel Study of Income Dynamics (PSID), estimates show that being in preschool relative to being cared for by one's own parents is associated with 25% higher odds of attaining a high school degree relative to having a GED/No qualifications. These results are also replicated when comparing the odds against graduating high school vs GED and attaining a GED vs No qualifications with an odds increase of 11% and 60% respectively. These results highlight the unique role of quality education when compared to other types of care.

Finally, chapter 5 summarizes the contribution and main findings of chapters 2-4. I then discuss the limitations of each study and offer some possible directions for future research.

Bibliography

Assembly, UN General (1948). "Universal declaration of human rights." In: *UN General Assembly*.

Barros, R Paes de and Rosane Mendonca (1998). "The impact of three institutional innovations in Brazilian education." In: *Organization matters: Agency problems in health and education in Latin America*.

Borman, Geoffrey D, Gina M Hewes, Laura T Overman, and Shelly Brown (2003). "Comprehensive school reform and achievement: A meta-analysis." In: *Review of educational research* 73.2, pp. 125–230.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos (Feb. 11, 2011). *Making Schools Work: New Evidence on Accountability Reforms*. World Bank Publishing. 251 pp. ISBN: 0821386794. URL: https://www.ebook.de/de/product/14536372/barbara_bruns_deon_filmer_harry_anthony_patrinos_making_schools_work_new_evidence_on_accountability_reforms.html.

Caldwell, Brian John (2005). *School-based management*. Vol. 3. International Institute for Educational Planning Perth, Australia.

Di Gropello, Emanuela (2006). *A comparative analysis of school-based management in Central America*. 72. World Bank Publications.

Fullan, Michael and Nancy Watson (Dec. 2000). "School-Based Management: Reconceptualizing to Improve Learning Outcomes." In: *School Effectiveness and School Improvement* 11.4, pp. 453–473. DOI: 10.1076/sesi.11.4.453.3561.

Heckman, James (2006). "Skill formation and the economics of investing in disadvantaged children." In: *Science* 312.5782, pp. 1900–1902.

Heckman, James and Lakshmi K Raut (2016). "Intergenerational long-term effects of preschool-structural estimates from a discrete dynamic programming model." In: *Journal of econometrics* 191.1, pp. 164–175.

Heckman, James and Yona Rubinstein (2001). "The importance of noncognitive skills: Lessons from the GED testing program." In: *The American Economic Review* 91.2, pp. 145–149.

Jimenez, Emmanuel and Yasuyuki Sawada (1999). "Do community-managed schools work? An evaluation of El Salvador's EDUCO program." In: *The world bank economic review* 13.3, pp. 415–441.

Patrinos, Harry Anthony and Tazeen Fasih (2009). *Decentralized decision-making in schools: The theory and evidence on school-based management*. World Bank Publications.

Reynolds, Arthur J, Judy A tchapterle, Suh-Ruu Ou, Irma A Arteaga, and Barry AB White (2011). "School-based early childhood education and age-28 well-being: Effects by timing, dosage, and subgroups." In: *Science* 333.6040, pp. 360–364.

Van de Werfhorst, Herman G and Jonathan JB Mijs (2010a). "Achievement inequality and the institutional structure of educational systems: A comparative perspective." In: *Annual review of sociology* 36, pp. 407–428.

CHAPTER 2

How do we improve cognitive abilities? An unequal intervention of the AGE program in Mexico

Abstract: This chapter evaluates a World Bank experiment that aims to make parents involved in the decision making process of schools. The schools participating in the treatment group received US$1200 in quarterly payments through the parent's association to increase parental involvement and allow them to invest in school infrastructure and study materials for the students. In contrast, the *control* schools only received US$600 under the same conditions. Using data from an experimental design, the chapter explores whether the randomized monetary incentive to get parents involved in the decision-making of the school in four rural provinces in Mexico improved test scores in Mathematics. Compared to the baseline year 2007, the chapter estimates a phased-in treatment effect for years 2008, 2009 and 2010 using a bayesian ordinal model. The results show that the differences between treated and control groups for 2007 and 2008 are negligible, which was expected as the treatment needs time to fade in, but for the last two years the model estimates that treated units have increased odds of about $\sim$ 13% - 17% of scoring higher test scores than the control units. These results are robust even when controlling for the SES of the student and the type of school they are enrolled in. Moreover, the model's predictions show that this effect is strongly moderated by the SES origin of the student where students in the treatment group coming from high SES background had a privileged "boosting" effect relative to the same treated low SES students. These results hold for 4th graders in Mathematics and Natural Sciences but disappear for 5th graders in the same schools.

2.1 Introduction

The World Bank has long been interested in implementing high quality interventions that allow countries to decentralize their educational decision making process. More specifically, school-based management (a World Bank educational policy designed to decentralize educational decision making from the government to school level) has been implemented in several countries with the aim of improving school management and student performance.

Mexico has been a country known for their efforts in decentralizing their educational decision making to regional and local stakeholders. This is evident with programs such as *Apoyo a la Gestión Escolar* (AGE) which raised incentives for parents to take part in the decision making process of schools. School-based management (SBM) interventions have been implemented in several countries around the world and have been evaluated many times. However, most research suggests that it has a mixed effect in its impact in cognitive abilities. Some well known studies show that it has had positive effects while others show that it has little to no effect (Bruns et al., 2011). However, the same literature highlights that evaluating SBM interventions is very difficult because it is very hard to design an experiment for this type of treatment. In fact, Bruns et al. (2011) analyze most SBM research using very high standards and only about 5 experiments out of 22 had reliable and valid randomization criteria.

To raise attention on the validity and effectiveness of this policy, this chapter studies an experiment conducted in Mexico from 2007-2010 by the World Bank. The schools involved were all participating in a Mexican program named AGE which gave schools US$600 in quarterly payments through the parents association. The goal of this grant is to get parents involved in school affairs, prompt them to take part in school decisions and get involved in children's school activities. This grant has been shown to increase parental participation in the decision making process of schools (Gertler et al., 2012). In the current experiment, World Bank investigators increased the grant to US$1200 for the treated schools to test whether parental involvement in school affairs increased even further. They expected this extra income to have an effect in lowering dropout rates, repetition rates and increase test scores at the school-level. This chapter brings forward two questions with individual level data: did the AGE program influence cognitive skills and did this effect fare differently for different SES groups?

The present study has two main strengths: its randomized design and its methodology. Educational interventions are often hard to study in a randomized fashion. In fact, a handful of school-based management evaluations have had something close to a randomized design (Bruns et al., 2011). Having said that, this is one of the few evaluations of school-based management under a rigorous experimental randomized design. This chapter takes advantage of this to make causal claims of the effect of the program on tests in Mathematics, Natural Sciences and Spanish.

The chapter uses a fully bayesian multilevel ordinal model to model the test scores of children during the experiment accounting for the clustering of students within schools and between years. This is beneficial because it allows to counteract the role of measurement error and uncertainty by specifying a prior distribution of possible values for the treatment effect on test scores. This prior distribution has particular importance in the context of this study given that the values that generate the prior distribution are based on theoretical and empirical considerations. Additionally, the advantages of this adjustment are evident once we consider that test scores are plagued by measurement error and uncertainty due to the fact that assigning a single mark to a student is done through subjective decision making.

Using 2007 as a baseline relative to the years 2008, 2009 and 2010, the results show that treated and control groups are practically the same in terms of test scores in the first two years but for the last two years the model estimates that treated units have higher odds of about $\sim$ 13% - 17% of scoring higher test scores than the control units. The predictions of the model show that this effect interacts with the SES origin of the student where students in the treatment group coming from high SES background had a greater effect relative to the same treated low SES students. These results hold for 4th graders in Mathematics, Natural Sciences and Spanish but disappear for 5th graders in the same schools.

The reasons why this might have happened are numerous and the paper speculates two possible explanations for the results. First, the mechanisms through which cognitive abilities increased were not directly affected by parent's heightened participation in school activities. This might have had positive relevance in parent-teacher relations but not particularly on test performance or cognitively enriching activities. However, test scores could have have increased due to the indirect spillover effect of parental participation towards more quality time with the students in tasks such as homework or reading. This is a possible explanation considering that parents also received training on how to help their children in academic activities as a broader initiative of the ongoing program. Note that both treatment and control groups received the same training but perhaps being more involved at school affairs through the parent's association increased overall parent-child interactions.

The chapter also explores whether the treatment effect was moderated by the SES origin of the student. The results show that for treated schools, high SES students benefited more from the treatment as they had higher chances of scoring in the top marks than low SES students. That is, receiving exactly the same treatment, high SES students saw larger probability gains than lower students in terms of scoring in the top groups. More concretely, treated high SES students had an odds increase of higher grades of about 23%-26% and the same low SES students in the same years saw an increase in odds of about 14%-13%. Conversely, high SES students saw smaller probabilities of scoring in the bottom marks than their low SES counterparts. As all students were receiving the same treatment and were very similar in over 190 characteristics at the school-level, the likely explanation that I provide is related to the interactions with their parents.

Secondly, there is evidence that parental participation in the parents association increased in 4th grade with evidence from the head of the association as well as from the principals of the schools (Gertler et al., 2012). However, this is not clear for 5th graders. Further research should attempt to confirm this as it would explain not only the lack of effect for 5th graders but the effectiveness of increasing parental participation in school affairs.

The chapter is outlined as follows. The first section introduces relevant research on education decentralization and highlights the interventions which have had positive as well as mixed effects. It also focuses on the literature on parental involvement at home, as the proposed mechanism at play is thought to be related to parent-child interactions at home. The second section describes the experiment and hypothesis of the study. The third section introduces the methodology and data while the last two sections present the descriptives of the data, the results and the conclusion.

2.2 Literature Review

Decentralizing educational systems is a trend that has been gaining ground for the past decade. The OECD, through the Program for International Student Assessment (PISA), has found that those countries who have allowed teachers and principals to have an ample range of decision making, have decreased differences in test scores between schools, and outperform countries with less decentralization (OECD, 2013).

A benchmark example is the Finnish case. Sahlberg (2007), from the World Bank, proposes that the main success of the Finnish system is that it grants complete responsibility to teachers and principals to make their decisions in a vast array of educational aspects. When that happens, schools start to offer different educational experiences, increasing competition among schools, while still maintaining quality education for everyone (Sahlberg, 2007).

To put it differently, starting from a simple demand-supply model, as the product of education has more quality, more emphasis is placed on diversification of the product and thus schools become more encouraged to exceed each other. However, as Sahlberg (2007) emphasizes, this worked seamlessly only because everyone, regardless of income or social background, had free access to all schools. The schools were different, but all offered virtually the same level of quality. Hanushek and Woessmann (2007) have also uncovered that learning encouragement can be driven by school institutional factors. Allowing teachers to decide how to teach the curriculum and to choose their own personalized pedagogical techniques will promote and encourage a sense of responsibility in them. When this mechanism is at play, then it could likely strengthen the accountability framework given that teachers will feel responsible for their outcomes.

School based management (SBM from now on) is an educational policy that has been implemented in developed as well as developing countries to decentralize educational systems. The objective of the policy is to decentralize decision making to the state, regional or school level, depending on the specific goal of the reformers (Bruns et al., 2011). The logic runs in this manner: educational systems are organisms which are far too complex to be efficiently handled in a centralized fashion. Producing educational quality and distributing it equally is almost impossible when decision making is held solely by a single actor (King and Guerra, 2005). By transferring responsibility to local stakeholders, who know much more about what their community needs, it then becomes an efficient means of making decisions. After transferring authority to the schools, a council composed of parents, teachers and principals is to be created on which consensus is reached on the priorities of that school. To prevent undesirable results, the council is allowed a limited set of responsibilities. According to Bruns et al. (2011), the decision making that is allowed to teachers, parents or principals is limited to: budget allocations, deciding on whether to dismiss or keep teachers, administrative decisions such as buying textbooks, adjusting the curriculum, rebuilding or improving school infrastructure and evaluating teacher and student outcomes. However, some contend that this type of decentralization is still ineffective given that key aspects of the decision making process, like how the national curriculum is planned, are still unmovable from the central organism (Fullan and Watson, 2000). Furthermore, another issue is also whether parents are prepared to make such decisions. It could well be that they do not possess the necessary skills to make important educational decisions and thus keeping some national competencies at the government level is the right decisions.

In addition to the transfer of responsibilities, SBM implements an accountability framework. Those in charge of decision making are held accountable for their decisions but also those in charge of the decision making make those below them accountable. Leithwood and Menzies (1998) find, after studying hundreds of cases of SBM, four types of SBM: administrative control, professional control, community control and balanced control. The first one implies that principals are in sole ownership of decisions. The second one allows only teachers to be in charge of the decision making. The third just includes parents and community stakeholders and lastly, the fourth one is a balanced control which involves parents and teachers. However, the general program has evolved broadly and nowadays there are SBM programs which involve school directors, parents, teachers and even students themselves. SBM has been implemented in many countries, among which are the U.S. (D. L. Taylor and Teddlie, 1992), Nicaragua (King and Guerra, 2005), Guatemala, Honduras and El Salvador (Di Gropello, 2006). All of these authors have found that the program has had an impact in lowering dropout rates, increasing student enrollment and student achievement. On this last outcome, however, there's mixed evidence. Findings by Barros and Mendonca (1998) show that achievement did not increase in Brazil even after measuring it 11 years later. As a matter of fact, Gertler et al. (2012) note that SBM evaluations should be interpreted with great care, as most of them have not had rigorous experimental designs. In contrast, they find that the results of those few studies with strong experimental designs are ambiguous, ranging from strong improvements of test scores to statistically insignificant improvements. In short, SBM proves to have consistent and significant positive results on intermediate school outcomes such as repetition rates, dropout rates and graduation rates. When student achievement is concerned, however, mixed results blur the landscape.

Another concern that hampers the effectiveness of SBM is whether parents are ready to provide a good education to their children in their family environment. Intermediate school outcomes can be easily altered by being more attentive on their children and their school matters, but other skills, such as cognitive and non-cognitive ones, might need a more thorough approach. They need nurture in specific high-quality activities like reading and discussions. It is important to raise this concern considering that most SBM reforms have ignored the issue and do not give it the necessary attention (For a review of almost all SBM evaluations and their outcomes of interest, see Bruns et al., 2011).

The benefits of SBM depend greatly on the type of the reform. In principle, the potential benefits of SBM can be better management of school affairs, be it economically, administrative and extracurricular. Community involvement in decision making can also drive to improvements in the quality of the classes and the demands that are placed on the teachers. Lastly, the most studied outcome and the one with the most research is student achievement measured as test scores. If teachers are developing new pedagogical techniques, as well as improving their skills due to accountability feedback, then we should expect a change in a wide variety of outcomes, ranging from cognitive skills to even students' school perception. If that is combined with the role of parents, then we should expect an increase in student performance in school.

In theory, SBM makes sense. However, when parent's are not trained to help their children to achieve their fullest potential, then the result of an SBM reform might be very small or inexistent. For example, researchers have pointed out that for the Head Start program (Reid et al., 2001), teachers are being trained to better educate the children but not enough instruction is given to the parents of the children. The most successful SBM program implemented so far is El Salvador's EDUCO program, which granted a high degree of authority to parents and trained them in school management as well as on developing their capacity to help their children with their homework. An in depth analysis by Jimenez and Sawada (1999) attributes the success of the program to the high parental participation. In a similar line, most of the work of by James Heckman and his colleagues (J. Heckman and Kautz, 2013; J. Heckman, 2006; Cunha et al., 2006) suggests that all of the landmark early intervention programs, such as the Perry Preschool program and the Abecedarian Program, would not have yielded their strong and robust return rates if parents had not participated in weekly 90 minute sessions on how to raise their children [1]. In fact, it is not only about instructing parents but the frequency and importance that is placed on this education. Fullan and Watson (2000) review the most successful SBM programs and find several common denominators among which is that the community agrees unanimously that education is one of the most important tenets of success and are willing to embark in the process of building the community around education. All in all, there should be a community-wide driven reform where all agents agree on the direction and tasks of the SBM policy.

Mexico is one country that has adopted SBM and decentralizing policies as a remedy to their unequal educational system. From 1991 to 2003, the Mexican government collaborated with the World Bank and the Inter-American Bank to improve teacher education and increase enrollment of disadvantaged children into schools as part of a bigger program named CONAFE. Acevedo and Paqueo (2003) found that this collaboration yielded positive results, as it increased enrollment of indigenous children in the program, as well as improved students test scores across the economically disadvantaged population. In the same line, the Mexican government organized and implemented the project *Apoyo a la Gestión Escolar* (AGE). The program provides low performing schools (usually indigenous schools or schools located in marginalized communities) with the equivalent of $600 American dollars to the school budget through the parents association in quarterly payments. This can be seen as a form of School Based Management given that parents

[1] The parent weekly training time varies per intervention, but the Perry Preschool program and the Abecedarian had 90 minute sessions per week. See Schweinhart and Weikart (1981) and Campbell et al. (2002) for a review of the experiments

are now expected to get involved in school activities and contribute to decision making (H. A. Patrinos and Fasih, 2009). In addition to the budgetary increase, participating parents received training on how to make school decisions and how to help their children with their homework. This is considered a big step for the educational landscape of less developed Mexican regions considering that, generally speaking, parents are scarcely involved in school matters (Gertler et al., 2012). The parents association is supposed to outline a working plan of school priorities with the teachers and principals, and plan a budget to adhere to for the rest of the year. Parents are allowed to allocate the budget to small civil works and infrastructure improvements as well as supplying the schools with the necessary materials. Contrary to the EDUCO program from El Salvador, on which parents were given the power to hire, monitor and dismiss teachers, this program only allowed parents to make budgetary decisions and plan school activities.

However, there are concerns on whether the parental training is indeed effective on the parents of the students. As it is likely that these parents are poorly educated, low income earners and might not have that much time availability, the training that was given to them was not enough to have a significant improvement in the child's family environment [2].

In addition to the training of the parents, teachers education was not very thorough either. Anderson et al. (2003) mentions several components that made one SBM reform effective: classroom-based in service teacher training, strengthening the capacity of the teachers association to have constant developments, management training for head teachers and, lastly, parental involvement and financial support at the school level. A separate branch of research, dedicated to measuring to what extent teacher training and teacher quality influences student achievement, has found similar results and concludes that teacher experience and content focused development are strong predictors of student achievement (Harris and Sass, 2011).

From all of these components, the AGE implemented in Mexico only has parental involvement, financial support and some type of development for teachers, as they are getting feedback from the school council. But the important strength of AGE is that

[2] However, it must also be acknowledged that, as J. Heckman and Kautz (2013) wittily point out, academic research looks at disadvantaged children as those coming from poorly educated and low income families when they should instead pay attention to the adverseness of the environment on which they were raised

teachers and parents are interacting much more than before (Gertler et al., 2012), and as we know from other SBM reviews, this has been linked to improved student outcomes (Fullan and Watson, 2000).

2.2.1 Parental, school and community involvement and it's relationship to student achievement

School-home linkage in terms of support for students has been well researched and offers important lessons on how to understand what works for SBM interventions. In particular, there's been evidence that already links the benefits of the combination between the home, school and community environments at improving the academic achievement of students.

The work of J. L. Epstein and Sheldon (2002) presents evidence that suggests that community partnerships and school-home interactions can reduce daily absenteeism among students. Moreover, it suggests that it not only reduces daily absenteeism but also chronic absenteeism, a more severe phenomena. This piece of research served as an important finding in the literature given that it showed empirically the importance of community involvement in a student's well being. In fact, these results explain why a lot of the SBM literature has focused on these intermediate indicators, including the work of Gertler et al. (2012) which focused on absenteeism and drop out rates in the same SBM intervention used in this paper.

However, this paper will explore it's relationship relative to student achievement and in that aspect there's also been a lot of evidence. Focusing on student's mathematics achievement, the main focus of this paper, Sheldon and J. L. Epstein (2005) explore the notion that family and community involvement can improve student performance using longitudinal data for elementary and secondary schools. Their findings show that increasing family support at home for their children, is associated with improvements in test scores. More importantly, it improves the chances of scoring above the proficiency levels set in Mathematics, the minimum standard in the standardized test in their study. These findings have been also replicated in other settings including second generation immigrants and children from all ages such as Galindo and Sheldon (2012), Altschul (2011) and Xu et al. (2010).

The concept of school-home interactions is a complex one and has been very controversial because there doesn't seem to be a clear cut explanation as to what the role of parents and schools should be in educating children (J. L. Epstein et al., 2018). More in particular, the dilemma comes from whether home and school environments have complementary roles, separate roles or sequential roles. The handbook by J. L. Epstein et al. (2018) has attempted to counteract the mixed findings by providing a theoretical framework that explains how children benefit from home-school interactions through six types of involvement:

1. Parental involvement that establishes home environment's that support children's development
2. Design effective forms of school-to-home and home-to-school communications about school programs and children's progress
3. Recruit and organize parent's help and support
4. Provide information and ideas to families about how to help students at home with homework and other curriculum-related activities, decisions, and planning
5. Include parents in school decisions, developing parent leaders and representatives
6. Identify and integrate resources and services from the community to strengthen school programs, family practices, and student learning and development

These six theoretical pillars are not completely independent and sometimes play competing and endogenous roles, somethings discussed in depth in J. L. Epstein et al. (2018) and J. Epstein (2018). These six types of involvement are particularly relevant in the context of the SBM interventions because most, if not all, are present in the AGE program. In particular, the SBM program under study seeked to get parent's involved in the decision making (involvement 5 and 6) and get parents more involved at home (involvement 1 and 2) through parental training (involvement 4). More specifically, the parental training was implemented with the aim of improving the student's development. However, as J. L. Epstein et al. (2018) clearly states, in order for this step to be effective, there needs to be parental competence and home availability, something which precedes the real objective: having knowledge on how to set up an appropriate home environment that fosters development. This makes sense theoretically from the family's perspective yet the work of Sheridan et al. (2019) provides among the best up to date evidence that from the school's point of view, teachers also need to be trained to effectively communicate with parents. This point seems to highlight that the competence of the parents, as well as the teachers, as they are both important for improving a child's development. As argued before in the description of the AGE program, the training given to parents in this specific intervention does not seem to be of very high quality given it's low frequency and ubiquity (in the context of this AGE program, this is described in detail in the next section), so it could be the case that it did not have an important effect on the development of the home environment.

The work of Patall et al. (2008) offers an empirical perspective when it comes to parental training on children's development at home. In their meta-analysis of over 20 studies, they find that manipulating parental training aimed at helping their children with their homework had several benefits among which is improved academic performance. This finding fits very well within J. L. Epstein et al. (2018) theoretical framework as it states that to have an effective type of involvement, parent's need information and ideas on how to help children with their homework and their curriculum-related ideas. Moreover, using more recent evidence from children aged 0 to 18, Boonk et al. (2018) review 75 studies from the years 2002-2017 and find some clear cut correlations regarding parental involvement. The authors criticize the somewhat abstract concept of parental involvement sometimes used in the literature and look to provide specific types of involvement that shed some light on the construct of parental involvement. In particular, they find that reading at home, holding high expectations for the children's academic achievement and parental encouragement and support for learning are important predictors of academic achievement. The first and third are of particular relevance to SBM interventions because the parental training aimed at improvement child-parent interactions specifically concentrated on training parents to get involved in the development of a supportive home environment which included reading and support for learning at home.

All of these findings suggest that a somewhat positive environment at home set by the parents helps improve the student achievement across all homes. Yet this relationship does not operate in isolation and could be moderated by other variables, such as the initial cognitive skills of the child (Rogers et al., 2009; S. Phillipson and S. N. Phillipson, 2012). However, the most discussed variable by far is the socio-economic status of family, sometimes measured as income, other times as the education of the parents and sometimes has been linked to ethnicity in certain settings. For example, using ethnicity in the United States, not a perfect measure of SES but at least informative, the findings by Fan et al. (2012) offers an interesting contrast to the aforementioned ideas. Their research shows that although parental advising to children has a somewhat positive effect when pooled together, it has mixed associations when disaggregated by ethnicity. For example, they find that parental advising worked positively on Hispanic children towards academic self-efficacy but negatively associated with the Asian American sample. These results serve to highlight that there might be a heterogeneous relationship once disaggregated by other variables, in this case ethnicity.

In contrast, the work of Wang and Sheikh-Khalil (2014) finds that the association between parental involvement and student achievement did not vary significantly between ethnicity groups but it did so between SES groups. Their results show that low SES families tend to be *less involved both at school and at homework related activities at home*. This is contrary to the results of the high SES group, which boosted high levels of involvement at home and did employ more home-based activities related to student engagement. Their results offer some evidence which shows that when low SES families provide structured and rich home environment aimed at improving academic engagement, low SES students also benefit a lot from these interventions. Having said that, there is overwhelming evidence pointing out to the lack of involvement at home from low SES families. In particular, the work of N. E. Hill and L. C. Taylor (2004) and N. E. Hill et al. (2004) suggests that lower income parents tend to be less involved in their children's education even if they have the same high aspirations for their children as high SES families. They point that the explanation may come from the fact that they feel they are less effective at changing school achievement by being involved in their children's academic and home environment.

This piece of evidence is crucial for understanding the effectiveness of school-based management. If we assume that low SES parents did not get as involved as high SES families, then the importance of the intervention could be affected by the lack of help the low SES families can actually give. More seriously for the case of SBM interventions, low SES families might not have the time, resources, or information to have greater involvement in the development of the academic home environment. As Jordan and Plank (2000) put it, this could possibly be related to the fact that low SES groups have a perceived lack of competence in their power to improve their children's educational outcomes.

SBM interventions are usually placed in a context of developing countries, specifically in vulnerable populations which clearly fit the profile described above: scarce time to participate in such important developmental situations, lack of resources to dedicate to developing a rich academic environment at home and equally important, a lack of information on how to improve their interactions with their children.

2.2.2 Design of the experiment

Despite all of the criticisms of the AGE program, the program implemented in Mexico is an important initiative from the Mexican government. Interventions such as this one should be fostered and measured more frequently to improve the quality of education of a country. That is why in early 2007, the World Bank, together with the Ministry of Education of Mexico, implemented an experiment in which they selected 250 random schools that were already participating in the AGE program and granted 125 of those schools an extra $600 dollars altogether. This was accompanied with parental training to better help their children. I describe the experiment next.

The World Bank, in collaboration with the Ministry of Education of Mexico, chose to design an experiment to test if an increase in the budget of the schools through the parents association would increase intermediate school outcomes and test scores. After obtaining the complete list of all schools participating in the AGE program in Chiapas, Guerrero, Puebla and Yucatán [3], the World Bank investigators excluded boarding schools, schools that did not participate in ENLACE 2006 [4], and schools that joined the urban school-based management program called PEC. Although this does not bias the randomization procedure in any way, it does mean that the schools that participated in the analysis are probably representing the average schools in the regions rather than the well performing (boarding schools) or the worst performers (school that did not participate in national tests).

In the 'Design of the experiment' documentation found on the World Bank's website

[3]These were the provinces with the highest population of indigenous people

[4]A national test given in all Mexico every year

(see http://microdata.worldbank.org/index.php/catalog/1039), this excerpt is all they discuss: "From the universe of AGEs schools in the four states we excluded boarding schools, schools that did not participate in ENLACE 2006, and schools that joined the mostly urban school-based management program (PEC)". It is perhaps informative that schools in Oaxaca (a region they did not include) did not participate in ENLACE 2006 because there were teacher strikes and this region is the one with the highest proportion of indigenous population, something that highly correlates with poor schools.

After the exclusion of these schools, 250 were randomly chosen to participate in the design. 125 of the schools were randomly assigned to the treatment and the remaining 125 to the control group. Having said that, both selection of schools and assignment of the treatment comply with the randomization assumption. Looking at the distribution of general and indigenous school within the treatment and control groups, the team that designed the experiment also confirmed that this distribution was very much representative of the actual distribution of schools (Gertler et al., 2012) [5].

All 250 schools received US$600 dollars as they were participating in AGE. The treatment design increased the budget to US$1200 dollars for the parents association of the selected schools in quarterly payments and left the control group with the baseline US$600 dollars; schools are subject to random audits to make sure the money is being spent correctly.

The experiment was implemented in 3rd, 4th and 5th graders and the treatment was equally applied across grades, i.e, same amount of money, training and liberty to take decisions. All parents were welcomed to participate in the parent's association and no information mentions whether there are quotas or any selection mechanism into the parent's association. Although there is no way of knowing precisely how many parents got involved in the AGE program thanks to the extra funding, previous research on AGE has documented how much parents got involved in the decision-making process once 'Apoyo a la Gestion Escolar' got implemented in their school. H. Patrinos (2006) interviewed many of the principals of the participating schools in all of the provinces in this study and they all said that parental participation increased throughout the school. Moreover, the direct evidence from the principal investigators found that the parent association did increase in

[5] The reader can access all of the experiment design documentation at the World Bank's website: http://microdata.worldbank.org/index.php/catalog/1039

number of participants, according to the head of the parent's association (Gertler et al., 2012). All parents which participated in the parental association received the training and in the committees that managed the extra funding. H. Patrinos (2006) found that most parents got interested in their child's performance at school and got motivated to follow their child's progress more closely. This prompted them to have increased communication with other parents, teachers and principals.

Students received questionnaires that asked them about their experiences at home, at school, scores on their latest tests and their demographic information. Parents, teachers and principals were also given a questionnaire which asked questions about the school environment and the household environment. In particular, the principals were asked whether they saw increase in parental participation in the parent's association. As per the parent's, no information was asked on how they spent their budget as the budget was strictly limited to a few tasks that were monitored by the school, teachers and parent's themselves.

In all AGE schools (and in the experiment), the parents association is supposed to outline a working plan of school priorities with the teachers and principals, and plan a budget to adhere to for the rest of the year. Parents are allowed to allocate the budget to small civil works and infrastructure improvements as well as supplying the schools with the necessary materials.

There is the concern that some parents might influence the budget allocation to children from some specific groups but this seems unlikely. The funding was given to the parent's association as a whole and the funding allocation was voted by the parent's association in consultation with the teachers and the principal. Moreover, there were random audit checks from the government making sure that the budget was adhered to what was supposed to. This gives greater support to the idea that funding was not concentrated on a given group (for example, high SES children).

Just as with the initial AGE program, the parents receive training on school management as well as on how to help their children with their homework. The frequency with which these are conducted is around every three weeks for a 1 hour meeting with the parent's association. From a child's perspective, the treatment they were receiving was in the form of more involvement from their parents in school, as well as on leisurely activities.

According to the objective of the principal investigators, more involvement from parents in school affairs would be reflected in more and better involvement by the parents with their children (Gertler et al., 2012). To clarify, the treatment here is only an increase in US$600 dollars for the treatment group; both groups were already receiving training for parents on how to help with their children. The program lasted from 2008 to 2010 with a baseline survey in 2007, with the surveys being carried out at the end of each school year. In the last two years of the experiment the team of principal investigators added two new control groups for further comparison. The first group of schools did not receive any money but received the same parental training that other schools were receiving and the second group was not participating in AGE nor receiving any parental training. In other words, a pure control. Although the inclusion of these two new groups did not have an impact on the quality of the data (all schools were still interviewed with the same questionnaires and same regularity), it did pose a financial burden. The principal investigators decided then to lower the sample size in each school rather than drop any school because of money constraints. This decision leaves all schools representative yet the power of the analysis should in theory be reduced due to decrease of the sample size.

Although these two groups seem very interesting for comparison, they are only available for the last two waves, which are actually very close together [6] so there is not enough time to properly estimate any treatment effects. In addition to this, there are no baseline observations for comparison in any of these schools as we have for the initial treated and control groups. For these reasons, these two extra groups are excluded from the analysis.

To ensure school homogeneity between treatment and control groups, the World Bank team compared 188 school characteristics for each grade for each school for three years using census data from the Ministry of Education, and found that 91% of them were similar (Gertler et al., 2012). This evidence strongly supports the assumption of balance and homogeneity between treated and control groups.

This is among the few randomized experiment of SBM and it complies with all assumptions to reach a causal conclusion of the program. However, some limitations should be noted. Since there is no panel data at the individual level, no causal claims can be made individually but rather at more aggregate levels. Despite that, there are no reasons

[6]The 2009 survey was moved to 2010 due to teacher's strike. The field work for the last two years were actually only 5-6 months apart.

to believe that students, for example, in 2009 are different from students from 2007.

The aim of this chapter is to study whether the AGE experiment described above (1) improved mathematics test scores for the treated schools and (2) whether the effect of the treatment differed by the SES origin of the child. I hypothesize that the AGE program might have had an unequal effect between families. This is because after reading the documentation of the AGE program and comparing this to previous research of parental training, it does not seem that the quality of the training was of high standards. First, the composition of this sample is of parents from very low educational qualifications. Most parents in the sample come from a farming or agricultural background and very little have completed university degrees. The training was done every three weeks for only one hour and the training was given to the parents association as a whole with no specific one-to-one interactions nor home visits.

This bears a stark contrast once we compare it to the frequency, of for example, more rigorous training such as the Early-Head Start Visiting program. Head Start has been found to be effective, specially when it's component of home based visitation has been present (for example, in Raikes et al. (2006) which found that the number of home visits, duration in the program, length of visits and intensity of service were among the most important quality components of the visit). It's effectiveness has been concretely linked to the EHS home-based services which includes (1) weekly 90-minute home visits and (2) two group socialization activities per month for parents and their children. Home visitors are required to have knowledge and experience in (1) child development and early childhood education; (2) principles of child health, safety and nutrition; (3) adult learning principles; and (4) family dynamics (taken from the official documentation at `https://eclkc.ohs.acf.hhs.gov/policy/45-cfr-chap-xiii/1302-22-home-based-option`).

But even more, the quality of the Head Start home visits program pales even more compared to other interventions such as the Abecedarian Project and the Perry Preschool program, which fostered even more personalized training with more focused and personalized training sessions. Having said all of that, the quality of training in the AGE training programs does not seem to have high standards when compared to other training programs. More concretely, the one hour training session every three weeks, considering the already low educational qualifications of the parents, seems of low standards when compared to other initiatives that aimed also to train parents on how to help their children.

It should be noted that most of the interventions mentioned above are aimed at children younger than the children in the AGE program. Having said that, it still seems of very low frequency, low intensity and scarcely focused on the child when contrasted to other interventions.

Considering that, there are reasons to believe that better educated parents might have benefited much more than other families in this intervention. This is the case given that better educated parents have the necessary skills to be more involved in the decision making and in the trainings.

It is also important to highlight that given the age of children, in order for the experiment to take effect, more time is needed than traditional early education interventions. The mechanisms are mainly the ones described throughout the paper but factoring in the cumulative life cycle model described by Carneiro and J. J. Heckman (2003) which shows a high return to early interventions and a low return to remedial or compensatory interventions later in the life cycle.

To put it into more context, what they describe as early interventions are children between ages 0-3 whereas the children in this sample are much older (8 years old or older). Assuming their argument is at stake here, I would expect for the effect of the treatment to take more time than traditional early childhood compensatory programs which do find improvements even in the first year (Elango et al., 2015).

2.3 Research questions and hypothesis

This chapter will seek to answer two questions related to the effectiveness of the program and whether it had varying effects depending on the SES of the students.

More to the point, this chapter follows a similar question as in Gertler et al. (2012) but through different means. They study the effect of the treatment on school dropout rates, repetition rates and test scores. However, all of these indicators (including test scores) were calculated at the school level using the public school census. That is, they matched each school in the experiment to their corresponding dropout rate, repetition rate and test scores averaged at the school level. At this point, they performed all of their statistical analysis only using school-level information. This means that they did not leverage all of the student-level information gathered for every student-school-year combination and cannot study whether the treatment varied by the SES of the student.

The first research question is concerned in replicating the results of Gertler et al. (2012) for test scores but using a different dependent variable. Given that one of the key aims of the paper is to look at the changes in SES in the treatment effect, the first thing to do is to study whether there is indeed a treatment effect at the individual level. Given that the experiment in Gertler et al. (2012) used only 250 data points, in order to leverage a higher sample size and a similar but completely exogenous indicator of child performance, the paper used children's score from the last test at the individual level (there is no reason to believe that a particular child's score is endogenous to an aggregated school-level score). This has several advantages. First, it proves as a replicability of the Gertler et al. (2012) experiment under similar scenarios. This serves as a good confirmatory analysis to their previous findings, something well needed in empirical research due to high non-replication rates Collaboration et al. (2015) in social science research. In this scenario, although the same exact experiment is used, a replication becomes plausible given that it uses a exogenous dependent variable when compared to the school-level scores but measuring a similar underlying variable. That is, the sources are independent of each other and child-level scores should not be strongly correlated with aggregated school-grade level scores.

All of the above would give even greater credibility to the findings in Gertler et al. (2012) and would open a way to test for an interaction with the SES of the students. Given that the SES interaction cannot be tested with the data on test scores used by Gertler et al. (2012), it is for this reason that the paper choose to keep the dependent variable at the individual level. Although this variable has several limitations described in the results and limitations section, it opens a way to replicate the findings, increase sample size (and reliability) and test for the heterogeneity of the treatment effect through an SES interaction.

Below I describe both questions with their respective justifications.

1. Did the increase in funding on treated schools led to higher test scores for treated units?

- Hypothesis 1: Considering that both treated and control schools were already receiving extra funding and had parents participate in the school decision-making process, it can be expected that the effect of the treatment on test scores will be positive but small. The main argument why this happened comes from previous research on the topic. As discussed above, there is qualitative evidence from H. Patrinos (2006) on the specific provinces under study that when the AGE programs were introduced, parents got very much involved at school. On top of this, they were more demanding with teachers when it comes to their children's performance and also got very interested in the academic well being of their children. 6 years later, this was also among the questions that Gertler et al. (2012) tested and they found an overall increase in parental participation at school affairs and on being more demanding with their children's performance. Equally important, Gertler et al. (2012)'s findings show that this increase in interest was present with the double AGE, and not the traditional AGE which was found in H. Patrinos (2006), six years before. With all of this being said, it then becomes very plausible that these increases in attention and interest towards their children's well being, were also paralleled by some increased attention at home.

Theoretically, a school intervention that improves parental participation both at school and at home gives reasons to believe that this will translate into better parent-student interactions related to homework and school affairs and this will lead to higher test scores. Gertler et al. (2012) acknowledge this indirectly but fail to evaluate it.

In a study of the AGE program in the same provinces in the study from Gertler et al. (2012), H. Patrinos (2006) offers qualitative evidence that when the AGE programs were introduced, parents got very much involved at school. Equally important, they were more demanding with teachers when it comes to their child's performance and also got very interested in the academic well being of their children. Additional evidence from Gertler et al. (2006) also suggests that parents got very interested in their child's performance and increased both pressure to teachers and attention at home. This study was also a qualitative survey on many of the parents participating in AGE programs in the year 2006 in many of the provinces that also participated in the study in 2012.

According to their survey, parents noted that the AGE program helped to reduce child labor. Since the increase in funding through the parents association has nothing to do with the prices of the school, one possible explanation is that parents decided to stop sending their children for labor and concentrate on school, a trait related to more home involvement and interest in their child's well being at school. But the above can be described as too ambiguous and unsupported by evidence, which is certainly true. However, parents also listened to advice from the teachers on how to help their children. It is also likely that this advice was somewhat related to some type of support at home. An excerpt from the study of Gertler et al. (2006) says:

> "...[Parents reported that they]... are careful to listen to teachers on how to improve their child's performance..."

And even though the parents interviewed by Gertler et al. (2006) suggested many likely mechanisms for the improvement of AGE, the authors clearly mentioned the argument discussed in the hypothesis section as a possible mechanism:

> "...The AGEs meetings are important for the school as they facilitate dialogue between parents, teachers and school directors; consequently improving school climate. This is believed to further foster parental involvement in the school, as well as at home with their children's school work..."

Six years later, this was also among the questions that Gertler et al. (2012) tested and they found an overall increase in parental participation at school affairs and at home. This is not to say that other mechanisms were at play, as the parents suggest in Gertler et al. (2006):

> "...Parents believe that the AGEs at least to some extent help motivate their children to study more, probably through the increased teacher's effort and motivation..."

It is in fact very plausible that the increase in performance could be related to improved teacher performance, more teacher working hours, stronger accountability from parents to teachers or simply more resources to spend at school. I am sure that all of these explanations played a role in the increased performance and are well documented in the qualitative discussion of Gertler et al. (2006).

However, none of these arguments seem to explain whether there would be an SES effect. Why would some SES groups benefit more from all these interventions? The only two reasons I can think of are (1) schools are completely segregated such that high SES schools pushed their teachers more than others and thus the increase in performance happened across schools and not within or (2) parental SES influenced performance through other mechanisms, such as a better and more fruitful relationship at home that improved their child's education.

In fact, Gertler et al. (2012) recognize this and suggest that the change might have actually happened at home:

> "More work is needed in determining if the effect comes from an incidence of changing within household activities for supervising children improvements in school and studying practices. In this sense, some studies, like Valdes et al (2009) have found for Mexico, using qualitative methods, that parents circumscribe their support for children improvements at school only *within the household and not necessarily interacting with teachers and principals*."

The first explanation (1) is certainly possible but as will be seen from above, all of the modeling from the paper adjusts for the percentage of mid-high SES students in the school. This means that the all estimates account for the fact the there might be over concentration of SES groups in particular schools.

A variant of the second explanation above (2) is that not much happened at home, but different SES parents exerted different types of pressures to teachers and thus teachers helped improve, for example, high SES children more than they helped improved low SES children. This argument could certainly explain the phenomena. Yet the argument loses strength once we account that many of the meetings and activities happened through the parent's association. This means that the interactions were happening with all parents together and the possibility of a given parent exerting more pressure than others is likely to be lower than if they were one-to-one meetings with the teachers frequently.

Having said all of that, the paper resorts to what I think is the most likely mechanism, which is also suggested (but untested, both in Gertler et al. (2006) and Gertler et al. (2006)) in Gertler et al. (2006), that in some way, SES factors influenced the way in which children improved their performance. It is justified that this mechanism does not play a central role in the literature cited above given that they do not examine any SES effects. However, once we factor in the SES variable, their above mechanisms loose some of their strength and other possible mechanisms become relevant.

2. Does the treatment effect vary depending on the SES categories of students?

    - Hypothesis 2: I expect for the effect to vary by SES origin. I presume that children coming from high SES backgrounds will benefit more from these interventions as the parent-student interaction will be more meaningful in terms of cultural and cognitively enriching activities. This is the case in many of the school-level studies documented by Van de Werfhorst and Mijs (2010b) but also evident in life course research concentrated on children's cognitive abilities (Bradbury et al., 2015). As a counterpart, students from low SES groups will benefit less from the intervention considering that their parents have less to offer when it comes to decisive cognitive and non-cognitive skills on average. This is essentially the thesis put down in Kautz et al. (2014) but not applied in the context of a developing country.

2.4 Methods

2.4.1 Data

The experiment was carried out in 2007, 2008, 2009 and 2010 in which the same questionnaire was applied to random students in the selected 250 schools. The questionnaire contained hundreds of questions ranging from demographic characteristics to attitudes towards school. In addition to the student questionnaire, the teachers, parents and principals also received questionnaires asking about their role and relationship to the children. Although a rich source of information, in theory there is no need to control for any variable given that the intervention was randomized. However, more variables need to be considered because the sampling of schools was stratified and thus it can be a strong confounder. Below I describe each of the variables used in the analysis.

As dependent variable I will use their scores on their latest Mathematics test which was less than a month ago. Students were also asked about their latest test score in Spanish and Natural Sciences but I will use Mathematics because the population under study is indigenous and some of them might be more fluent in Spanish than others. I include all modelling results for Spanish and Natural Sciences in the appendix.

The mathematics test score is an ordinal variable with scores censored at the lower range of the categories. The variable has six categories ranging from a score of '5 or less' (bottom censored) to '10'. Given that the scale of the test scores is censored at the bottom grade with five or less, it is difficult to ascertain whether this variable can be treated purely as a continuous variable. I proceed the analysis treating it as an ordinal categorical variable in order to obtain more reliable estimations [7].

It is usual in educational research to recode test scores into groups in order to average out specific effects into low, middle and high groups. However, given that the variable has very few categories, I choose to keep it in it is original form in order to maximize variation in test scores and not constrain students to be even more similar into hierarchical groupings. This strategy works well here because there are enough observations in each category to be able to estimate effects at each original grade.

As independent variables I will include

- a dummy for treated and control groups
  - 1 = Treated
  - 0 = Control (reference)
- dummy variables identifying the four years of the study
  - Year 2010
  - Year 2009
  - Year 2008
  - Year 2007 (reference)

[7] One could argue that the spaces between the scores are substantively equal and thus it should be treated as a continuous measure but the censoring combined with the few categories available make the variability small. This, if treated as a continuous measure, can throw unexpected estimations because the errors are clustered into specific groups due to the non-continuous nature of the variable.

- a categorical variable denoting low, middle and high SES groups
  - 3 = High SES
  - 2 = Middle SES
  - 1 = Low SES (reference)
- a dummy for whether the school is a general school or an indigenous school
  - 1 = General school
  - 0 = Indigenous school (reference)
- proportion of students of high-mid SES within each school
- a continuous variable for the age of the student
- an interaction term between the treatment dummy and the dummy variables for the years of the study

In principle, the important variables here are the treatment dummy, the variables for the years of the study, the categorical variable for the SES of the student and the interaction term between the two. However, I also include a dummy for the type of school because the randomization was stratified to compare general and indigenous schools. Additionally, I also include the age of the students because given that there is high repetition rates in most of these schools; some grades might have students with different ages and that might explain the variation in the treatment effect within a grade.

The data is freely accessible through the website of the World Bank. For the 2007 data, data can be found in http://microdata.worldbank.org/index.php/catalog/1036, for 2008 http://microdata.worldbank.org/index.php/catalog/1037, for 2009 http://microdata.worldbank.org/index.php/catalog/1038 and for the end line data access http://microdata.worldbank.org/index.php/catalog/1039.

2.4.2 Defining SES in developing countries

One of the main objectives of this chapter is to look at how the treatment affected children from different socioeconomic levels. Questions like parent's education, income or parent's occupation are standard proxies of SES in Western societies. In this case, it is different. Aside from the fact that the four provinces under study are among the poorest in Mexico, the social structure is mainly populated by low-level jobs, farmers and informal labor. This means that the validated and traditional social class schemas (Erikson and Goldthorpe, 1992; Breen and Goldthorpe, 1997) are probably not trustworthy in this scenario. The children interviewed in the survey range from 3rd to 6th grade, which also increases the probability of recall error and imprecision when it comes to questions such as education and occupation of the parents.

For example, figure 2.1 shows the average score for Mathematics, Natural Sciences and Spanish for all levels of father's education in 2007.

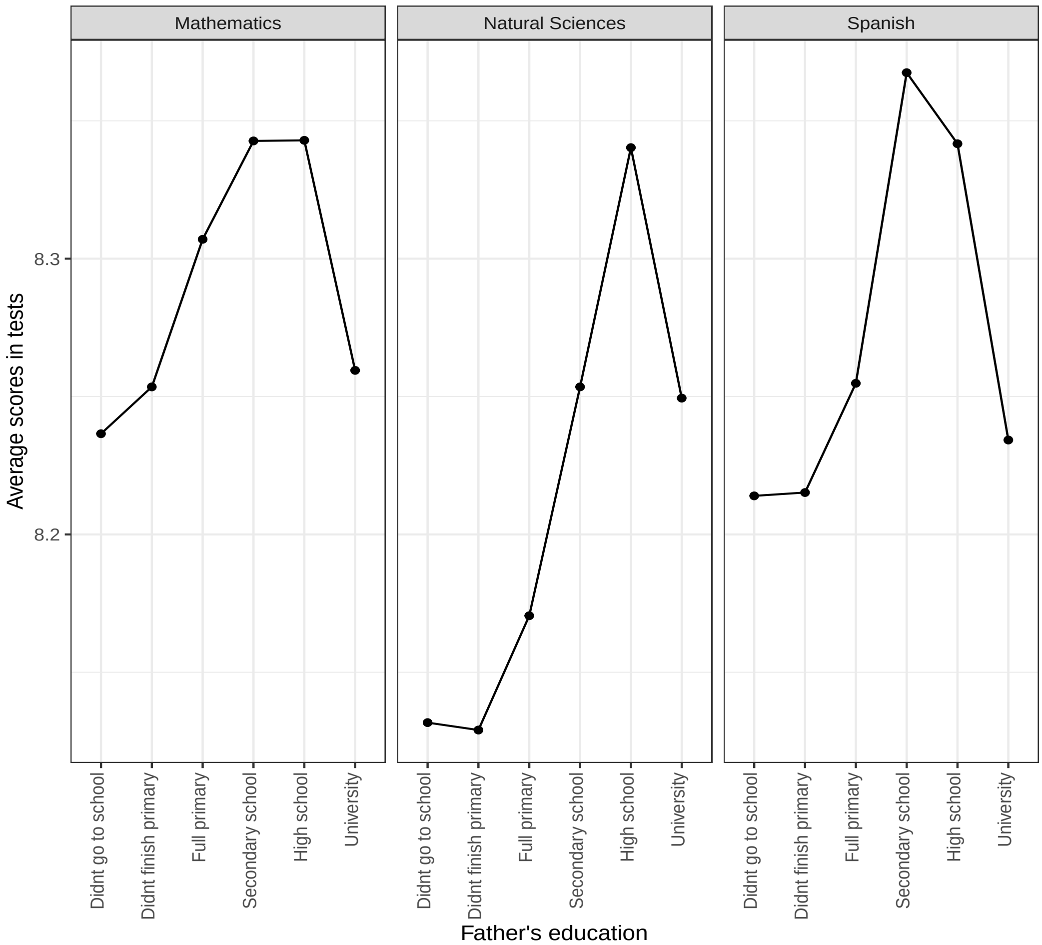


Figure 2.1: Average score in Mathematics, Natural Sciences and Spanish for different levels of father's education in 2007

There is some sort of positive correlation but according to the student's response, children from highly educated parents have lower scores than children from families where parents only completed primary school. I find this very difficult to believe and acknowledge that this is a well known limitation of studies in developing countries, specially where young students answer questions with incomplete information. However, researchers have resorted to other measures of SES. For example, the work of Bollen et al. (2001) describes several strategies to calculate proxies of SES in developing countries, and more concretely Vyas and Kumaranayake (2006) suggest which variables to use based on what the literature has found to be important.

The categories of variables used to measure SES are consumption/income, education, ownership of goods and demographic conditions. From all the previous categories there are many variables of interest ranging from income/ownership of property to how many rooms in the household are used for sleeping. Although there is data on the education and occupation of the parent's of the child, I am hesitant to rely *only* on these because of imprecision in the answers. That is, they are most likely to represent differences in socio-economic status but given that they were answered by the child, they have too much recall error to be used reliably in isolation. Instead I use this variable in combination with variables related to the availability of goods in the household as well as the household structure. Note that I discard father's occupation because in order to develop a valid hierarchy of all jobs I would need to have detailed knowledge of the class structure of indigenous regions in Mexico.

Following the approach of Kolenikov and Angeles (2009) in combination with Filmer and Pritchett (2001) I will estimate a joint Principal Component (PCA) of these variables:

- What is the level of education of your father?
- How many rooms are used for sleeping in your household?
- How many people live in your house aside from yourself?
- How many of these do you have at home?
  - Car / truck
  - Refrigerator
  - Microwave

- Washing machine
- Telephone line
- Cellphone
- Television
- Videocassette
- DVD player
- Computer
- Educational programs in your computer
- Educational resources such as encyclopedias.

To capture more meaningful variation from the selected variables I apply two transformations. First, I calculate the ratio of rooms to people living the household instead of using both variables. This has been effectively used before in the study of Vyas and Kumaranayake (2006) and has been shown to work well in other PCA estimations in developing countries. In addition, I create a single variable that adds all the items that the child has at home (such as number of cars, microwaves, etc..). This captures the average material goods that the household has. Considering that the simulations performed by Kolenikov and Angeles (2009) showed that PCA is not affected greatly by using ordinal variables, I will apply the estimation using father's education which is not completely continuous.

However, given that the reduction algorithm is using several variables for the estimation, this reduces the effective sample size by about 10% because all of them have missing values. Throughout the main text I provide all estimations with the raw unimputed data. However, in the appendix I impute the missing values through using the MICE (Multivariate Imputation by Chained Equations) approach (Barnard and Rubin, 1999; Buuren and Groothuis-Oudshoorn, 2010) and provide all results with this imputed data sets. These results are in section 2.8.2 in the appendix.

Table 2.1 and table 2.2 in the appendix show the composition of some selected variables used in the PCA for treated and control units for years 2007 and 2010 before imputing the missing with the most frequent value. Both tables show not only that there is no difference between treatment and control units but that there are no important differences between both tables, suggesting that the index should be stable across time.

I estimate the factor loading of these variables by calculating a singular value decomposition of the data matrix and not by using an eigenvector on the covariance matrix. The first and most important factor loading explains 43% of the variability of all variables in 2007 and 41% for 2010, suggesting that these variables are indeed somewhat correlated. These values are also reassuring relative to other studies where the proportion of variance explained was somewhere between 30% and 50% (Vyas and Kumaranayake, 2006). Having said that, the main concern of this dimensionality reduction is whether the results of the predicted index makes sense in terms of representation of the SES groups. Below I present the index compared to more traditional measures of SES. For parsimony and lack of space the results are only presented for the baseline year 2007.

Figure 2.2 plots the distribution of the index as well as a categorization of the index into three groups representing the low, middle and high SES groups. Each of these groups contains 30%, 40% and 30% of the data respectively. Separating these groups allows to capture each SES group separately and test it with the predictions for it's validity.

In the top plot the distribution looks vastly concentrated on the lower values (between -1 and 5, positively skewed) with a small number of outliers, just as traditional income measures. Moreover, the plot does not show any particular gaps in the distribution, suggesting that the variable was smoothly calculated. The skewness looks particularly severe in the last two years of the study, so it is something we need to consider when interpreting our final results [8]. We also see that when dividing the distribution into low/mid/high groups based on quantiles in the bottom plot, the proportions lie as expected with the vast majority in the bottom/mid group and only a decreasing share in the top group. Do note that even though it seems as if the SES index has the same value for different SES groups (for example, the maximum for the low SES group overlaps with the lowest from the middle SES group), they do not overlap. This happens due to the nature of the density plot which does not cut the end of the density right after the value. The true values align with what is expected: the maximum of the low SES is lower than the minimum value of the middle SES group, and so on.

[8] In fact, I have deleted a few outliers in the last two years because they were shifting the distribution upwards by about 10%

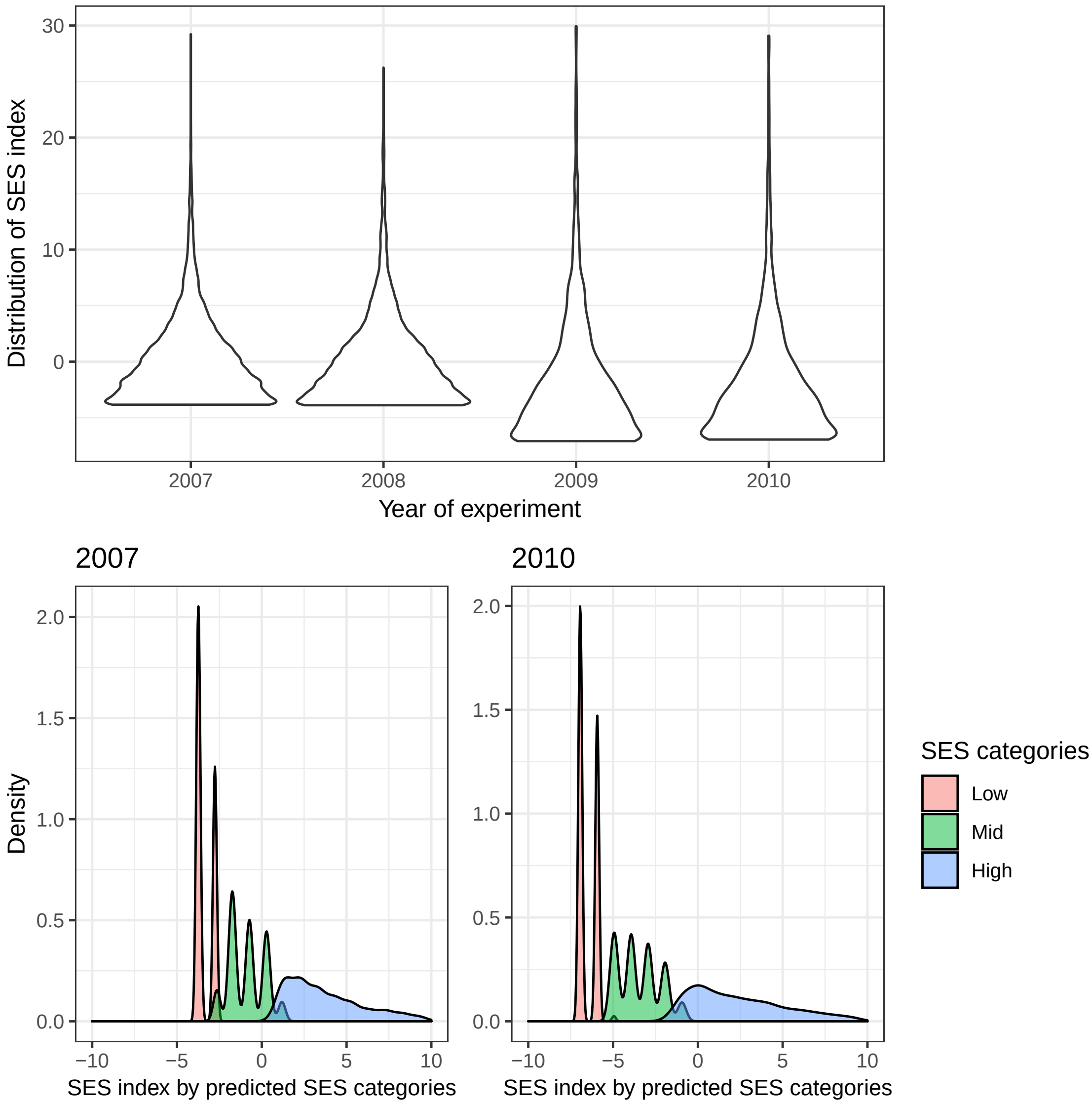


Figure 2.2: Predicted SES categories and SES index

Since the metric of the SES distribution is meaningless given that it is standardized, we need to compare the SES categories against the actual values of the father's education. Although I believe father's education is not a completely reliable indicator of SES in this context, there should be some degree of correspondence between both indicators. Figure 2.3 plots the distribution of father's education by the different SES categories.

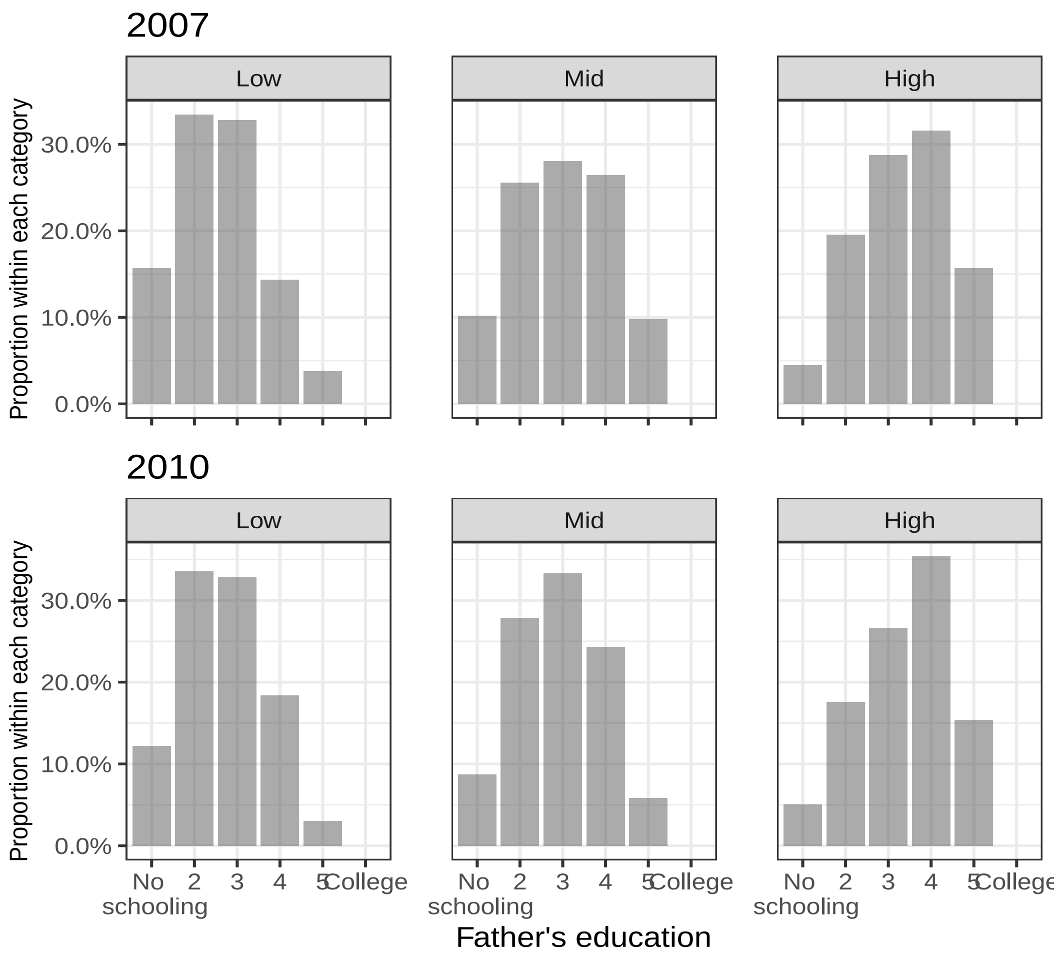


Figure 2.3: Predicted SES categories (top) by father's education

For the low SES, the average education is close to no schooling, whereas in the middle group the average moves to the center. For the high SES group, we see a shift in the distribution with very little respondents in the lower levels of education and higher shares in the top levels of education. In other words, there is an evident positive correlation between both measures.

As outlined above, the SES categories seem to be robust and well identified in terms of its distribution and what it measures. The tables presented before in the appendix also give support to the similarity of the index across time given that the variables have similar composition across all years and between treated and control groups. Despite this, the results should be interpreted carefully for the high SES group, particularly for the last two years considering that the results might be shifted by other outliers.

The main concern for these outliers is the fact that some people with very high SES index are also grouped with people of high SES index, but much lower than their value in the high SES category. This can shift the results for the high SES group given that we have respondents in the high SES category which could have higher values in unmeasured categories such as income and cultural capital, among other things.

The additional analysis added to section 2.8.2 in the appendix contrasts this relationship between the SES categories and the test scores and the results line up as expected: there is positive linear relationship between ordered SES categories and test scores.

2.4.3 Methodology

As the dependent variable is an ordinal variable, the most appropriate way to model the outcome is using an ordered categorical regression model also widely known as the proportional odds model. More formally, ordinal outcomes fall in one of J categories. In this case, these are 5 categories. One way to express an ordinal model is to introduce a latent variable y^*, that is related to the observed outcomes by estimating a latent variable from the outcome of interest:

$$y = \begin{cases} 1 & \text{if } y^* < \zeta_1 \\ 2 & \text{if } \zeta_1 \leq y^* < \zeta_2 \\ \vdots & \\ J & \text{if } \zeta_{J-1} \leq y^* \end{cases},$$

where $\boldsymbol{\zeta}$ is a vector of cut points of length $J-1$. We can think of this as estimating a linear model where we regress the latent variable y^* on the covariates of interest in a model similar to

$$y^* = X_i\boldsymbol{\beta} + \epsilon,$$

where $X_i\boldsymbol{\beta}$ is a data matrix with all covariates and ϵ is an independent error term which usually has a logistic probability distribution as in classical logistic regression.

Given that the data cannot really distinguish the intercept from the cut-off points, the model has no intercept. The coefficients of the model are indeed estimated with a reference category but the reference category is not present in the model output. As described in the description of the experiment, students are nested into schools and into years and this might actually confound the effect size, leading to severe overestimation of the coefficients. Moreover, the fact that the same school is repeated over time might also induce some overestimation. One solution would be to use a cross-classified multilevel model where it is possible to specify that a school is present in every year. However, it's not entirely appropriate in this setting because there are only four time points (with few groups both clustered standard errors and random intercepts are problematic). Having said that, the final multilevel model will contain the matrix of covariates with the intercept varying by the school and including a dummy for each year (fixed effects) to control for differences between years. This accounts both for the clustering within schools and adjusts for differences between years. Note that all estimation are only interested in the random effects as a way to adjust for clustering so the coefficients of the random effects will not be discussed.

Classically this model is estimated via maximum likelihood through a frequentist approach. However, recent advances in statistical computing have made possible the use of bayesian inference without the need of expensive computing environments. More concretely, the development of the Stan programming language (Carpenter et al., 2017) has allowed for full bayesian statistical inference using a variety of algorithms, including Markov Chain Monte Carlo (MCMC). The reasons for choosing a bayesian approach over frequentist are many and I will only discuss the benefits briefly in the context of this study. For a discussion on frequentist and bayesian methods, see Gelman and Shalizi (2013). For more concrete criticism of frequentist methods see Gelman (2013) and Gelman and Loken (2013).

The two main reasons for choosing a bayesian framework in this study is because it is possible to incorporate prior knowledge for the coefficients and because the interpretation of the posterior distribution is much more straightforward than frequentist point estimates (Gelman and Shalizi, 2013). Bayesian inference allows to weight the estimation of the coefficients by specifying a prior distribution on each of the covariates. That is, not only does it leverage the strength of the data to estimate each coefficient but it maximizes its accuracy by allowing to incorporate theoretical and empirical evidence through the specification of a prior distribution of possible values.

Whenever an experiment is carried out without any sampling error and all variables in the data matrix are completely free of measurement error, then allowing each coefficient to have a prior uniform distribution from $-\infty$ to ∞ is harmless [9]. In most social science research, like this experiment, that is not the case. For example, the main dependent variable of this study is the grade of the student. However, estimating a single grade for a student is based on evaluating a series of exercises in a test that carries together uncertainty and subjective decisions. Moreover, the question was asked directly to students, which adds even more uncertainty and error because they could have exaggerated the test scores [10]. A possible remedy is to acknowledge the uncertainty in the variables under study and provide informative priors for each of the coefficients based on past research and theoretical reasoning.

From the experiment by Gertler et al. (2012) there is evidence that in this specific scenario, test scores increased, although at the school level. Based on the previous work outlined in the literature review I have also showed that, albeit theoretically, scores should improve with these type of interventions, there is mixed evidence. Following this logic, then it is appropriate to provide a distribution that allows for a small positive effect as well as a very small negative effect (due to the mixed evidence). This distribution should not be strong enough to constrain the estimates towards one specific side of the distribution. At the same time, this distribution should allow for some type of bigger effects but their probability should be very low. I will model the distribution for all coefficients using a t

[9] Frequentist estimations assume that each coefficient in a model has a prior uniform distribution from $-\infty$ to ∞, so all values are equally likely. In social sciences, especially in survey data, this is inappropriate as there is considerable amount of measurement error and randomness in the data. Bayesians counteract this problem by specifying a likely distribution for the coefficients

[10] As will be shown in the descriptives section, the distribution of the grades suggests that there is no over concentration of grades, something which would suggest that students are reporting scores in the top marks only. Despite this, the previous arguments add up for considerable measurement error

distribution with three degrees of freedom [11]. This is what Gelman et al. (2008) call a weakly informative prior.

For the hypothesis in the paper this distribution is informative because it allows the distribution to concentrate around zero (to allow for small effects) but also provide wide tails (allowing positive and negative values more widely, unlike the normal distribution), leaving room for any abnormal big effects with very low chances of occurring. The t(1) distribution provides an even wider distribution (with even wider tails) than the t(3) and thus a slightly concentrated distribution (that is, the t(3)) is more appropriate because most of the effects discussed in the literature review are small. Finally, the t(3) is appropriate in this case given that it narrows the tails such that the effects are constrained toward zero, as the literature suggested some null findings.

The recommendations for the using the t distribution comes from the work of Gelman et al. (2008), where they propose this distribution along with the cauchy distribution as standard weakly informative prior distributions for logistic and other regression models. I do not use this distribution merely because it is fairly accepted as a standard weakly informative prior but because the description above fits it very well with what was discussed in the literature section and what theory should expect.

More formally, the dependent variable is distributed as:

$$y \sim OrderedLogistic(X_i\boldsymbol{\beta}, K-1)$$

where $X_i\boldsymbol{\beta}$ is a vector of coefficients corresponding to each covariate where all of the coefficients are modeled using a t distribution with three degrees of freedom and the $K-1$ cut-off points are assumed to have a prior uniform distribution. The prior uniform distribution (of y^*, the latent variable from which the model estimates the cut off points) makes sense because it is expected that scores are equally likely to be either 5 or less, 6, 7, 8, 9 or 10. That is, we have no reason to believe that students are more likely to score 9 or 10 rather than 6 or 7 or vice versa [12].

[11] Gelman et al. (2008) uses a a cauchy distribution, a slightly different parametrization of the t distribution. All in all, the t distribution is very similar

[12] Note that the prior should reflect the theoretical and empirical expectations and not the direct distribution of the dependent variable. Even though the distribution of test scores scores is slightly skewed, it does not mean that the prior should reflect exactly the same distribution

There is also the possibility of using non informative priors or "flat" priors. These priors would give each of the coefficients the same likelihood of being $-\infty$ to ∞. If I were to give flat priors to all of the coefficients in the models, the results would be fairly similar to classical frequentist approach. The leverage and importance of Bayesian methods only work when prior knowledge is incorporated through prior distributions. Using flat priors is just fine when not much is known about a phenomena and the research agenda is first attempting to study it. However, for this study this is not the case and given the previous empirical evidence I try my best to incorporate as prior information.

In the next section I show the descriptives of the main variables to inspect the distributions and show the results of this estimation.

2.5 Results

2.5.1 Descriptives

As discussed above, the experiment was implemented in the 3rd, 4th and 5th graders systematically, and 6th graders only in 2009. However, the experiment was not completely successful either for 3rd graders due to a teacher strike in 2009. The principal investigators discuss the problems that were associated with this strike (Gertler et al., 2012). Through the rest of the paper the analysis will concentrate on studying 4th graders, to match a comparison to Gertler et al. (2012), but also replicate a summary of the statistical analysis on 5th graders in the appendix to avoid making the chapter too dense.

The descriptives table below contains frequencies and percentages for the main categories of the independent variables only for 4th graders. These numbers relate to all student who completed questionnaires collected from the treatment and control schools. Since all models exclude missing values from any of variables present in the model (dependent as well as independent variables), all descriptives and plots below already exclude missing values in order for the total sample size to match between model summaries and descriptives.

Table 2.1 shows percentages within each category of the main variables used in the analysis. To begin, let us confirm both treated and control units are balanced. In terms

	2007		2008		2009		2010	
Variables	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage
Treatment								
- Treated	395	0.54	607	0.51	375	0.54	555	0.54
- Control	342	0.46	572	0.49	322	0.46	472	0.46
SES categories								
- Low SES	254	0.34	333	0.28	216	0.31	293	0.29
- Middle SES	289	0.39	467	0.40	256	0.37	449	0.44
- High SES	194	0.26	379	0.32	225	0.32	285	0.28
Type of school								
- General	380	0.52	638	0.54	344	0.49	554	0.54
- Indigenous	357	0.48	541	0.46	353	0.51	473	0.46
Type of grade								
- 4th grade	737	1.00	1179	1.00	697	1.00	1027	1.00
Age								
- 7	4	0.01	0	0.00	0	0.00	0	0.00
- 8	30	0.04	10	0.01	2	0.00	2	0.00
- 9	0	0.00	0	0.00	0	0.00	0	0.00
- 10	423	0.57	756	0.64	485	0.70	715	0.70
- 11	163	0.22	243	0.21	122	0.18	180	0.18
- 12	77	0.10	111	0.09	52	0.07	80	0.08
- 13	24	0.03	40	0.03	28	0.04	34	0.03
- 14	15	0.02	13	0.01	7	0.01	10	0.01
- 15	1	0.00	6	0.01	1	0.00	6	0.01

Table 2.1: Descriptives of main variables used in the analysis for 4th grade

of sample size, for all years the sample size is between 340 and 600 for both groups. As mentioned in the section about the design of the experiment, the last two years of the study included two other control groups which consequently reduced the total sample size for the initial treated and control schools. Despite this, the percentages for each year are practically the same. Although I do not present it here, Gertler et al. (2012) document in detail how both treated and control units are not only balanced in terms of sample size but also on around 190 school features, which they tested to be significantly similar. The predicted SES categories from the PCA also seem balanced for all years.

Another important variable here is the age of the students. Considering that the sample contains students from 4th and 5th graders across all years, we see significant variation in the age of the students. This is most likely related to high levels of repetition within grades. In fact, the principal investigators (Gertler et al., 2012) did study the repetition rate of the school as one of their most concerning of their outcomes. This same descriptive table is present for 5th graders in table 2.3 in the appendix, showing a balance sample across the four years.

The final sample size for the treated group across time is 395, 607, 375, 555 respectively for 2007, 2008, 2009 and 2010. The sample size for the control group for the same years is 342, 572, 322 and 472. The model controls for all of these variables except for grade in order to capture any residual variation left from the randomization [13].

It is also important to make sure that the dependent variable is balanced across all years and the distribution is well balanced for all grades. Figure 2.4 visualizes the distribution of the test scores separated by treatment and control groups and across all years of the experiment.

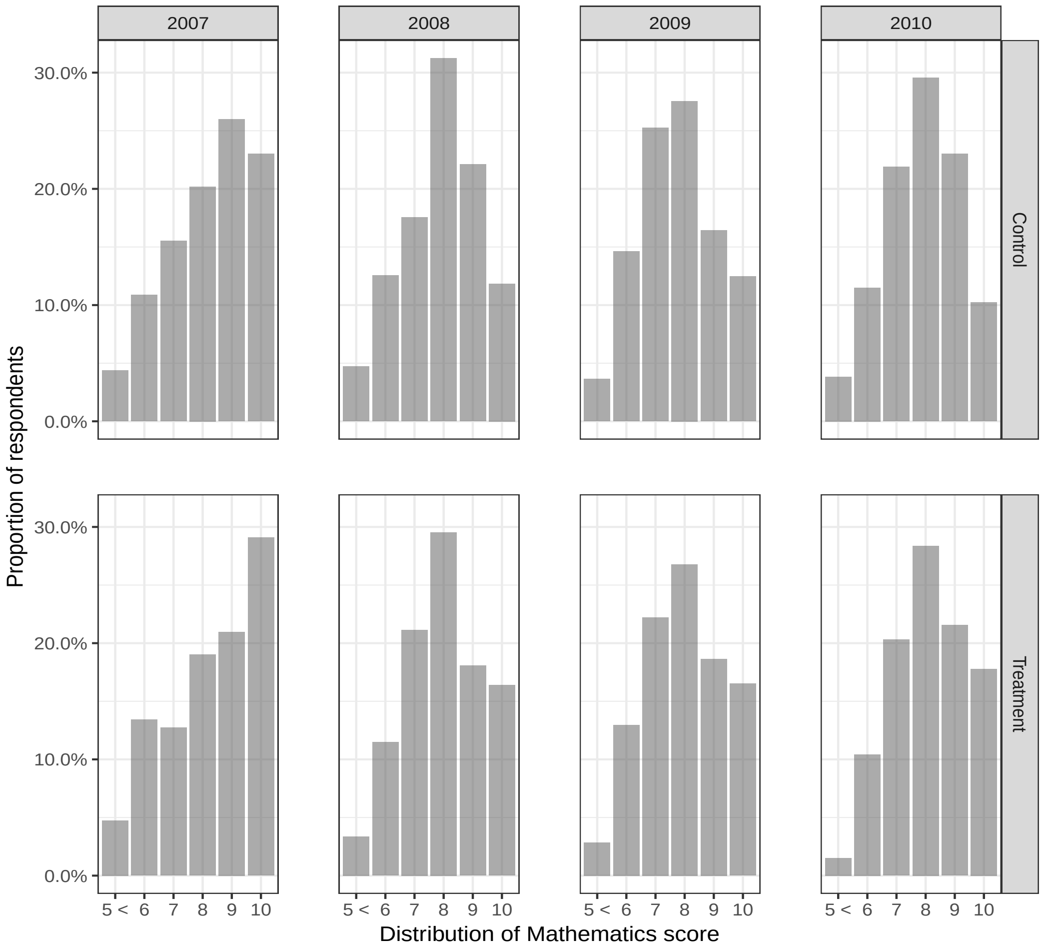


Figure 2.4: Distribution of Mathematics test score for treated and control groups for all years for 4th graders

With the exception of 2007, the distribution is centered at 8 with a normal shape. Considering that test scores were asked to the students rather than recorded independently

[13]Theoretically speaking it is not necessary to control for anything except the type of school because the treatment was completely randomized. However, I do control for additional variables because there are enough degrees of freedom for it to be harmless and improve accuracy through residual variation

by teachers, this distribution is reassuring. There is no way of disentangling whether a child lied when responding about test scores but this distribution does not have extreme over concentration in either the top or bottom scores. In fact, it shows a natural variation where very little students scored neither very high nor very low. This distribution seems to have changed slightly across time. For example, we see that the number of students scoring in 10 for the treated group relative to the control group is higher for the years 2009 and 2010. As we see later on, this effect is confirmed in the modeling section. Despite this, we also see that the distribution for 2007 is different. It is unclear what the explanation for a changing distribution between 2007 and all other years is but what's important is that the treated and control units have similar shapes although with treated units having an over concentration on scoring the highest mark. We should take this caveat into consideration when interpreting our results.

2.5.2 Modeling and results

Figure 2.5 presents the results of the model described in the methodology section. As described, the model is fitted using a bayesian ordered multilevel logistic regression, so the coefficients obtained by the posterior distribution are expressed as logistic odds. I transform them to odds ratios for easier interpretation. As an example, an odds ratio of 1.20 for the treatment group relative to the control group can be interpreted as 25% increase in the odds of scoring higher in the dependent variable for the treatment group. Conversely, an odds ratio of 0.80 for the treatment group relative to the control group can be interpreted as having 25% less odds of scoring higher in the dependent variable. Considering that the bayesian approach turns away from specific point estimates and instead embraces variability, I present the coefficients in a plot with their distribution of plausible values. [14].

To begin, note that the model includes an interaction term between the treatment indicator and the years of the experiment. This interaction reflects the difference-in-difference estimation described in the methodology section, in which the coefficient on the interaction effect is taken as the difference-in-difference estimate of the causal effect. The intercept for the model (not present because it cannot be estimated) contains the reference

[14] For those interested in the classical regression table, see table 2.4 in the appendix

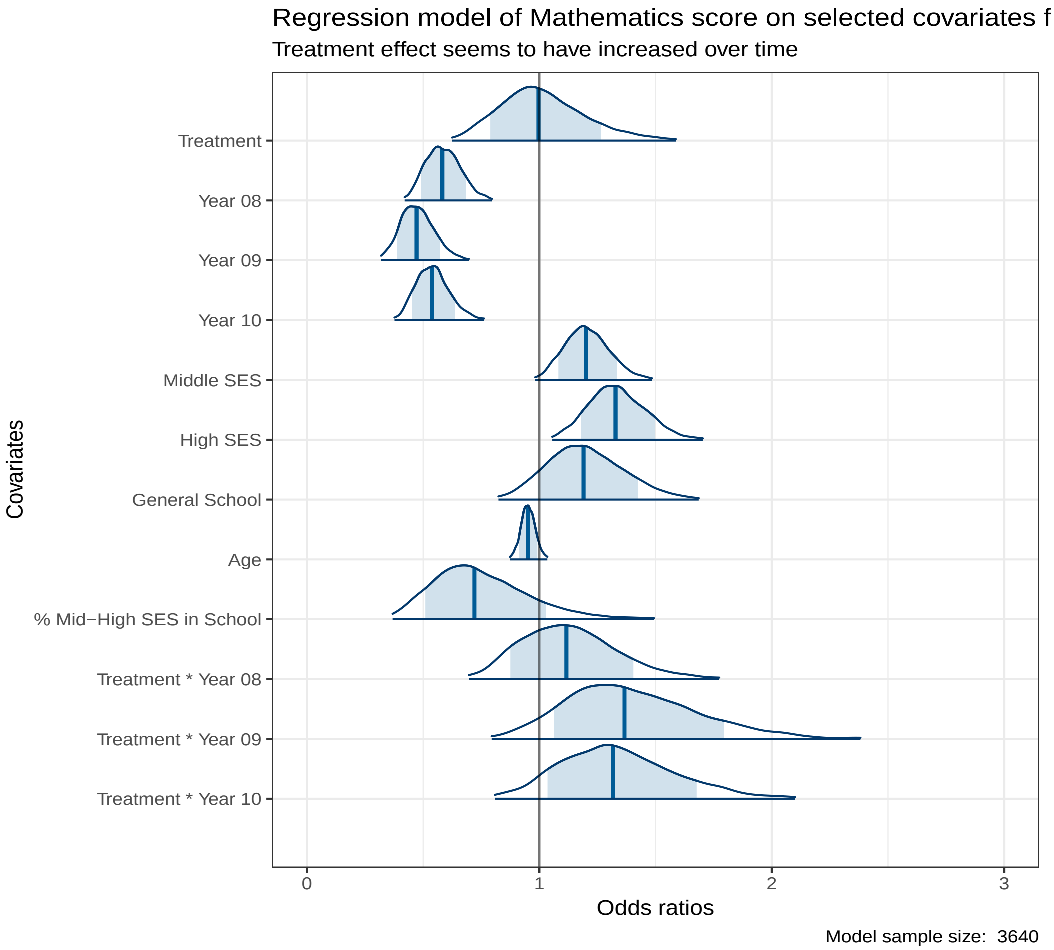


Figure 2.5: Distribution of the coefficients of the model expressed as odd ratiosin Mathematics for 4th graders

categories for treatment and years and all other covariates with their reference categories. In terms of the two variables of interest, the intercept contains the odds for control units in 2007 which have low SES, are indigenous schools, have 0 % of mid-high SES students within a school and have an average age of 0. The treatment coefficient indicates the extra odds of achieving higher grades for the treated group relative to the control group in 2007 with all the other covariates in their reference categories. As expected, the coefficient is as likely to be negative as it is to be positive because in the baseline year both groups of schools were very similar. In a similar line, the main effects of Year 08, Year 09 and Year 10 are the added effect for each year relative to being in the control group in 2007.

These coefficients are not readily interpretable because we do not have the intercept to make the appropriate addition/subtraction to calculate the main effects. Instead, we will interpret the interaction terms instead. The interaction can be interpreted as the difference between treatment and control in one year relative to the difference between treatment and control in another year. For instance, the difference in scores between the treated and control groups in 2007 was ~ 0.00 grade points. The same difference between both groups in 2008 was of about ~ 0.11 as well. The interaction term Treatment * Year 08 estimates the difference between these two differences. If the coefficient in the model is positive for the interaction, it means that the treatment group in 2008 had greater scores over the control relative to the same difference in 2007. The interaction term Treatment * Year 08 (and all other interaction terms) can be interpreted as the difference between treated and control in 2008 versus the difference between treated and control units in 2007.

Let us begin with the coefficient Treatment * Year 08. The difference is not clearly positive but it does seem to nudge towards the positive spectrum with an average increase of about $\sim 11\%$ odds of achieving higher grades. That is, treated schools are 11% more likely to score higher test scores than the control groups in 2008 relative to the same difference in 2007. This difference is small and very uncertain as the intervals are very big. On the other hand, the interaction term for Treatment * Year 09 has a much bigger effect. The treatment seems to have faded in about two years later with reasonable magnitude. To be more precise, treated schools are about 37% more likely (versus control) to score higher test scores in 2009 relative to treated schools (versus control) in 2007. In the same line, the interaction term for Treatment * Year 10 has a very similar shape and treated schools are about $\sim 32\%$ more likely to score higher test scores than control schools in 2010 relative to the same difference in 2007. These effects seem to point out that the coefficients are mostly positive with intervals going between 1 and 2 odd ratios. The uncertainty is reasonably big but the direction and magnitude seem as expected (even when the prior allows for negative values).

The coefficients for the Middle and High SES do line up as expected as they increase linearly. This gives further support for the validity of the SES index. The odds ratios for the general school are also higher than for indigenous school, also expected as these last ones are more disadvantaged. Lastly, we see that a one year increase in age leads to small decrease in odds of achieving higher grades. This makes sense as the higher the age it means that there is a high repetition rate or that children are lacking behind in previous grades. Do note that the model outlined above does not exactly comply with the proportional odds assumption, classically assumed by the model. In table 2.5 in the appendix I present a crude approximation of the assumption using the Brant test (Brant, 1990). The evidence is slightly mixed: whenever the P-value is below 0.05, the assumption is not met. Many of variables do not meet this assumption while only the High SES and Age variables comply with it (note that Low SES is not included because it is used as a reference). It is for this reason that I have estimated the above model by imposing additional restrictions on the thresholds used to estimate the cutoff points in the latent variable. More formally, I relax the assumption by constraining the thresholds to be equally spaced (or equidistant) which forces the underlying response scale in such a way that there is the same distance between adjacent response categories (Long et al., 2006; Christensen, 2015; Bürkner and Vuorre, 2018). This model is replicated with the multiple imputed data set described in the methodology section in figure 2.4 in the appendix. Results look even stronger with the multiply imputed data set.

Do note that the coefficients of the interactions are adjusted for the fact that some students are in general schools, their SES group, their age and the proportion of Mid-High SES students within each school. Considering that the interpretation of the interaction term is not straightforward given that the intercept is not estimated, I will present the probabilities of achieving certain grades by the groups in the interaction term later on in the chapter.

In bayesian estimations it is common practice to evaluate how the model fits the data. Usually, the posterior distribution of the model in the outcome variable is compared to the actual distribution of the outcome variable in the data. Visualizing the difference between both distributions can help assess the strengths and weaknesses of the model. Figure 2.6 shows two plots. The top one depicts the distribution of the average treatment effect averaged over all years together. The bottom plot shows the distribution of scores with the average of the posterior distribution on top of each bar.

The first estimation shows that the latter years seem to predominate in the treatment effect as the average odds increase is of about 15% for treated schools across all years. Furthermore, the model is very precise in the estimation of the Average Treatment Effect (ATE) given that the bulk of the distribution is far from 1, which would represent no relationship whatsoever. Both the results of the model and the ATE give support for the first hypothesis that states that the treatment had a positive effect on increasing grades.

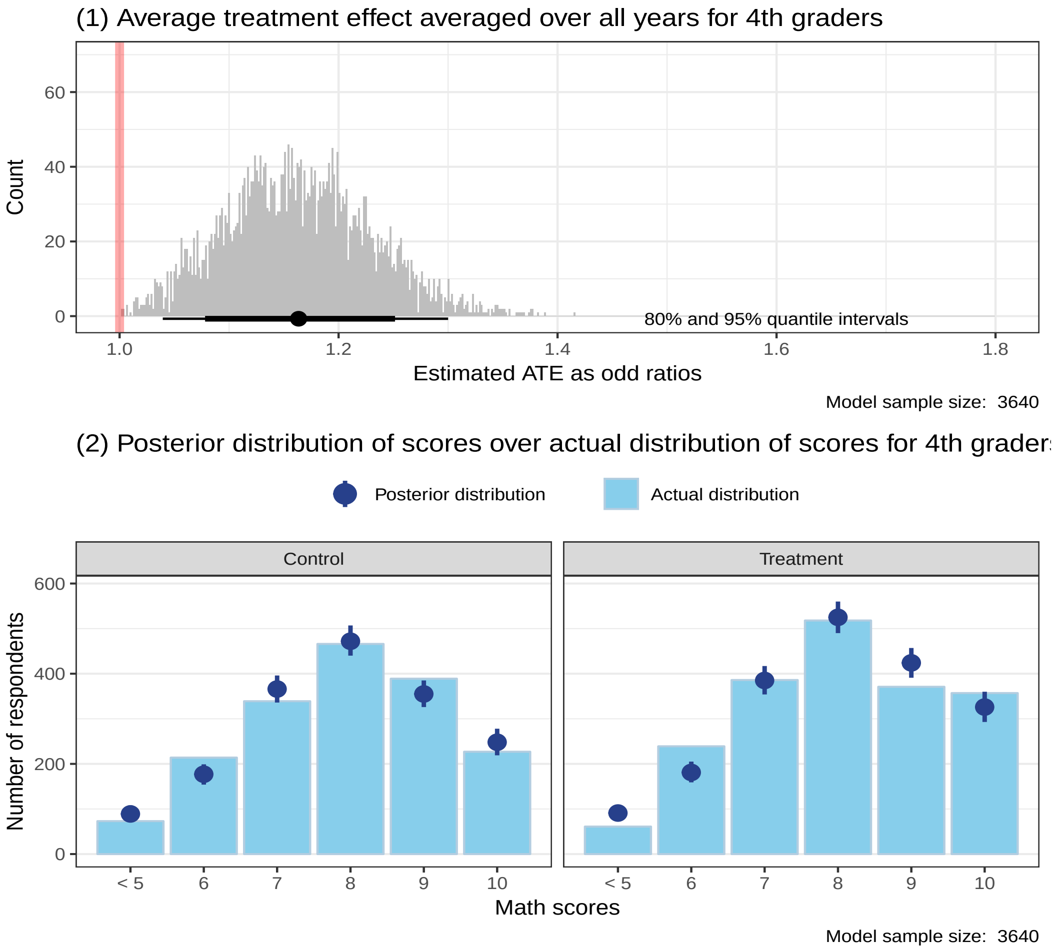


Figure 2.6: Treatment effect averaged over all years and model fit between posterior and actual distribution of test scores

Moreover, it had an even greater effect than expected, as the hypothesis stated that the effect would be small in terms of odds ratios. It averages an increase in odds between $\sim$ 13% and $\sim$ 17%, a quite substantial range. In the bottom plot we can also see that the average posterior distribution for each category fits very well within the categories of the grade variable. This means that the predictions of the model line up very well with the actual data. However, the model has slight deviations when it comes to predicting the bottom two grades.

To summarize, the model seems to be stable and robust in terms of it is prediction and the results show that the treated schools seemed to have a phased in effect in the last two years. These two last years had an overwhelming effect as the average treatment effect is positive and very precise towards the positive spectrum of the distribution. The same results are available for the imputed data set in figure 2.5 in the appendix. As well as the previous graphs, results look even stronger with the multiply imputed data set. As described in the methods section, the model accounts for both clustering within schools through a random intercept and within years, indirectly by keeping the year dummies introducing a fixed effect. However, the model does not specifically account for the fact that schools are present in all years as well. This means that the model could be overestimating the effects slightly and should be taken into account for interpretation.

Next I present the predicted probabilities of achieving certain groups of grades between the treated and control schools. Figure 2.7 plots the results. Let us interpret the meaning of each of the lines. The circled red line in the control group represents the probabilities of achieving either a '5 or less' or a '6' in the test. As it is evident for the control group, the probabilities are increasing steeply as time passes by. In contrast, the same line for the treated schools seems to have stopped increasing as the treatment began and even decreased in the last year of the experiment. Note that the uncertainty intervals for the control and treated schools overlap in the first two years of the study (as we interpreted from the model) but become more distant in the last two years, when the treatment effect became stronger. The squared blue line represents the probabilities of achieving either a 9 or a 10, so the higher marks in the test. For the control group, the probabilities seem to decrease as time passes with a sharp increase in the last year. The treated schools, however, show an opposite pattern. They stop decreasing relative to the control schools and increase more steeply in the last year, much more than the control group.

Despite these clear results, the green triangled line shows that the probabilities of achieving the middle grades, that is 7 or 8, are similar between both groups. It is puzzling as it seems as though the treatment had an effect only on both extremes of the grade distribution.

To explore whether the treatment effect was moderated by the SES of the student, I run the same model from above separately for each SES group. To understand whether the impact of the treatment over time was moderated by the SES, I would need to include a three-way interaction in the specification. For the sake of interpretability I run the same specification separately and show the tables with the corresponding coefficients. Table 2.2 presents the model for the low SES group, table 2.3 presents the model for the mid SES group and table 2.4 presents the model for the high SES group.

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	1.06	0.74	1.51
Year 08	0.62	0.45	0.87
Year 09	0.52	0.35	0.76
Year 10	0.6	0.41	0.87
General School	1.17	0.86	1.6
Age	1.08	0.99	1.19
% Mid-High SES in School	0.7	0.38	1.28
Treatment * Year 08	1.06	0.77	1.48
Treatment * Year 09	1.14	0.73	1.79
Treatment * Year 10	1.13	0.72	1.77
N. Observations	1096		
WAIC (similar to AIC)	3638.73		

Table 2.2: Bayesian ordinal model of Mathematics test score on covariates for Low SES 4th graders. Estimates expressed as odds ratios.

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	1.11	0.85	1.47
Year 08	0.63	0.49	0.81
Year 09	0.53	0.39	0.73
Year 10	0.65	0.49	0.86
General School	1.02	0.79	1.31
Age	0.89	0.81	0.97
% Low-High SES in School	0.82	0.39	1.79
Treatment * Year 08	1.08	0.92	1.27
Treatment * Year 09	1.14	0.86	1.51
Treatment * Year 10	1.19	0.91	1.56
N. Observations	1461		
WAIC (similar to AIC)	4816.13		

Table 2.3: Bayesian ordinal model of Mathematics test score on covariates for Middle SES 4th graders. Estimates expressed as odds ratios.

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	1.06	0.74	1.53
Year 08	0.56	0.41	0.75
Year 09	0.47	0.33	0.66
Year 10	0.45	0.32	0.66
General School	1.51	1.08	2.11
Age	0.95	0.84	1.06
% Low-Mid SES in School	2.24	1.01	4.9
Treatment * Year 08	1.04	0.85	1.28
Treatment * Year 09	1.23	0.94	1.61
Treatment * Year 10	1.26	0.88	1.79
N. Observations	1083		
WAIC (similar to AIC)	3522.49		

Table 2.4: Bayesian ordinal model of Mathematics test score on covariates for High SES 4th graders. Estimates expressed as odds ratios.

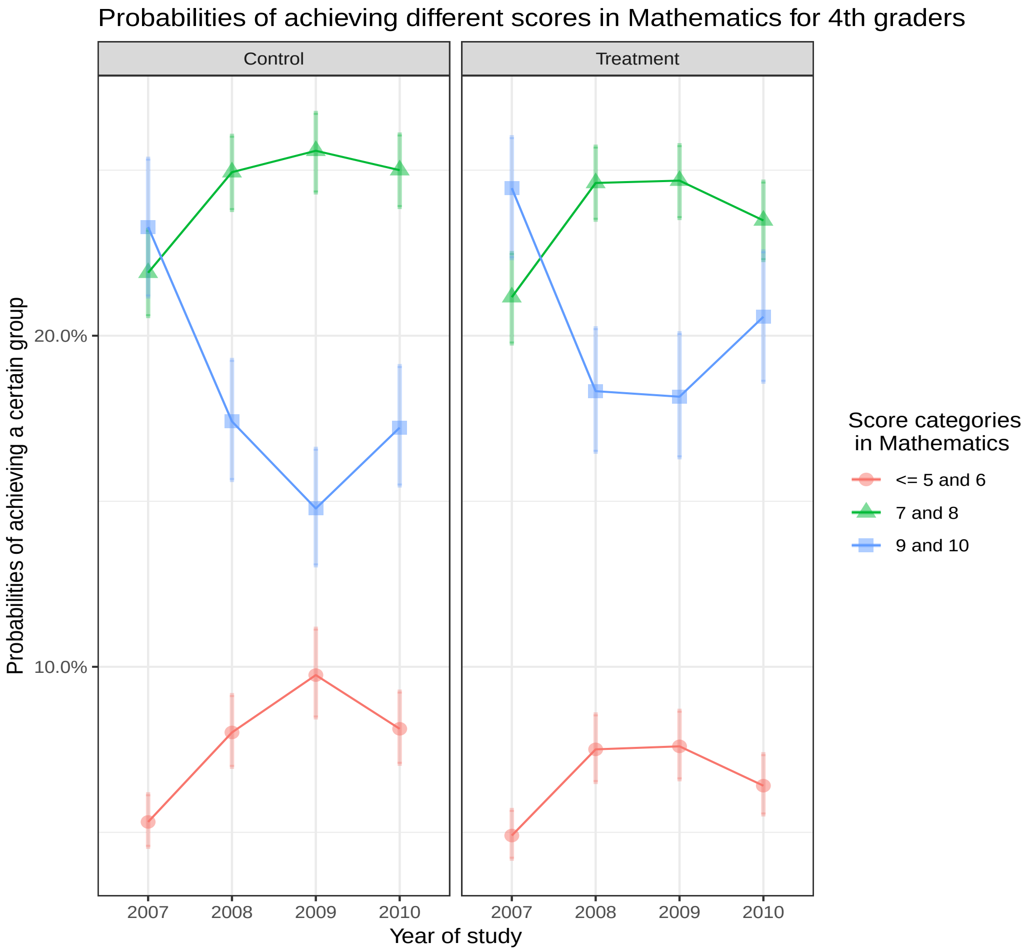


Figure 2.7: Probabilities of achieving certain grades across all years for treated and control units for 4th graders in Mathematics

Let us concentrate only one the odds ratios for the interaction. For the low SES interaction, the treatment group in 2008 are about 1.06% more likely to score higher than the control group relative to the same difference in 2007. For 2009 and 2010, the odds increased even more to 1.14% and 1.13% respectively. As it can be seen, the treatment seems to have had an effect for the low SES sample. Moving on to table 2.3, we see that the coefficients are at 1.08%, 1.14% and 1.19% respectively for 2008, 2009 and 2010, higher than the effect that low SES students had. However, they are indeed very close as the first and second year overlap with the estimates of the low SES model. Despite this, there seems to be some degree of correspondence between what we would expect: higher overall effect for the mid SES group. Finally, table 2.4 shows the same odds ratios for the high SES group. For this group, the effect was much more substantial. For the year 2008, the expected treatment effect was of about 1.04%, yet for 2009 and 2010 it was at 1.23% and 1.26%. These estimations are quite uncertain, as it can be seen from the credible intervals. However, they are no more uncertain than the ones for the other groups, where in all estimations the intervals are very wide. It is not surprising to find this much uncertainty considering the complex estimation of interaction effects with such small sample size within each group.

The results show here seem to point out that there was indeed an SES gradient in the impact of the program. Despite this, they are not bullet-proof and are accompanied by a great deal of uncertainty. This is something to account for when replicating similar designs in order to test for this type of effect in future research. In order to provide an alternative visualization of these models, I compute the predicted probabilities for each model. Figure 2.8, Figure 2.9, Figure 2.10 show the predicted probabilities respectively for the low, mid and high SES models.

Let's begin with figure 2.8. The Y axis presents the predicted probabilities of achieving a certain grade. The red line refers to the probabilities of achieving less than or equal to 5 and 6, the green line the probabilities of achieving a 7 or an 8 and the blue line the probabilities of achieving a 9 and a 10. As can be seen between the right and left panel, the effect for the low SES group is not that evident. The blue line seems to be higher for the last two years, but the probabilities are not that much apart. The same is evident for the red line, the chances of scoring in the bottom marks. This lines up as expected with the model, which showed an uncertain yet small effect.

Moving on to the mid SES plot, figure 2.9 shows a stronger effect. Treated students from the mid SES group are more likely to score in the higher marks in the last two years than the control group. Conversely, treated students seem less likely to score in the bottom marks, particularly in the last year. This lines up as well with the model coefficients. However, it should be clarified that the credible intervals overlap in some of these probabilities. This means that the results carry a great deal of uncertainty which should be considered when making final conclusions.

Finally, figure 2.10 shows the results for the high SES group. We see an even stronger effect for scoring in the top marks. There is a clear trend showing on how treated high SES students stopped decreasing their probabilities for the high marks (blue line) in the last two years. Conversely, the control units decreased their probabilities steeply. These

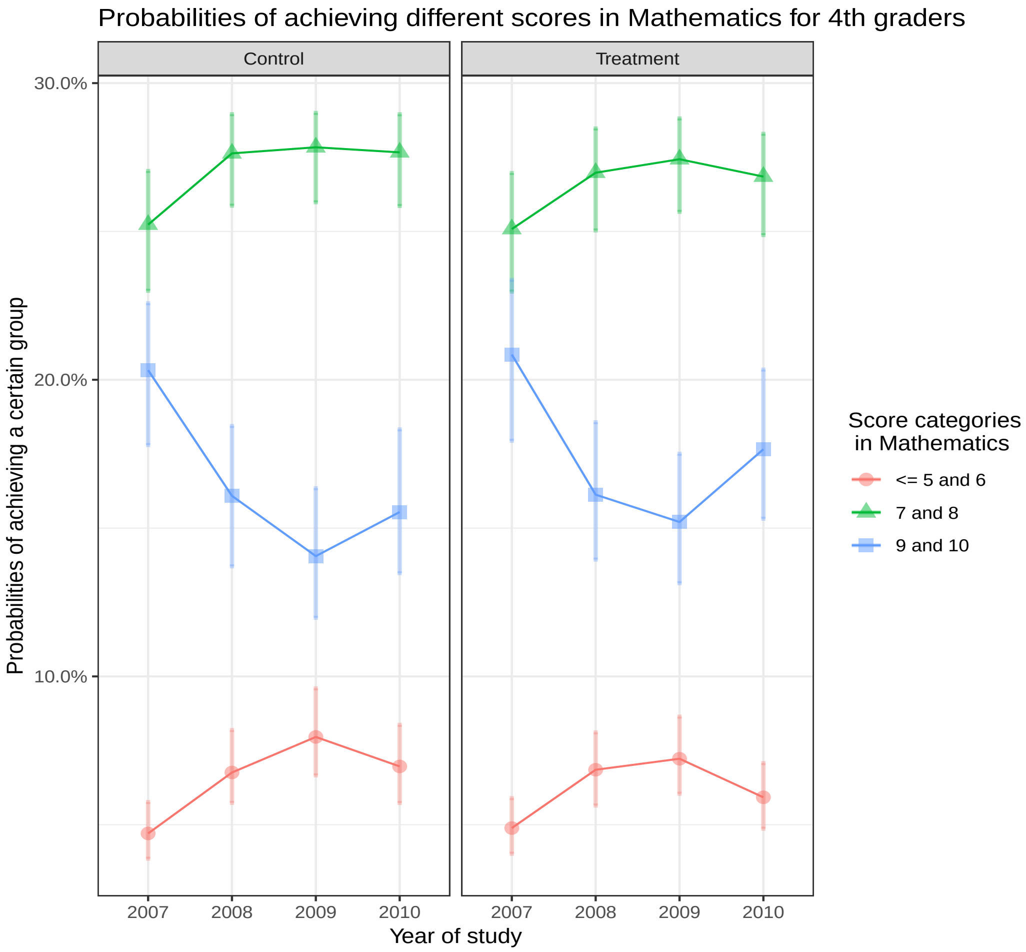


Figure 2.8: Probabilities of achieving certain grades for Low SES students between treated and control units in Mathematics for 4th graders

results are reversed when scoring in the low SES marks. The chances of scoring in the low marks seemed to have stopped for the treated students relative to the control group and decreased in the last year relative to the control group.

These plots confirm the second hypothesis that states that the treatment effect was moderated by the SES origin of the students. These results show indeed that the benefit that a child will receive from this treatment is related to the education and training of the parents. This is true because we assume that all students receive similar education at school before the treatment due to the randomization.

As robustness check, I run the initial model for test scores in Natural Science and Spanish in table 2.7 and table 2.9 for 4th graders in the appendix. The results hold and look very similar to the ones showed for Mathematics, which provide robustness to the

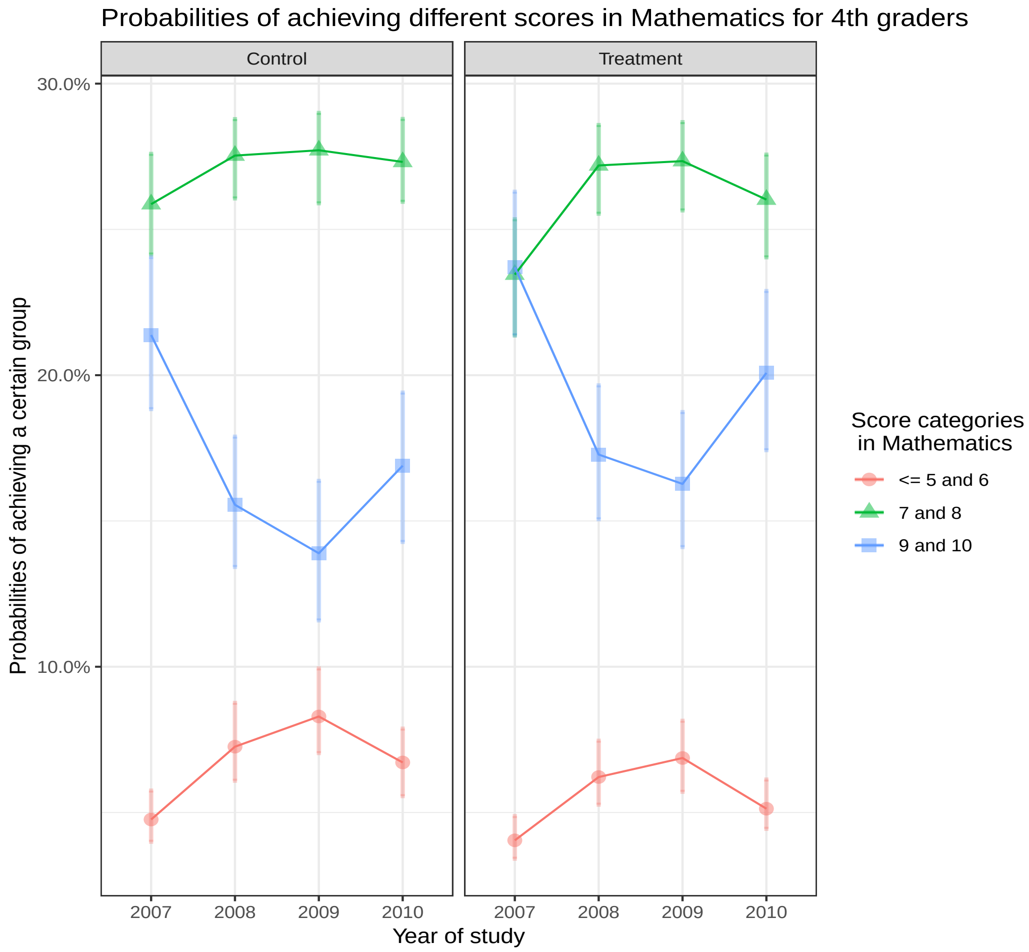


Figure 2.9: Probabilities of achieving certain grades for Middle SES students between treated and control units in Mathematics for 4th graders

results.

Figure 2.1, figure 2.2, table 2.6, table 2.8 and table 2.10 in the appendix replicate the analysis for 5th graders and the results are mixed. For example, figure 2.2 shows that the distribution of the ATE is most likely to be positive yet on the other hand, figure 2.1 does not show strong support for the interaction as all interaction terms seem to be as likely to positive and negative with wide intervals. The results point out that the model does not fit the data very well for 5th graders but does very well for 4th graders. However, the treatment effect seems to be quite substantial for 4th graders and the effect seems to be moderated by the SES origin of the students. It is puzzling why this effect is not as strong or robust for 5th graders despite them sharing the same school and teachers.

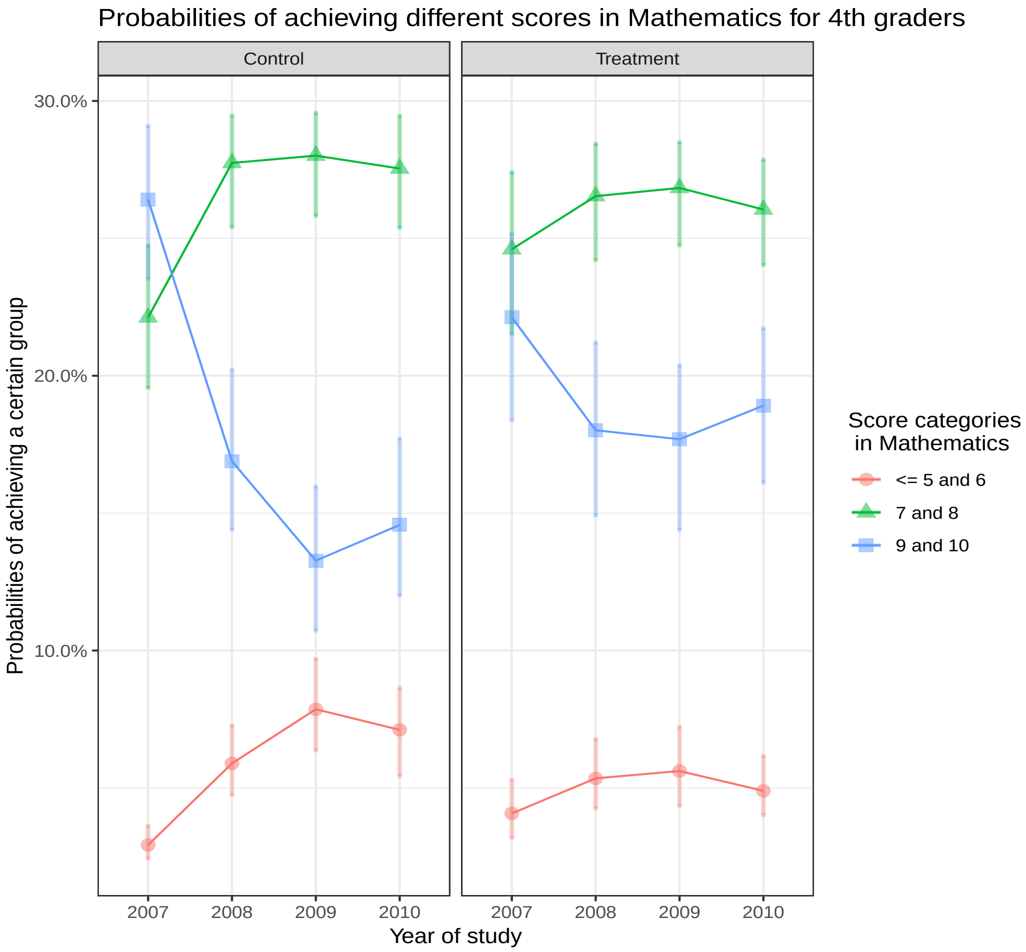


Figure 2.10: Probabilities of achieving certain grades for High SES students between treated and control units in Mathematics for 4th graders

2.6 Limitations

This study has several limitations which need to be discussed. First, the model defined for the estimation has slight limitations in terms of predicting the top two marks in the test. I do not expect for the model to change dramatically if the predictions become more precise in the top marks. The previous is evidenced by the fact that the actual and posterior distribution of the outcome variable are very close even though they differ slightly.

Another possible concern is whether children exaggerated their scores in the test since the question was asked directly to them. In principle, there is nothing that can be done about that. However, even if they did that, the distribution of the scores seems fairly natural with no over concentration in the top marks, the expected behavior if children were overestimating their scores. All in all, the main outcome has considerable measurement error that I attempt to remedy by providing a likely posterior distribution based on theoretical and empirical insights in the literature of the topic.

One key mechanism that the text has not adjusted for is the concept of dosage effects. For example, perhaps the reason for the lack of effects in the first year of the treatment – 2008 – could be because effects accumulate with effects of exposure. All children in 2008 had spent a single year under the program. However, 4th and 5th graders in 2009 would have been exposed twice (in 2009 but also as 3rd and 4th graders in 2008). Similarly, 4th and 5th graders in 2010 would have been exposed three times. The implications of the change in dose over the years is an important component and could explain why there is no effect in 2008, for example. Further research should extend these results to include some type of panel effect that accounts for the dosage of the treatment received by each student. In a similar line, all the models presented in the paper account for both clustering within schools and within years. However, the models does not specifically account for the fact that schools are present in all years as well. This means that the model could be overestimating the effects slightly and should be taken into account for further research.

Another limitation, although not very serious is that the design does not have panel data. The results cannot make a definitive causal claim at the individual level but are indeed valid when aggregating between years and inside specific groups such as the SES origin of the students.

2.7 Conclusion

The present study focuses on the importance of cognitive skills in a World Bank policy that aims to increase parental involvement in school's decision making. 250 schools were chosen among the universe of schools in Mexico in four different provinces to participate in the experiment. The parental associations of these 250 schools were receiving $600 dollars in Mexican pesos to motivate parents to get involved in the decision making process of the school. The experiment under study increased the $600 dollars to $1200 for 125 schools and left the remaining 125 schools with the initial $600 dollars. With this money parents were allowed to improve school infrastructure, make teachers accountable, program school activities and buy resources for the students. Both groups of parents were already receiving training to improve their interactions with their children and be able to better help them with their homework. Empirically, the treatment can be defined as the incentive for parents to get more involved at school and with their children. However, theoretically, students should also benefit in terms of their school performance as it is expected for parents to help them with their homework and school duties as they get more involved at school.

Using the the baseline year in 2007 which had 737 students, I estimate a phased-in treatment effect for years 2008, 2009 and 2010 which had 1,179, 697 and 1,027 students correspondingly. Using a fully bayesian ordinal multilevel regression model, the results show that the differences between treated and control units for 2007 and 2008 are negligible, which was expected as the treatment needs time to phase in, but for the last two years the model estimates that treated units have an odds increase of about $\sim 13\%$ - 17% of scoring higher test scores than the control units.

More concretely, the probabilities of scoring higher test scores are concentrated in the top two marks of the test while the control group is increasing their probability of scoring the bottom two marks. Once these results are broken down by groups of test scores, the pattern of the evolution is interesting. For both treated and control units the probability of increasing in the bottom marks is increasing with time. As the treatment fades in the treated group stabilizes the probabilities and even starts to decrease in the last year. Conversely, the control schools keep increasing their probabilities of landing in the bottom marks of the test. The exact opposite pattern happens for the top marks in the test. Control units are decreasing their probabilities of landing in the top marks while the treated schools seem to stabilize and increase their probabilities to score a top mark in the last years. Although it should not matter because of the randomized design, these results hold even after including several sets of control variables which look to make the estimates more precise.

As argued throughout the chapter, I develop a possible mechanism through which these results come about. First, the mechanisms through which cognitive abilities increased were not directly because parent's participated more in school activities. This might have had positive relevance in parent-teacher relations but not particularly on test performance or cognitively enriching activities. However, it might have increased due to the indirect spillover effect of parental participation towards more quality time with the students in tasks such as homework or reading. This is a possible explanation considering most parent-child developmental research finds that increasing such activities benefits greatly students in their early ages (Waldfogel, 2006). This mechanism becomes more plausible when we consider that parent's also received training on how to help their children in academic activities as a broader initiative of the ongoing program. Note that both treatment and control groups received the same training but perhaps school involvement triggered other improvements related to activities at home.

These estimations are applied in Mathematics, Spanish and Natural Sciences for 4th graders and all results point in the same direction. The chapter also explores whether the treatment effect was moderated by the SES origin of the student. The results show that for treated schools, high SES students benefited more from the treatment as they had higher chances of scoring in the top marks than low SES students. That is, receiving exactly the same treatment, high SES students saw larger probability gains than lower students in terms of scoring in the top groups. More concretely, treated high SES students had an odds increase of higher grades of about 23%-26% and the same low SES students in the same years saw an increase in odds of about 14%-13%. Conversely, high SES students saw smaller probabilities of scoring in the bottom marks than their low SES counterparts. As all students were receiving the same treatment and were very similar in over 190 characteristics at the school-level, the likely explanation that I provide is related to the interactions with their parents.

But more concretely, why should this effect be different between high and low SES parent? There is evidence from the parent-child interaction literature which suggests that it is more effective to increase involvement for the low SES parents because it is much difficult to change (Domina, 2005). In the same line, there is evidence from (Park and Holloway, 2017) which shows that parental involvement with their children was more strongly related to high SES children than low SES children. There are several meta-analysis which suggest that this is the case such as Wang and Sheikh-Khalil (2014) and Dearing et al. (2006). However, the strongest evidence comes from a recent study by Boonk et al. (2018) which evaluated 75 studies (ranging from before 2000's to more recent ones) on parent-child interactions both from the psychological and sociological perspective and documented findings that high educated mothers were generally more effective in their involvement activities in terms of academic achievement:

> ”...According to these findings higher educated mothers are in general more successful in their involvement activities compared to lower educated mothers. This could be explained by more effective involvement skills of educated mothers (e.g., Englund, Luckner, Whaley, Egeland, 2004; Fekonja-Peklaj, Marjanovic-Umek, Kranjc, 2010).”

The results from this paper serve to highlight that training parents is a well intentioned effort. However, different parents need different trainings. This was deemed evident by Deming (2009) in programs such as Head start but the lesson is not yet standard in policymaking. As can be interpreted from the results presented here, high SES students probably received better feedback and support from their parents as opposed to the low SES students. These effects are in the expected direction and showcase the fact that when implementing policies aimed at improving the well being of students and their schools, we need to consider that not everyone is at the same playing field and policymakers need to develop more specialized solutions targeted towards specific populations.

However, it should be noted that all of these results carry a great deal of uncertainty and should not be interpreted without caution. Further research should attempt to follow up on similar results to give more credibility to the findings. In fact, assuming these results hold and are replicated under more stringent specifications, then it should be imperative for reformers to actually re-conceptualize SBM to stimulate a more active approach by parents, specially low educated parents. What I mean by an active approach is to not just be active in school related activities, but to actually train parents to know how to treat and behave around their children to help them reach their best potential.

As a robustness check, the study replicates all of the findings for 5th graders and finds that the results are mixed. There is no clear interaction effect but there is an average increase in probabilities for the treated group. However, these are not as clear as for 4th graders.

The results from this chapter are both interesting and puzzling. In line with the findings by Barros and Mendonca (1998) which find that SBM had no impact on test scores, I find that the treatment had no effect on 5th graders. However, for 4th graders, the results line up with the work of Gertler et al. (2012). Gertler et al. (2012) also found some puzzling effects when comparing consecutive grade-year comparison. That is, they found an effect for 3rd graders, which in the next year of the experiment were not completely evident for 4th graders. We should expect this considering that a similar composition of 3rd graders was the one that passed to 4th grade the next year. These are puzzling questions that should motivate future research on the nature and scaling of these programs.

Further research should concentrate on exploring which specific activities in SBM relate to the findings presented here. For example, the data from this study has information on the parent-children interactions in terms of reading and revising homework. On top of that, all children can be linked to their school questionnaire which has hundreds of questions related to the quality of the school's infrastructure, the teacher's education, the frequency in which parent's got involved in the parental meetings, among other things. Finally, follow-up research should also disentangle why the effect is clear and robust for one grade but not for another. Some research points out that parental participation in 5th grade was not as high as in 4th grade (Gertler et al., 2012) and it could explain why the interaction is not as clear.

This chapter finds that an increase in funding for the parent's association helped children improve Mathematics grades, however this improvement was mediated by the SES origin of the student. The paper hypothesizes that this unequal effect was due to parental involvement at home, which given the difference in education between SES groups, helped some groups more than others. Given that this mechanism is still speculative, further research should attempt to test it directly. If it is the case that this mechanism is indeed playing an important role, then it is important for educational reformers to take a new look at this type of policy and make parents more capable of giving their children the support they deserve. Receiving a good education at home is one of the most important assets a child can have and concentrating on improving the education of parents should be an imperative topic in a policymaker's agenda.

Bibliography

Acevedo, Gladys Lopez and Vicente Paqueo (2003). *Supply-Side School Improvement and the Learning Achievement of the Poorest Children in Indigenous and Rural Schools: the Case of PARE*. Vol. 3172. World Bank Publications.

Altschul, Inna (2011). "Parental involvement and the academic achievement of Mexican American youths: What kinds of involvement in youths' education matter most?" In: *Social Work Research* 35.3, pp. 159–170.

Anderson, Laurie M, Carolynne Shinn, Mindy T Fullilove, Susan C Scrimshaw, Jonathan E Fielding, Jacques Normand, Vilma G Carande-Kulis, Task Force on Community Preventive Services, et al. (2003). "The effectiveness of early childhood development programs: A systematic review." In: *American journal of preventive medicine* 24.3, pp. 32–46.

Barnard, John and Donald B Rubin (1999). "Miscellanea. Small-sample degrees of freedom with multiple imputation." In: *Biometrika* 86.4, pp. 948–955.

Bollen, Kenneth A., Jennifer L. Glanville, and Guy Stecklov (Aug. 2001). "Socioeconomic Status and Class in Studies of Fertility and Health in Developing Countries." In: *Annual Review of Sociology* 27.1, pp. 153–185. DOI: 10.1146/annurev.soc.27.1.153.

Boonk, Lisa, Hieronymus JM Gijselaers, Henk Ritzen, and Saskia Brand-Gruwel (2018). "A review of the relationship between parental involvement indicators and academic achievement." In: *Educational Research Review* 24, pp. 10–30.

Brant, Rollin (1990). "Assessing proportionality in the proportional odds model for ordinal logistic regression." In: *Biometrics*, pp. 1171–1178.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos (Feb. 11, 2011). *Making Schools Work: New Evidence on Accountability Reforms*. World Bank Publishing. 251 pp. ISBN: 0821386794. URL: https://www.ebook.de/de/product/14536372/

barbara_bruns_deon_filmer_harry_anthony_patrinos_making_schools_work_new_evidence_on_accountability_reforms.html.

Bürkner, Paul-Christian and Matti Vuorre (2018). “Ordinal regression models in psychological research: A tutorial.” In:

Buuren, S van and Karin Groothuis-Oudshoorn (2010). “mice: Multivariate imputation by chained equations in R.” In: *Journal of statistical software*, pp. 1–68.

Campbell, Frances A, Craig T Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson (2002). “Early childhood education: Young adult outcomes from the Abecedarian Project.” In: *Applied developmental science* 6.1, pp. 42–57.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017). “Stan: A probabilistic programming language.” In: *Journal of statistical software* 76.1.

Christensen, Rune Haubo Bojesen (2015). “ordinal-Regression Models for Ordinal Data. R package version 2015.6-28.” In: *See http://www. cran. r-project. org/package= ordinal.*

Collaboration, Open Science et al. (2015). “Estimating the reproducibility of psychological science.” In: *Science* 349.6251, aac4716.

Cunha, Flavio, James Heckman, Lance Lochner, and Dimitriy V Masterov (2006). “Interpreting the evidence on life cycle skill formation.” In: *Handbook of the Economics of Education* 1, pp. 697–812.

Deming, David (2009). “Early childhood intervention and life-cycle skill development: Evidence from Head Start.” In: *American Economic Journal: Applied Economics* 1.3, pp. 111–134.

Di Gropello, Emanuela (2006). *A comparative analysis of school-based management in Central America*. 72. World Bank Publications.

Epstein, Joyce (2018). *School, Family, and Community Partnerships, Student Economy Edition: Preparing Educators and Improving Schools*. Routledge.

Epstein, Joyce L, Mavis G Sanders, Steven B Sheldon, Beth S Simon, Karen Clark Salinas, Natalie Rodriguez Jansorn, Frances L Van Voorhis, Cecelia S Martin, Brenda G Thomas, Marsha D Greenfeld, et al. (2018). *School, family, and community partnerships: Your handbook for action*. Corwin Press.

Epstein, Joyce L and Steven B Sheldon (2002). “Present and accounted for: Improving student attendance through family and community involvement.” In: *The Journal of Educational Research* 95.5, pp. 308–318.

Erikson, Robert and John Goldthorpe (1992). *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford University Press. ISBN: 978-0198273837. URL: https://www.amazon.com/Constant-Flux-Mobility-Industrial-Societies/dp/0198273835?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0198273835.

Fan, Weihua, Cathy M Williams, and Christopher A Wolters (2012). "Parental involvement in predicting school motivation: Similar and differential effects across ethnic groups." In: *The Journal of Educational Research* 105.1, pp. 21–35.

Fullan, Michael and Nancy Watson (Dec. 2000). "School-Based Management: Reconceptualizing to Improve Learning Outcomes." In: *School Effectiveness and School Improvement* 11.4, pp. 453–473. DOI: 10.1076/sesi.11.4.453.3561.

Galindo, Claudia and Steven B Sheldon (2012). "School and home connections and children's kindergarten achievement gains: The mediating role of family involvement." In: *Early Childhood Research Quarterly* 27.1, pp. 90–103.

Gelman, Andrew (2013). "Commentary: P values and statistical practice." In: *Epidemiology* 24.1, pp. 69–72.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su, et al. (2008). "A weakly informative default prior distribution for logistic and other regression models." In: *The Annals of Applied Statistics* 2.4, pp. 1360–1383.

Gelman, Andrew and Eric Loken (2013). "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time." In: *Department of Statistics, Columbia University*.

Gelman, Andrew and Cosma Rohilla Shalizi (2013). "Philosophy and the practice of Bayesian statistics." In: *British Journal of Mathematical and Statistical Psychology* 66.1, pp. 8–38.

Gertler, Paul, Harry Anthony Patrinos, and Eduardo Rodrıguez-Oreggia (2012). "Parental Empowerment in Mexico: Randomized Experiment of the "Apoyo a La Gestion Escolar (AGE)" Program in Rural Primary Schools in Mexico." In: *Society for Research on Educational Effectiveness*.

Hanushek, Eric and Ludger Woessmann (Feb. 2007). *The Role Of Education Quality For Economic Growth*. The World Bank. DOI: 10.1596/1813-9450-4122.

Heckman, James (2006). “Skill formation and the economics of investing in disadvantaged children.” In: *Science* 312.5782, pp. 1900–1902.

— (2013). *Giving Kids a Fair Chance (Boston Review Books)*. The MIT Press. ISBN: 9780262019132. URL: https://www.amazon.com/Giving-Chance-Boston-Review-Books/dp/0262019132?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0262019132.

Heckman, James and Tim Kautz (2013). *Fostering and measuring skills: Interventions that improve character and cognition*. Tech. rep. National Bureau of Economic Research.

Heckman, James and Alan B. Krueger (2004). *Inequality in America: What Role for Human Capital Policies? (Alvin Hansen Symposium Series on Public Policy)*. The MIT Press. ISBN: 9780262083287. URL: https://www.amazon.com/Inequality-America-Capital-Policies-Symposium/dp/0262083280?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0262083280.

Hill, Nancy E, Domini R Castellino, Jennifer E Lansford, Patrick Nowlin, Kenneth A Dodge, John E Bates, and Gregory S Pettit (2004). “Parent academic involvement as related to school behavior, achievement, and aspirations: Demographic variations across adolescence.” In: *Child development* 75.5, pp. 1491–1509.

Hill, Nancy E and Lorraine C Taylor (2004). “Parental school involvement and children’s academic achievement: Pragmatics and issues.” In: *Current directions in psychological science* 13.4, pp. 161–164.

Jimenez, Emmanuel and Yasuyuki Sawada (1999). “Do community-managed schools work? An evaluation of El Salvador’s EDUCO program.” In: *The world bank economic review* 13.3, pp. 415–441.

Jordan, Will J and Stephen B Plank (2000). “Talent loss among high-achieving poor students.” In: *Schooling students placed at risk: Research, policy, and practice in the education of poor and minority adolescents*, pp. 83–108.

King, Elizabeth M and Susana Cordeiro Guerra (2005). “Education reforms in East Asia: Policy, process, and impact.” In: *East Asia decentralizes: Making local government work* 292, p. 179.

Kolenikov, Stanislav and Gustavo Angeles (Mar. 2009). “Socioeconomic status measurement with discrete proxy variables: is Principal Component Analysis a reliable answer?" In: *Review of Income and Wealth* 55.1, pp. 128–165. DOI: 10.1111/j.1475-4991.2008.00309.x.

Leithwood, Kenneth and Teresa Menzies (May 1998). "Forms and Effects of School-Based Management: A Review." In: *Educational Policy* 12.3, pp. 325–346. DOI: 10.1177/0895904898012003006.

Long, Scott J, J Scott Long, and Jeremy Freese (2006). *Regression models for categorical dependent variables using Stata*. Stata press.

Magnuson, Katherine, Christopher Ruhm, and Jane Waldfogel (2007). "Does prekindergarten improve school preparation and performance?" In: *Economics of Education review* 26.1, pp. 33–51.

OECD (Dec. 2013). *Strong Performers and Successful Reformers in Education: Lessons from PISA 2012 for the United States*. OECD Publishing. DOI: 10.1787/9789264207585-en.

Patall, Erika A, Harris Cooper, and Jorgianne Civey Robinson (2008). "Parent involvement in homework: A research synthesis." In: *Review of educational research* 78.4, pp. 1039–1101.

Patrinos, H (2006). "Mexico: AGE (Apoyo a la Gestión Escolar)–School management support: a qualitative assessment." In: *Processed, World Bank, Washington DC*.

Patrinos, Harry Anthony and Tazeen Fasih (2009). *Decentralized decision-making in schools: The theory and evidence on school-based management*. World Bank Publications.

Phillipson, Sivanes and Shane N Phillipson (2012). "Children's cognitive ability and their academic achievement: The mediation effects of parental expectations." In: *Asia Pacific Education Review* 13.3, pp. 495–508.

Raikes, Helen, Beth L Green, Jane Atwater, Ellen Kisker, Jill Constantine, and Rachel Chazan-Cohen (2006). "Involvement in Early Head Start home visiting services: Demographic predictors and relations to child and parent outcomes." In: *Early Childhood Research Quarterly* 21.1, pp. 2–24.

Reid, M. Jamila, Carolyn Webster-Stratton, and Theodore P. Beauchaine (2001). "Parent Training in Head Start: A Comparison of Program Response Among African American, Asian American, Caucasian, and Hispanic Mothers." In: *Prevention Science* 2.4, pp. 209–227. DOI: 10.1023/a:1013618309070.

Rogers, Maria A, Jennifer Theule, Bruce A Ryan, Gerald R Adams, and Leo Keating (2009). "Parental involvement and children's school achievement: Evidence for mediating processes." In: *Canadian Journal of School Psychology* 24.1, pp. 34–57.

Sahlberg, Pasi (Mar. 2007). "Education policies for raising student learning: the Finnish approach." In: *Journal of Education Policy* 22.2, pp. 147–171. DOI: 10.1080/02680930601158919.

Schweinhart, Lawrence J. and David P. Weikart (Oct. 1981). "Effects of the Perry Preschool Program on Youths Through Age 15." In: *Journal of the Division for Early Childhood* 4.1, pp. 29–39. DOI: 10.1177/105381518100400105.

Shapiro, Joseph and Jorge Moreno Trevino (2004). *Compensatory education for disadvantaged Mexican students: An impact evaluation using propensity score matching.* Vol. 3334. World Bank Publications.

Sheldon, Steven B and Joyce L Epstein (2005). "Involvement counts: Family and community partnerships and mathematics achievement." In: *The Journal of Educational Research* 98.4, pp. 196–207.

Sheridan, Susan M, Tyler E Smith, Elizabeth Moorman Kim, S Natasha Beretvas, and Sunyoung Park (2019). "A Meta-Analysis of Family-School Interventions and Children's Social-Emotional Functioning: Moderators and Components of Efficacy." In: *Review of Educational Research*, p. 0034654318825437.

Taylor, Dianne L and Charles Teddlie (1992). "Restructuring and the Classroom: A View from a Reform District." In:

Vyas, S. and L. Kumaranayake (Aug. 2006). "Constructing socio-economic status indices: how to use principal components analysis." In: *Health Policy and Planning* 21.6, pp. 459–468. DOI: 10.1093/heapol/czl029.

Waldfogel, Jane (2006). "What do children need?" In: *Juncture* 13.1, pp. 26–34.

Wang, Ming-Te and Salam Sheikh-Khalil (2014). "Does parental involvement matter for student achievement and mental health in high school?" In: *Child development* 85.2, pp. 610–625.

Xu, Min, Susan N Kushner Benson, Renee Mudrey-Camino, and Richard P Steiner (2010). "The relationship between parental involvement, self-regulated learning, and reading achievement of fifth graders: A path analysis using the ECLS-K database." In: *Social Psychology of Education* 13.2, pp. 237–269.

2.8 Appendix

	Treatment		Control	
Variables	Count	Percentage	Count	Percentage
Number of sleeping rooms at home				
- 1	1343	0.33	1548	0.37
- 2	1394	0.34	1342	0.32
- 3	773	0.19	755	0.18
- 4	368	0.09	344	0.08
- 5 or more	164	0.04	189	0.05
Number of people living in your house				
- 2	162	0.04	166	0.04
- 3	177	0.04	178	0.04
- 4	421	0.10	438	0.11
- 5	694	0.17	687	0.17
- 6	730	0.18	704	0.17
- 7	586	0.14	687	0.17
- 8	431	0.11	444	0.11
- 9	346	0.08	324	0.08
- 10 or more	526	0.13	519	0.13
Father's education				
- University	232	0.06	285	0.07
- High school	218	0.05	278	0.07
- Secondary school	599	0.15	615	0.15
- Full primary	713	0.17	776	0.18
- Didnt finish primary	701	0.17	607	0.14
- Didnt go to school	259	0.06	268	0.06
- I dont know	1364	0.33	1393	0.33
Number of cars				
- None	2783	0.73	3080	0.77
- 1	733	0.19	686	0.17
- 2	184	0.05	167	0.04
- 3	120	0.03	85	0.02
Number of cellphones				
- None	2515	0.68	2564	0.68
- 1	852	0.23	895	0.24
- 2	188	0.05	212	0.06
- 3	117	0.03	113	0.03
Number of DVDs				
- None	2102	0.54	2110	0.52
- 1	1535	0.40	1700	0.42
- 2	173	0.04	178	0.04
- 3	71	0.02	66	0.02
Number of computers				
- None	3446	0.90	3512	0.88
- 1	286	0.07	371	0.09
- 2	67	0.02	77	0.02
- 3	31	0.01	43	0.01

Table 2.1: Descriptives of selected variable in 2007 used in PCA by treatment and control indicator for 4th graders

	Treatment		Control	
Variables	Count	Percentage	Count	Percentage
Number of sleeping rooms at home				
- 1	1148	0.30	1219	0.33
- 2	1445	0.38	1345	0.37
- 3	758	0.20	700	0.19
- 4	291	0.08	261	0.07
- 5 or more	165	0.04	141	0.04
Number of people living in your house				
- 2	84	0.02	53	0.01
- 3	138	0.04	149	0.04
- 4	415	0.11	426	0.12
- 5	715	0.19	683	0.19
- 6	636	0.17	647	0.18
- 7	554	0.15	554	0.15
- 8	444	0.12	415	0.11
- 9	327	0.09	331	0.09
- 10 or more	441	0.12	409	0.11
Father's education				
- University	175	0.05	140	0.04
- High school	172	0.05	193	0.05
- Secondary school	557	0.15	594	0.16
- Full primary	772	0.20	610	0.17
- Didnt finish primary	609	0.16	572	0.16
- Didnt go to school	279	0.07	169	0.05
- I dont know	1214	0.32	1388	0.38
Number of cars				
- None	2755	0.74	2823	0.77
- 1	718	0.19	604	0.17
- 2	184	0.05	163	0.04
- 3	86	0.02	65	0.02
Number of cellphones				
- None	2143	0.58	2209	0.61
- 1	875	0.24	845	0.23
- 2	365	0.10	286	0.08
- 3	313	0.08	276	0.08
Number of DVDs				
- None	1816	0.49	1833	0.50
- 1	1606	0.43	1529	0.42
- 2	223	0.06	216	0.06
- 3	73	0.02	59	0.02
Number of computers				
- None	3304	0.89	3264	0.90
- 1	320	0.09	284	0.08
- 2	48	0.01	50	0.01
- 3	37	0.01	20	0.01

Table 2.2: Descriptives of selected variable in 2010 used in PCA by treatment and control indicator for 4th graders

	2007		2008		2009		2010	
Variables	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage
Treatment								
- Treated	862	0.52	782	0.49	707	0.49	706	0.51
- Control	791	0.48	820	0.51	738	0.51	689	0.49
SES categories								
- Low SES	469	0.28	454	0.28	435	0.30	441	0.32
- Middle SES	689	0.42	662	0.41	581	0.40	576	0.41
- High SES	495	0.30	486	0.30	429	0.30	378	0.27
Type of school								
- General	935	0.57	896	0.56	770	0.53	773	0.55
- Indigenous	718	0.43	706	0.44	675	0.47	622	0.45
Type of grade								
- 5th grade	1653	1.00	1602	1.00	1445	1.00	1395	1.00
Age								
- 7	0	0.00	0	0.00	0	0.00	0	0.00
- 8	0	0.00	0	0.00	0	0.00	0	0.00
- 9	0	0.00	0	0.00	0	0.00	0	0.00
- 10	784	0.47	255	0.16	667	0.46	338	0.24
- 11	486	0.29	767	0.48	519	0.36	693	0.50
- 12	251	0.15	364	0.23	153	0.11	225	0.16
- 13	98	0.06	148	0.09	74	0.05	95	0.07
- 14	24	0.01	50	0.03	25	0.02	35	0.03
- 15	10	0.01	18	0.01	7	0.00	9	0.01

Table 2.3: Descriptives of main variables used in the analysis for 5th graders

Variables	X2 - Chisq	Degrees of Freedom	p value
Omnibus	123.67	32.00	0.00
Treatment	10.36	4.00	0.03
Year 08	48.06	4.00	0.00
Year 09	15.54	4.00	0.00
Year 10	49.36	4.00	0.00
SES Mid	6.42	4.00	0.17
SES High	4.89	4.00	0.30
General School	6.52	4.00	0.16
Age	5.27	4.00	0.26

Table 2.5: Brant test for proportional odds assumption of Mathematics model for 4th graders

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	1	0.74	1.36
Year 08	0.58	0.47	0.71
Year 09	0.47	0.36	0.61
Year 10	0.54	0.43	0.67
Middle SES	1.2	1.06	1.37
High SES	1.33	1.14	1.55
General School	1.19	0.94	1.49
Age	0.95	0.9	1
% Mid-High SES in School	0.73	0.46	1.14
Treatment * Year 08	1.11	0.82	1.5
Treatment * Year 09	1.37	0.99	1.92
Treatment * Year 10	1.32	0.97	1.79
N. Observations	3640		
WAIC (similar to AIC)	11854.09		

Table 2.4: Bayesian ordinal model of Mathematics test score on covariates for 4th grade. Estimates expressed as odds ratios.

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	1.28	1.04	1.58
Year 08	0.53	0.46	0.62
Year 09	0.41	0.35	0.48
Year 10	0.46	0.39	0.53
Middle SES	1.15	1.04	1.28
High SES	1.28	1.13	1.44
General School	1.01	0.83	1.22
Age	0.94	0.9	0.98
% Mid-High SES in School	0.71	0.49	1.01
Treatment * Year 08	0.85	0.68	1.04
Treatment * Year 09	0.96	0.77	1.21
Treatment * Year 10	0.89	0.71	1.1
N. Observations	6095		
WAIC (similar to AIC)	19610.4		

Table 2.6: Bayesian ordinal model of Mathematics test score on covariates for 5th grade. Estimates expressed as odds ratios.

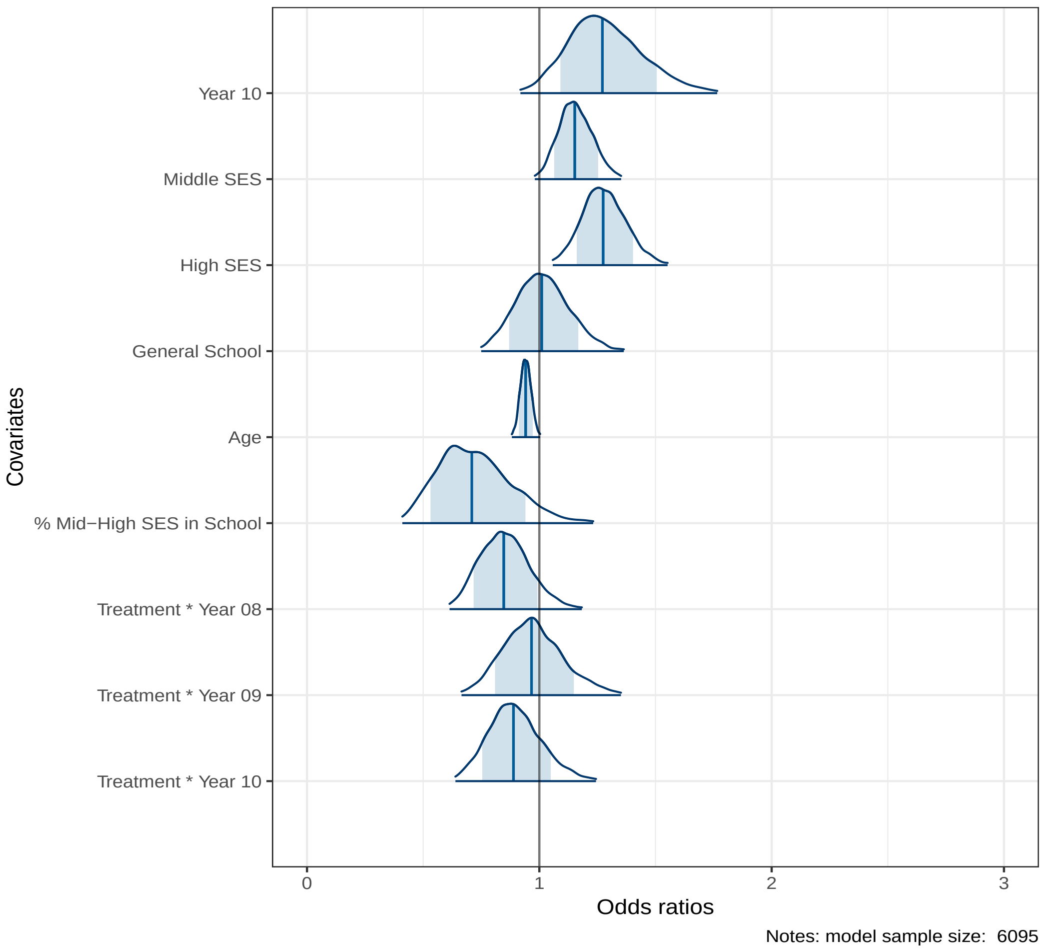


Figure 2.1: Distribution of the coefficients of the model expressed as odd ratios for 5th grade

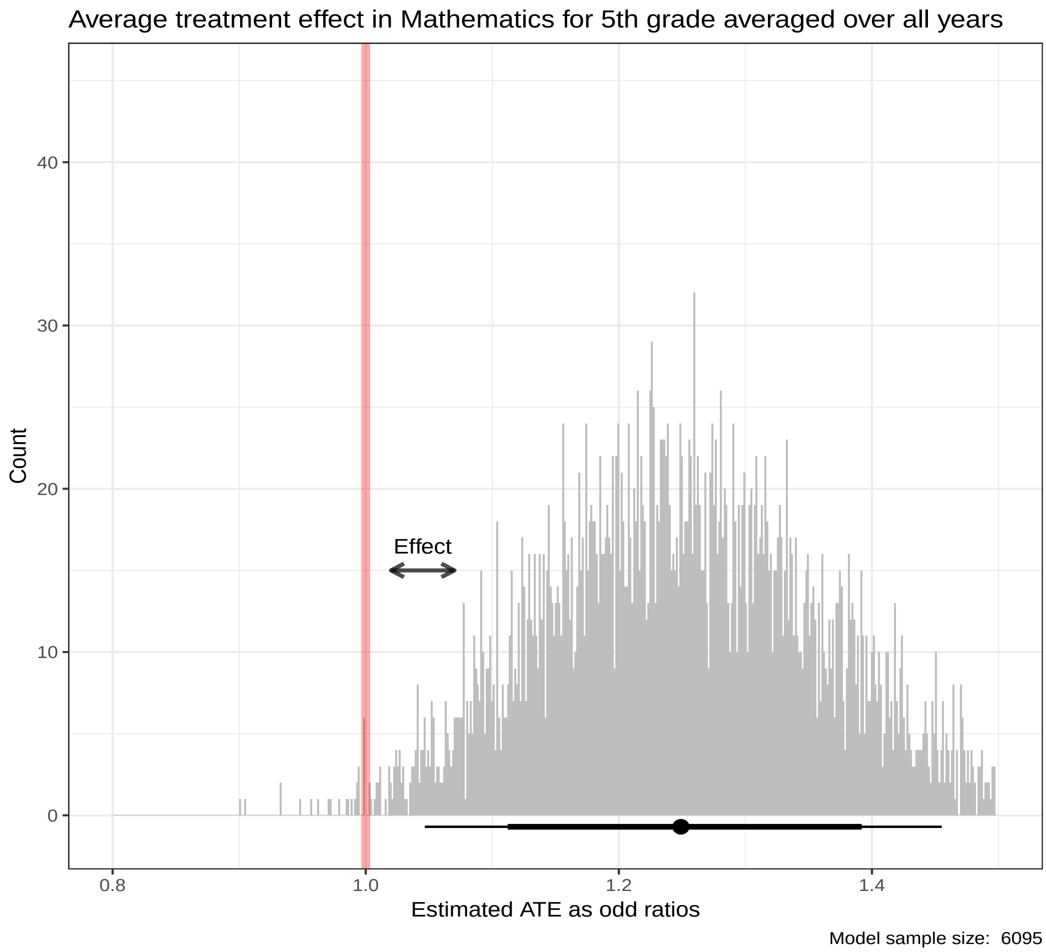


Figure 2.2: Treatment effect for Mathematics averaged over all years for 5th graders

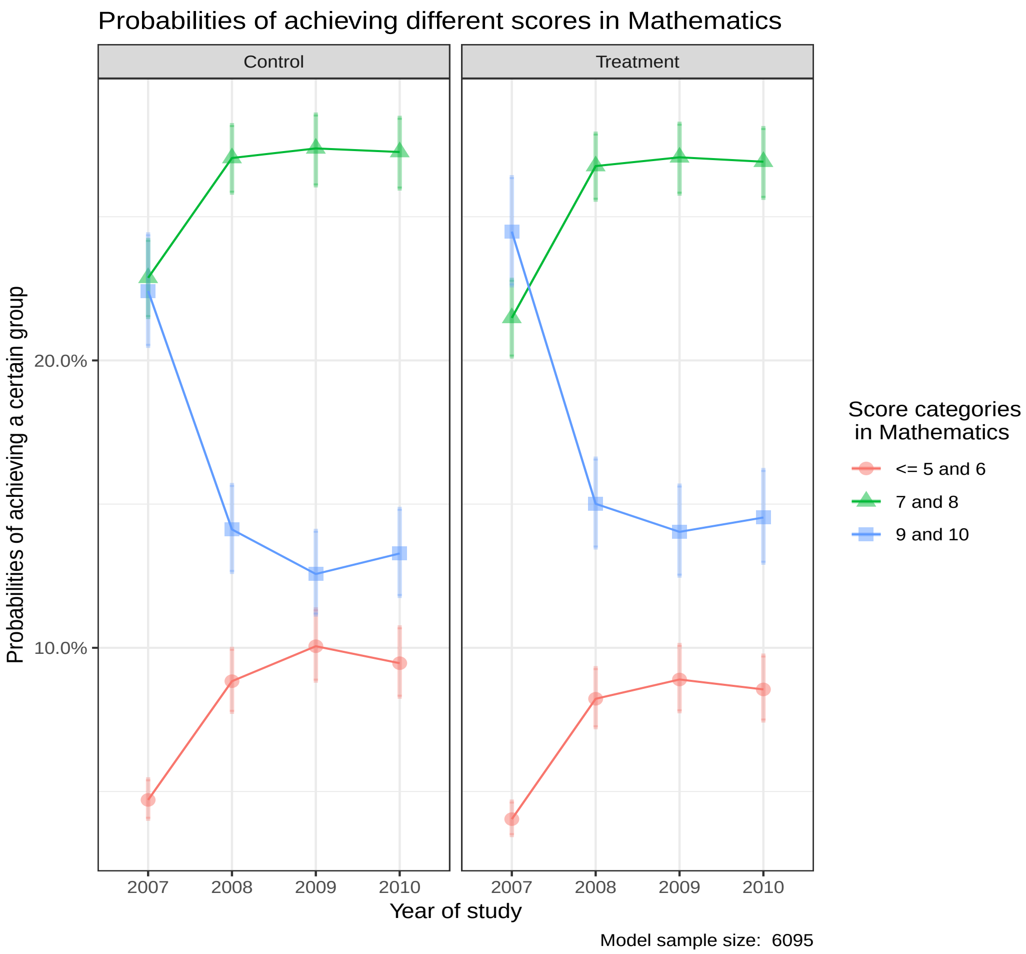


Figure 2.3: Probabilities of achieving certain grades in Mathematics across all years for treated and control units for 5th grade

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	0.84	0.63	1.12
Year 08	0.71	0.57	0.89
Year 09	0.64	0.5	0.82
Year 10	0.69	0.55	0.87
Middle SES	1.2	1.05	1.38
High SES	1.26	1.08	1.48
General School	1.15	0.93	1.41
Age	0.96	0.91	1.01
% Mid-High SES in School	0.75	0.48	1.14
Treatment * Year 08	1.07	0.79	1.44
Treatment * Year 09	1.69	1.19	2.38
Treatment * Year 10	1.38	1.01	1.88
N. Observations	3513		
WAIC (similar to AIC)	11444.95		

Table 2.7: Bayesian ordinal model of Natural Science test score on covariates for 4th graders. Estimates expressed as odds ratios.

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	1.21	0.98	1.5
Year 08	0.69	0.6	0.81
Year 09	0.54	0.46	0.63
Year 10	0.56	0.48	0.66
Middle SES	1.18	1.06	1.31
High SES	1.21	1.07	1.37
General School	1.2	1	1.43
Age	0.92	0.89	0.96
% Mid-High SES in School	0.75	0.53	1.05
Treatment * Year 08	0.9	0.72	1.12
Treatment * Year 09	1	0.8	1.26
Treatment * Year 10	0.86	0.69	1.07
N. Observations	5952		
WAIC (similar to AIC)	19355.84		

Table 2.8: Bayesian ordinal model of Natural Science test score on covariates for 5th graders. Estimates expressed as odds ratios.

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	0.91	0.74	1.12
Year 08	0.74	0.63	0.87
Year 09	0.55	0.46	0.65
Year 10	0.6	0.51	0.69
Middle SES	1.23	1.08	1.41
High SES	1.48	1.27	1.72
General School	1.08	0.86	1.36
Age	0.93	0.88	0.98
% Mid-High SES in School	1.06	0.68	1.67
Treatment * Year 08	0.97	0.83	1.14
Treatment * Year 09	1.32	1.17	1.48
Treatment * Year 10	1.21	1.15	1.27
N. Observations	3566		
WAIC (similar to AIC)	11467.37		

Table 2.9: Bayesian ordinal model of Spanish test scores on covariates for 4th graders. Estimates expressed as odds ratios.

Term	Odds ratios	95% lower bound	95% upper bound
Treatment	1.29	1.04	1.59
Year 08	0.72	0.62	0.85
Year 09	0.56	0.48	0.67
Year 10	0.62	0.53	0.72
Middle SES	1.14	1.03	1.26
High SES	1.31	1.16	1.48
General School	1.21	1.01	1.47
Age	0.89	0.86	0.93
% Mid-High SES in School	0.81	0.57	1.16
Treatment * Year 08	0.82	0.67	1.02
Treatment * Year 09	0.87	0.69	1.09
Treatment * Year 10	0.86	0.69	1.07
N. Observations	6023		
WAIC (similar to AIC)	19241.49		

Table 2.10: Bayesian ordinal model of Spanish test scores on covariates for 5th graders. Estimates expressed as odds ratios.

2.8.1 Results with imputed data set

These results come from the same models described in the main text but all main independent variables have been imputed using multiple chained equations. The results hold and become stronger with this imputed data set.

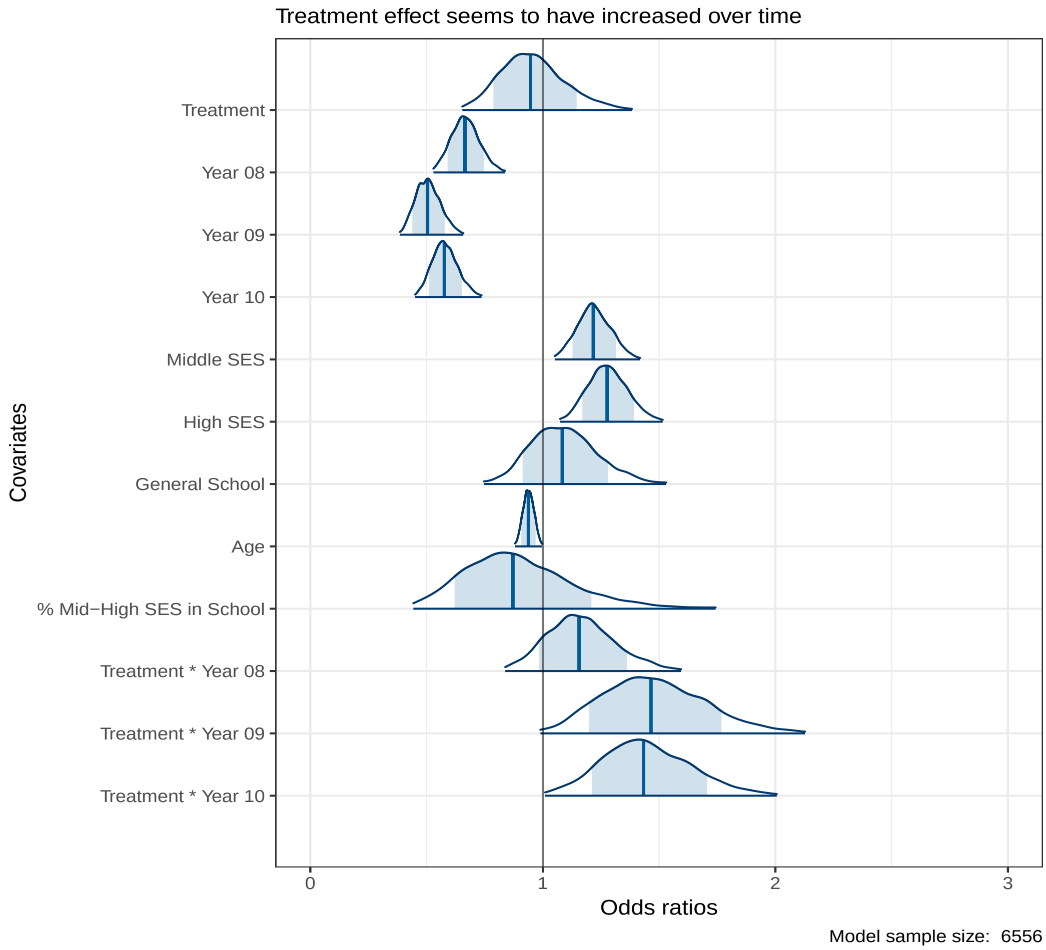


Figure 2.4: Distribution of the coefficients of the model expressed as odd ratios using a multiple imputed dataset

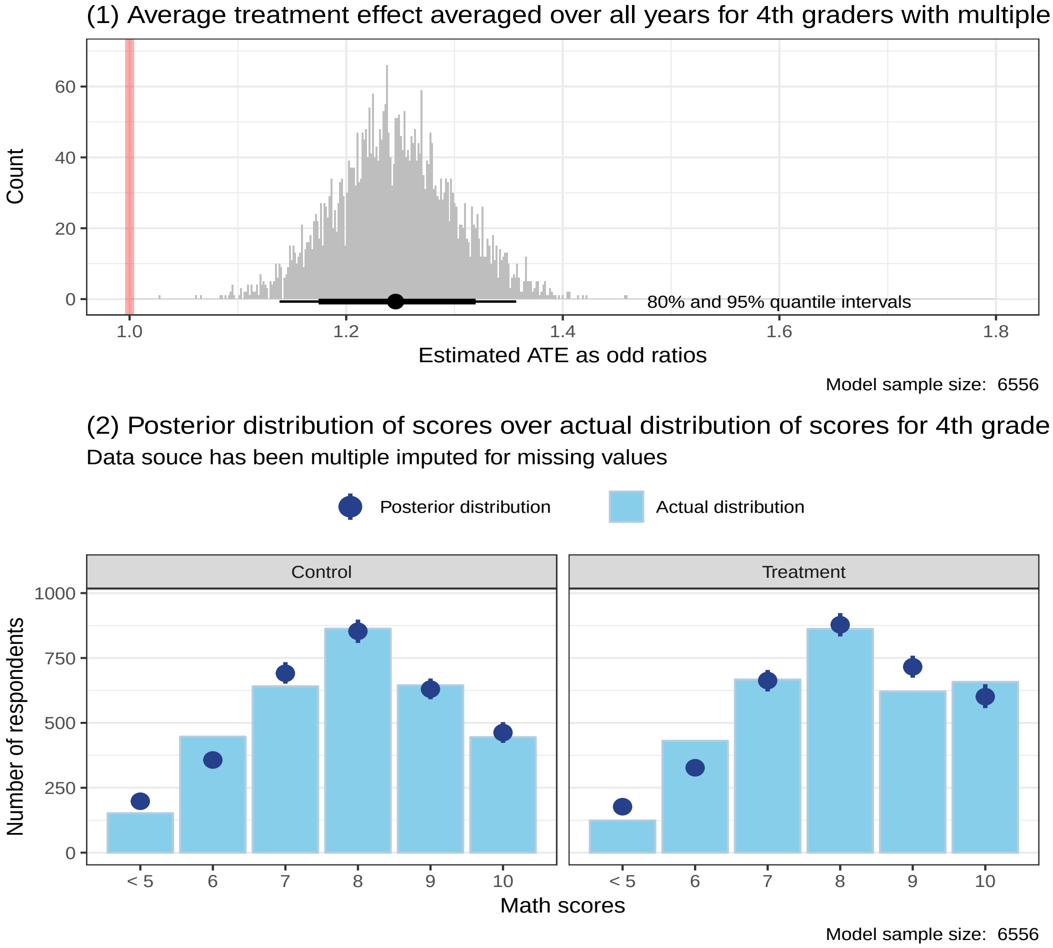


Figure 2.5: Treatment effect averaged over all years and model fit between posterior and actual distribution of test scores. Result for 4th graders in Mathematics using multiple imputed dataset.

2.8.2 Relationship between SES index and test scores

As a final step I also test whether the SES categories are highly correlated with cognitive abilities, a standard correlation in sociological research. Figure 2.6 presents the average score in three tests for all years and all SES categories.

With the exception of 2009, there is a clear positive correlation between test scores and SES categories for all years [15]. This is reassuring as the linear trend lines up much more credible than the same plot for father's education. We should be careful in interpreting the correlation after 2007 because the different shifts could very well be due to the effect of the treatment. It is also important to mention that given that many students have scores

[15] Note that I exclude the father's holding a university degree from the high SES group because they are just very few in the complete sample and they were influencing the SES categories to extreme numbers

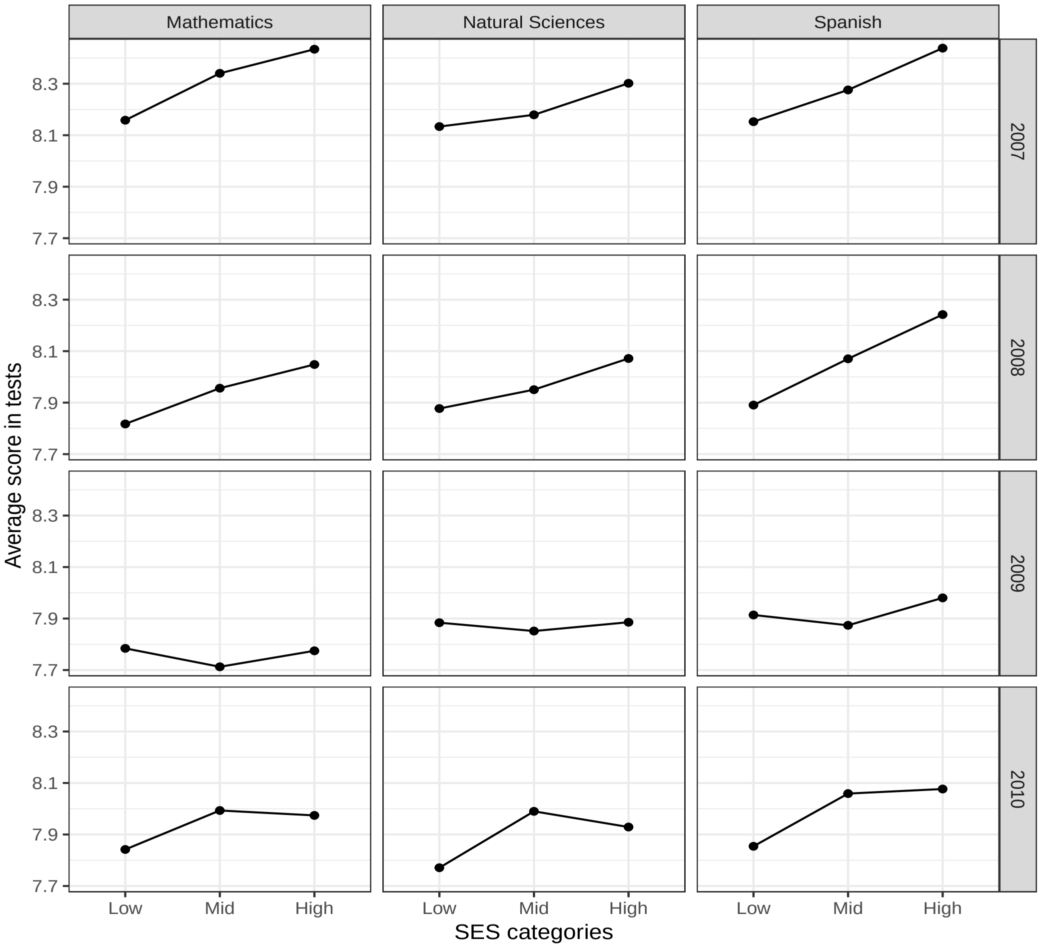


Figure 2.6: Predicted SES categories and average performance in different tests across all years

between 7 and 8, the differences between SES categories are not very big (in the order of 0.2/0.3). This plot is merely informative in that the SES categories seem to be measuring what we expect it to measure.

CHAPTER 3

Does curricular tracking explain global SES gaps? an international comparison of the SES achievement gaps from 2000 to 2015

Abstract: The literature on achievement inequality has recently started to focus on the dynamics of the socio-economic achievement gap in cognitive abilities. The main findings come from research in the U.S. revealing that the 90th/10th income achievement gap has widened 50% in the last 30 years. This chapter aims to investigate whether there are patterns in the evolution of the achievement gap from a comparative perspective. Using 15 years of data in 32 countries from the Program for International Student Assessment (PISA), I find that there is considerable variation in the way in which the gap between the average score of students above (and at) the 90th percentile and below (and) the 10th percentile is evolving. The prime examples come from the U.S. and Germany closing at about 50% and 30% in the last 15 years while France is widening at a similar rate. I find that curricular tracking and vocational enrollment explain 40% of the variance in the achievement gap between countries and show that the relationship is conditioned by a strong interaction. Low curricular tracking is associated with a small achievement gap, whereas high levels of curricular tracking is associated with wide achievement gaps. However, once tracking is coupled with high vocational enrollment this can remedy the potential adverse effects and reduce the gap by .6 standard deviation. I use simulations to show that switching to less curricular tracking can help decrease a country's SES gap

by about 10% while switching to more tracking would increase the achievement gap by about 51% percent.

3.1 Introduction

Inequality, be it economic or of non-pecuniary causes, is a crucial determinant of the opportunities a child will be exposed to. For the last decade or so, inequality has been rising and the literature on inequality has found evidence of increasing gaps between the rich and the poor in several dimensions.

Researchers have found evidence that, for example, for the U.S., economic inequality (Piketty and Saez, 2003), inequality of opportunity (Chetty et al., 2014) and inequality of achievement (Reardon, 2011) have increased in the last 20 years. Unfortunately, most of the studies to date mainly concentrate on the U.S.. There is a growing body of research that attempts to document and partially explain achievement gaps. However, there is still need for further explanations for the starking differences in achievement gaps between and within countries.

This chapter looks to study the cognitive achievement gap between the students above and at the 90th percentile and students at and below 10th percentile of the socio-economic distribution (SES from now on). The gap can be referred to as the achievement gap between the top and bottom SES groups. Using the Program for International Student Assessment (PISA) survey, I calculate the achievement gap for all available countries that participated in all the six available PISA waves, building a country pseudo-panel that shows a time trend of 15 years.

More concretely, I aim to study the evolution of this achievement gap and a possible explanation as to why there are such big differences between countries. The two questions of interest are: how have achievement gaps changed over time and what explains such changes. Building on findings from studies in educational institutions, I suggest the degree of curricular tracking of a country and its level of vocational enrollments as possible explanations for the high variation between countries.

The findings from this chapter suggest that there is reasonable variation in the evolution of the gap. For example, the U.S. has been closing its 90/10 gap by about 50% with a lower bound of nearly 35%. The reduction is also present in other countries such as Germany and Poland. Other countries such as France and Austria experience the contrary with persistent increases in the achievement gap. The results show that these gaps differ very much in their composition, where in some countries the top 90th percent of the SES gap are benefiting more whereas in others the bottom group is catching up at very rapid pace. Robustness checks show that these results replicate for the 80/20 and 70/30 gaps providing more strength to the validity of the findings.

I find that the degree of curricular tracking of a country seems to be a possible explanation for the changes in achievement gap; however it must be studied by also factoring in the degree of vocational enrollment. The results suggest that if a country has low levels of curricular tracking the gaps are considerably smaller than when tracking is highly predominant in the curriculum. When tracking is present, the gap is bigger by around .5 standard deviations. However, once curricular tracking enters the picture, vocational enrollment can help ease the burden of inequality and reduce the achievement gap significantly. Simulations show that if every country switched to little or no curricular tracking their gaps would shrink on average by about 10%.

The findings have two specific implications. First, researchers must avoid 'generalizing' or 'summarizing' inequality with one single indicator across many countries (Chmielewski, 2016). There are widespread contextual differences between countries, and each one should be studied with such details in mind. Second, further research should attempt to explain why some countries have experienced particular trajectories of inequality. I explore a possible explanation: the degree of curricular tracking and vocational enrollment of the country. Tracking seems to exacerbate inequality, but if coupled with a reasonable level of vocational enrollment it can help to alleviate the negative effects of curricular tracking.

However interesting these results are, there are still concerns on whether these trends are reflecting the true cross-country trends or whether they are also reflecting sampling variation. To understand these short comings, future research should model the trends statistically and test for whether they are substantially different in their magnitudes. I have avoided this type of analysis until this point because there are only 6 time points per country which is not enough to apply any proper time-series analysis. The paper attempts to begin to understand the phenomena by providing descriptives of the trends but it should be acknowledged that future work should prioritize modeling the trends statistically.

The chapter first discusses the literature on the SES achievement gap as well as on the known evidence of the effects of curricular tracking. It then introduces the research question together with the methodology. In the first part of the results section, the chapter explores how gaps have changed over time while the second part implements the modeling section. The limitation section acknowledges the main limitations of the study and the chapter concludes with an overall discussion of the results.

3.2 Literature Review

Educational inequality and its long-term impacts are topics that have been prominent in the social science literature for the past 30 years. The idea of meritocracy and intergenerational transfers has motivated, for a good part of the 20th and 21st century, much of the research on social mobility and social inequalities. When James Coleman released his famous Coleman report (Coleman et al., 1966), he helped to show that a family's socio-economic status and a student's performance are tightly linked. Since then, the topic has been studied extensively and several authors have contested whether the relationship is an invariable social law or a product of institutional arrangements [1]. Today, we have a much stronger understanding of the relationship.

Psychologists have been investigating child development for at least half a century and they find that the early stages in a child's life course are extremely important, if not the most important, for cognitive development and defining personality traits (Duyme et al., 1999; Waldfogel, 2006). The work of James Heckman helped to bring wider empirical attention to the subject. Heckman showed that cognitive and non-cognitive inequalities are present even before a student enters school (J. Heckman, 2006). To explain these, and other findings, Cunha et al. (2006) hypothesize that the cognitive level of a child at time t is a direct function of the experiences at time $t-1$. While it sounds straightforward, its importance is often missed. The model implies that investments into the education of a child compliment each other. It is difficult to easily compensate for an earlier lack of investment by investments at later stages. There are specific periods of skill formation in a child's life in which investment is particularly cost-effective. As a general rule, the earlier the investment in a child's education the greater the return. When tested against the

[1] For a detailed review of the long literature on educational inequality, please see Gamoran (2001).

data, Cunha et al. (2006) show that their explanation is consistent with other empirical studies.

Investments in individuals also support society as a whole as it boosts economic activity, it helps the labor market improve job conditions and maintain a rapid economic growth (Hanushek and Woessmann, 2007). Despite these findings and the recommended strategies to reduce the gap in cognitive achievement, we still find that a significant relationship between parental education and future destination is present in virtually all empirical studies of social mobility and inequality (Breen and Goldthorpe, 1997; Breen and Jonsson, 2007; Waldfogel, 2006; Bradbury et al., 2015; Chetty et al., 2016).

Since Coleman et al. (1966), educational researchers have spent most of their time studying the mechanisms through which this inequality has arisen. Naturally, they want to do that in order to reverse it and help every child reach his or her fullest potential. Despite the efforts, the literature still needs more studies that seek new explanatory mechanisms for these gaps, especially in terms of cognitive abilities. We need more evidence that seeks to explain the large and small cognitive gaps resulting from SES origins. And even more importantly, we have not assessed whether policy efforts to reduce inequality have actually had an impact in reducing the achievement gap over time.

Despite the shortcomings, we have indirect evidence of the relationships from the vast literature on social mobility. We know that virtually in all countries, developed or developing, there is inequality of opportunity. But there is considerable variation in the magnitude of inequality. For example, in the Scandinavian countries, particularly Denmark and Sweden, inequality is low compared to other countries and individuals have greater control over their destiny in terms of class mobility (Esping-Andersen et al., 2012; Breen and Jonsson, 2007; Shavit and Blossfeld, 1993).

From Denmark, researchers have learned a great deal about how to improve social mobility but that was possible by first learning that Denmark is by far one of the most mobile countries compared to other European countries; this is the case specially for families of low-SES origin (Bjorklund and Jantti, 2009; Jaeger and Holm, 2007). The important finding came when research discovered the main reason behind their social escalator: its educational system. For example, research by Esping-Andersen et al. (2012) and Bauchmuller et al. (2014), shows that the Danish education system has important and longstanding impact on improving opportunities. The education system is completely subsidized for all children, otherwise giving opportunities to families who would not have been able to pay. In addition to this, Denmark is recognized as a world-leader in terms of public support for its child-care system as it spends around 2% of GDP, and has among the highest enrollment rates for children under 6 years old (Esping-Andersen et al., 2012). Moreover, it separates students into different curricular tracks later than other European countries, at 16 years of age, something which has been linked to less educational inequality (Hanushek et al., 2006). These two traits make the Danish system effective in promoting equality relative to the other European countries. The early schooling experience attempts to put children on the same level and the process is not stopped by different curricular tracks as it starts at around age 16 when cognitive abilities are less malleable (Kautz et al., 2014). In short, the importance of this finding is that we should first study the presence and size of the effect and then proceed to find the causes behind it.

The first attempts to study the evolution of the achievement gap has found that the gap in cognitive abilities between high-SES and low-SES children has been widening over the years (Reardon, 2011). The literature on the topic has mainly concentrated on studying the case of the U.S. but other international evidence is emerging with a similar landscape. The U.S. is usually the case of study as it is the only country where cognitive testing is present from as early as 1940. Using this information, Reardon (2011) was the first to investigate the evolution of the cognitive gap and the results were surprising. Not only has the cognitive gap between the 90th income percentile and the 10th income percentile grown over time, but it has grown faster and to be wider than the highly contested white-black gap (Magnuson and Waldfogel, 2008). The gaps have actually reversed and we find that the income achievement gap is nearly twice as large as the black-white achievement gap (quite the opposite to 20 years back).

Reardon (2011) finds that the increase in the gaps has occurred predominantly for children born in the 1970's compared children born in the year 2000. In fact, the hard numbers suggest that the gap widened by 40-50%. The author also estimates the rate of change using data as early as 1940 and finds an even higher increase of 75%. Given that the studies before 1970 are less reliable in terms of comparability and sampling design, the author computes all results for before/after 1970. To provide a definitive answer to the size of the gap, Reardon (2011) concludes that the 90/10 income gap in the U.S. has a standard deviation (SD from now on) of 1.25 in tests scores for the year 2001. Using longitudinal data, Bradbury et al. (2015) find similar results. Their empirical analysis suggests that for 14 year old Americans, there is a SD of above 1, but lower than 1.25. Interestingly, Duncan and Magnuson (2011) find very similar results to the previous studies and confirm a gap with a SD of 1.25-1.50. To put these results into context, evidence from PIRLS shows that the predicted growth of a student for a year of school is of around 0.30 standard deviations (Beaton et al., 1996). PISA has also documented this type of metric in their annual reports (OECD., 2009). Having said, the magnitude of these gaps is very relevant.

Interestingly, the widening of the achievement gap has been paralleled by a growth of income inequality, which may be telling. Reardon (2011) offers several possible links, with the most reasonable being that family investment patterns have changed so that high income families now invest more resources on their children. The explanation lies in the fact that increasing income became more strongly correlated with other positive family traits related to time allocation and welfare services.

In a follow-up study, Reardon and Portilla (2015) uncovered a reversal of the trend. The follow-up study concentrated solely on kindergarten children in the U.S. for the years 1998, 2006 and 2010. They found that the 90th/10th income gap in readiness closed modestly. Furthermore, using data from fall and spring in the same kindergarten year, they calculated that the gap narrowed at a rate of 0.01 and 0.008 SD per year for mathematics and literacy between 1998 and 2010. They also calculated the same changes for a number of personality traits such as self-control and externalizing behavior and found similar results. In contrast, Reardon (2011) finds that in a 30-year span the gap was systematically increasing at a rate of 0.02, something reasonably close to the previous estimates. Their results not only hold for the income achievement gap, but they also found a decline in the white-hispanic gap (although not for the white-black gap). It should be noted that perhaps the reversal of the trend in Reardon (2011) would had been evident if data were available for years after 2000, the time-point from which Reardon and Portilla (2015) find the reversal.

The reasons why the authors find a reversal in the trend could be numerous and should be studied closely. They discuss a number of country-level indicators to explain this change and suggest that the reversal is likely due to the high increase of preschool enrollment.

They build on their previous argument by suggesting that in this same period (1998 - 2010) the income achievement gap in early schooling enrollment decreased substantially. Their conclusions, although suggestive, are speculative and have no *empirical support* which is why this is still an open question.

There have been other attempts to explain the achievement gaps with indicators such as economic inequality (Dupriez and Dumay, 2006), the difference in schooling hours and the tracking system (Duru-Bellat and Suchaut, 2005; Dupriez and Dumay, 2006), home and family factors (Marks et al., 2006) and fertility rates and expanding school access (Chmielewski, 2016). Each of these studies has made a contribution to understanding what works and what does not. The work of Dupriez and Dumay (2006), for example, explored the relationship between achievement gaps and economic inequality but without factoring in the multilevel structure of the students nested into schools. Moreover, it merely correlated achievement gaps with economic inequality. The work of (Duru-Bellat and Suchaut, 2005) is more comprehensive as it explores several indicators of the school system, among which is the differentiation structure of the secondary school system (tracking). However, as noted by Reardon et al. (2008), *'our understanding of the causes and patterns of these achievement gaps is far from complete'*. For this reason, the review by Van de Werfhorst and Mijs (2010b) gains particular relevance because it documents many instances in which tracking explains inequality between schools (one notable example is the work of Dupriez and Dumay (2006) which finds a strong correlation between tracking and achievement gaps).

Motivated by these recent results, other authors have taken this analysis to an international context in order to discover between-country trends. The work of Bradbury et al. (2015) employs a unique comparative analysis of the achievement gap between Australia, United Kingdom, United States and Canada. Their research design is distinctive in that they use longitudinal data from children as early as age 2 and study the evolution of the achievement gap up until age 14 [2]. The core finding of their study is that the American achievement gap is much wider than the gaps in Australia and Canada. They find that once the achievement gap is present in early school entry, it does not seem to narrow or widen much over the life course. In fact, they estimate that the quality of early childhood

[2] To the best of my knowledge this is not only the first study that uses panel data to study achievement gaps, but to also do it between countries

education can only explain about 30-40% of the high school SES gap. This suggests that once the achievement gap is present before entering school, it carries a social-scar effect [3]. One exception is the UK, which they found to be a country that helps close the gap in early primary years. This can likely be due to the comprehensive schooling and also the public support by the welfare state in dimensions like health and income support.

Even though their results are interesting and robust, one limitation of the work of Bradbury et al. (2015) is that their analysis is based on four surveys that have significant differences in terms of questions, sampling and populations and cannot be easily compared. They manage to harmonize the four surveys into a comparable format and their findings do seem to be reliable. But we should be careful at interpreting these findings causally without considering that the four surveys carry great deal of differences in terms of measurement and survey questions. For this reason, we should also pay particular attention to studies such as Chmielewski and Reardon (2016) and Chmielewski (2016) who have attempted to compare gaps between countries, and to evaluate whether there is a general increase in the gap using comparable surveys. However, it should be noted that the study of Chmielewski and Reardon (2016) tackles a completely different question from the above, namely to study cross-sectional differences between many countries, instead of over-time analysis of student gaps. Nonetheless, the work of Chmielewski (2016) does provide support for the overall finding that the achievement gap is widening over time.

A thorough review by Van de Werfhorst and Mijs (2010b) also sheds some light on the subject. First and foremost, they gather substantive evidence showing that countries which have a highly tracked curriculum tend to have high levels of inequality, measured in terms of achievement gaps. Hanushek et al. (2006) use the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) surveys to gauge whether highly tracked countries do indeed increase inequality after students pass the age at first selection of curricular tracking. The results suggest that early selection increases educational inequality. While less clear, there is also a tendency for early tracking to reduce mean performance. Micklewright and Schnepf (2007) using PISA but a different empirical strategy find that countries which have a high level of curricular tracking, are

[3]However, schooling could be preventing the gap from widening even more, and rigorous Randomized Controlled Trials (RCT) show that high quality schooling can indeed help ease the gap, in some instances even close it (Campbell et al., 2002).

distinctively unequal in the difference between the top 95th and bottom 5th performers. Their strategy simply calculates differences between the 90th and 10th percentile something different from the current paper, as I rely on the method developed by Reardon (2011) for removing biases in the achievement gap. In their results, the difference in test scores between these two groups is about 10 times higher than the average annual gain of a year of schooling.

Chmielewski and Reardon (2016), again using PIRLS, TIMSS and PISA, assess whether there are patterns of cross-national variation in the achievement gap. In other words, does the achievement gap differ between countries? Their work suggests that there is considerable variation in the achievement gap between top and bottom earning families across many developed countries. In comparison to the literature on achievement gaps, they find that the U.S. has a gap of 1.20 SD in 2001 which increase to around 1.30 in the year 2006 while Germany has a decreasing gap from around 1.25 to 1 SD in the same year-span. However, these numbers vary a lot and carry a great deal of overlapping uncertainty. They go even further and link this achievement gap to several country-level indicators related to income inequality, school differentiation and central exams, among other things. These correlations are suggestive as explanatory mechanisms but they are cautious in drawing causality.

One interesting question that is still missing from the literature is how these country gaps have evolved over time. With their data, Chmielewski and Reardon (2016) only have 3 countries which are present in all waves and also have very few waves as their question of interest (income categories) was only asked at three time points. The results are more about between country gaps rather than the magnitude and evolution of the gaps.

Chmielewski (2016), building on the work of Chmielewski and Reardon (2016) and Reardon and Portilla (2015) pooled together all the previously mentioned data, together with over 10 more studies ranging from the years 1964 to 2015 in order to discover differences between countries. With over 50 years of data, and over 100 countries, the author finds that there seems to be a general pattern of increasing achievement gap. However, once she disentangles the relationship by country, she finds a sizable amount of heterogeneity, with some countries experiencing a narrowing of the achievement gap, others no change at all, while others record a steady increase. This is revealing as it does not really pay off to look at a general average once each country has their own distinctive gap and evolution. This highlights the notion that the increasing achievement gap is clearly not universal and should be studied in context.

One limitation of the study of Chmielewski (2016) (as well as Reardon (2011)) is that they adjust for the age of each child in all studies. Although for their modeling purposes this is the correct thing to do [4], these modeling strategies are masking age-specific achievement gaps by controlling for age. We clearly see in Reardon and Portilla (2015) that there are age-specific gaps, and they do change at a fast pace in little time.

In fact, the evolution of high/low SES gaps for preschool children might be much less marked than the same gap for high school children. The explanation, although debated, has been gaining much support in recent years. For countries with high levels of curricular differentiation the transition from early schooling into the tracking system has been found to increase inequality of learning (Hanushek et al., 2006). Moreover, the vast sociological literature on educational transitions systematically finds that early selection tends to foster between-track inequality rather than erode their differences by tackling their specific needs (Van de Werfhorst and Mijs, 2010b). Based on this, it is difficult to simply assume that the achievement gap has been neither constant across cohorts nor the same between ages, as tracking/no tracking might exacerbate the achievement gap. Also, there is the possibility, theoretically, of self selection intro tracking, which forces some parents to invest more in their children *before* tracking is implemented to increase chances of selecting their children into the higher tracks (Jakubowski, 2010; Waldinger, 2006).

3.3 Research questions and hypothesis

The overarching aim of this chapter is to study the evolution of the high/low SES achievement gap in the past 15 years for several PISA participating countries and propose a likely explanation for the evolution. To maximize variation and comparability, I use all countries which have participated in PISA in at least 50% of all available waves.

I develop the sub-questions and their corresponding hypothesis separately in more detail.

[4]The differences in achievement could simply be due to changes in cognitive abilities across the lifetime. However, as we have noted before, Bradbury et al. (2015) find that the achievement gap is stable across the life time

1. The seminal work of Reardon (2011) suggests that achievement gaps change, and they do so much quicker than we thought after recording a SES gap increase of about 40% in 30 years. Reardon and Portilla (2015) stress that they also found a significant decrease in only 15 years of data, showing how important it is to study the changes in the achievement gap. First, I will concentrate on the evolution of the gap only for 15-year olds. As we have seen before, there are reasons to consider specific age-groups when estimating achievement gaps as there might be differences between age groups. This will serve as a comparison to the single year-country snapshot of Chmielewski and Reardon (2016) and the evolution of the kindergarten gap in Reardon and Portilla (2015). Secondly, I will compare the percentage change at which the gap widened/narrowed from the first to the last year available. This will give a general idea of the overall change over time. [5]. I posit no specific hypothesis for this question given that it is purely explorative.

   - Research question 1:

     (a) How is the achievement gap changing between countries?

2. The literature has concentrated narrowly on whether the gap is increasing because the top performers are getting ahead, because the bottom performers are falling behind or because both are changing at the same time. The work of Bradbury et al. (2015) is the only study that pays attention to the source of the achievement gap using longitudinal data that I am aware of. The findings are heterogeneous for the four countries in the study but the overall evidence shows that as children grow older, top and bottom SES groups seem to grow apart at a similar rate. These findings are very important for understanding inequality through out the life course of individuals but this still does not answer whether specific age groups have gaps that change over time. Moreover, their analysis is limited to four countries that have very little to no institutionalized curricular tracking. This type of question is important because it highlights whether there are specific policies that might prevent gaps from closing over time.

   - Research question 2:

[5] Although no study has performed this age-specific achievement gap for comparable tests over such a long time. The results will serve as comparison for other studies that use age-specific groups, such as 4th graders using PIRLS.

  - Is the gap originating as the top is gaining ground, the bottom is falling behind or because of a dynamic interaction between the two?

- Hypothesis 1:

  - The theoretical argument in favor of curricular tracking posits that in countries where there is a high degree of tracking we should expect the top and bottom to be evolving at a similar rate given that curricular tracking is thought to maximize the learning experience of both groups. The review by Van de Werfhorst and Mijs (2010b) discussed this critically and comment on the results from Hanushek et al. (2006) which suggests that tracking seems to be associated with greater inequality between SES groups. That being said, I expect to find that countries with highly tracked curriculums have growing achievement gaps as children from low-SES groups are at a greater disadvantage. Conversely, countries with low curricular tracking are expected to be associated with narrowing gaps as both SES groups are thought to be less segregated and thus equalizing their learning experiences. The specific mechanism from which this might be happening could be due to the varying investments between tracks, where the quality of teaching is different between the different tracks.

3. The work of Chmielewski (2016) shows that there are differences between countries in their overall levels of achievement gaps. This is different from the previous hypothesis because this chapter concentrates only on age-specific gaps rather than age-adjusted gaps. Given that most countries available in PISA participated in all six waves, this question attempts to find a possible explanation for why there are such stark differences in the SES achievement gap between countries. Chmielewski and Reardon (2016) perform a similar analysis but they concentrate only on *income* achievement gaps rather than on a more broad SES index. Moreover, they only perform their analysis on a handful of countries which limits the amount of between-country variability. The question outlined here tests whether several dimensions of tracking and vocational enrollment explain the differences in achievement gaps between countries. This has been explored before in the work of Jakubowski (2010) and discussed in detail in Van de Werfhorst and Mijs (2010b) and Hanushek et al. (2006). However, this paper uses student time data and a rich set of tracking and

vocational indicators not used before, which offers a new possibility to explain the phenomena.

- Research question 3:
  - Does curricular tracking and vocational enrollment explain differences between countries in achievement gap?
- Hypothesis 2:
  - According to the review by Van de Werfhorst and Mijs (2010b), the tracking setup of a country should play an important role in explaining the marked differences in achievement gaps between countries seen in Chmielewski (2016). I expect to find that a reasonable percentage of the variation between countries is explained by the fact that some countries have highly institutionalized tracking setups, while other countries have more flexible tracks.

4. The reasons why the gap is changing are still speculative. Some researchers have pointed out to the share of preschool enrollment as a possible mechanism (Reardon and Portilla, 2015) while others have tested the degree of economic inequality within a country (Chmielewski and Reardon, 2016). This chapter explores a different mechanism thanks to the work of Van de Werfhorst and Mijs (2010b). This question looks to uncover whether the curricular tracking setup of a country is also a possible explanation for the *evolution* of the achievement gap. This has theoretical and empirical justifications given that tracking might exacerbate achievement gaps once it is implemented (Van de Werfhorst and Mijs, 2010b). This could be possible by following this chain of logic: the introduction of early selection and tracking leads over time to a shift of resources to the academic track and thus to an exacerbation of inequality. This is because tracking can have dynamic effects given that the quality of instruction in different tracks can decrease/increase over time even if (available) tracking indicators stay constant. Together with this argument, there is observational evidence which suggests that a reduction of a tracking reform in Poland had a particularly positive and varying effect in reducing inequality in a time span of over 6 years (Jakubowski et al., 2010). In particular, this paper takes advantage of vocational tracking indicators which do change over time and tests whether these

interact with more classical tracking features. This advantage brings forward the plausibility of tracking explaining changes over time, something not done thus far.

In this question I cannot test for pre/post tracking reforms to investigate changes in gaps. However, it is possible to study whether tracking, together with the degree of vocational enrollment are valid explanations for the changing achievement gap within a country. To be clear, the data and methods in the study cannot support the offset hypothesis of tracking leading to dynamic changes over time. What it can do is to test whether tracking and vocational features can to a certain extent explain the dynamic changes of the achievement gap. Further research should attempt to find cases where pre/post observations are available to test the offset hypothesis more thoroughly.

- Research question 4:
  - Are curricular tracking and vocational enrollment related to the evolution of cognitive achievement gap?
- Hypothesis 3:
  - Building on the previous hypothesis, curricular tracking should play an important role on the evolution of the achievement gap. However, the relationship is not straight forward because tracking hardly changes over a period of 15 years. Curricular tracking might exacerbate gaps through mechanisms which are not seen through traditional tracking indicators. For instance, if the quality of instruction in the lower tracks worsen over time. I hypothesize that the degree of tracking of an educational system is related to changes in the gap, and the more curricular tracking, the more inequality. Moreover, the more vocational curricular tracking, the less inequality.

3.4 Methods

3.4.1 Data

To investigate the above mentioned questions, I will use the Programme for International Student Assessment (PISA). PISA is a survey carried out every three years that aims to evaluate education systems by testing the skills and knowledge of 15-year-old students. Currently, PISA has six waves starting in 2000 up until 2015, where recently, over half a million students were tested in mathematics, literacy and science in over 70 developed/developing countries.

PISA collects data through a two-stage stratified sampling design. With the help of governments, PISA randomly chooses 150 schools in each country, where they then randomly pick thirty 15 year olds to undertake the two hour tests. The sample size for each of the waves are 127,388 for PISA 2000, 276,165 for PISA 2003, 398,750 for PISA 2006, 515,958 for PISA 2009, 480,174 for PISA 2012, and 519,334 for PISA 2015. Together with the subject tests, PISA collects personal information from students, their families and their school environment (including teacher surveys), that serves as relevant background information that can be matched to the students performance. With the recent inclusion of PISA 2015, these six waves make up a time-series analysis of 15 years, enough to visualize changes in the structure of an educational system. None of the studies cited so far has used the last PISA wave, which was released in December 2016. This chapter takes advantage of these six waves to build a country pseudo-panel, making it possible to study changes in nearly 15 years for 32 countries. It should be noted that 8 out of the 32 countries did not participate in all waves, making a total of 183 country-year combinations. In order to maximize country variation, I have included countries which have at least participated in 50% of all waves. Table 3.7 in the appendix presents the countries and the number of years available for each country. Austria, Bulgaria, Chile, Israel, Japan, Slovakia and Turkey participated in 5 of the 6 waves and Slovenia only in 4.

To identify a family's SES, PISA collects several variables that measure different dimensions. Classically, they ask student's their parent's educational level. Scholars have considered this to be a reliable recall given that we expect fifteen year olds to know their parent's level of education (Reardon, 2011). This question has been asked in every wave and holds a somewhat similar coding across time, although the first two waves have small differences. In spite of this, a serious limitation is the fact that parent's education is measured using the ISCED classification, something that has changed over time. For example, until PISA 2009, the preferred framework was ISCED 1997, whereas the next wave switched to the newly developed ISCED 2011 classification. Both these classification schemes have equivalent look-up tables, but this requires a detailed inspection of the codings.

Another social background variable is the International Socio-Economic Index of Occupational Status (ISEI). This variable attempts to capture the social status of the family, without asking for income information. This index variable was developed by Ganzeboom and Treiman (1996) and later refined by Ganzeboom (2010) and it attempts to measure occupational status using a continuous measure. The indicator is a reliable alternative to the classical Erikson-Goldthorpe-Portocarero classification (Erikson et al., 1979). It has been scaled for comparability between waves and some authors have used it for inequality studies, finding expected results to be consistent with the social mobility literature (Chmielewski, 2016). PISA also includes a plethora of indicators on family wealth, home educational resources, the number of books in the home, among many other material resources in the household.

Yet one of the most relevant variables for this study is a composite SES index created by the PISA team. The index of economic, social and cultural status (ESCS) was created on the basis of the following variables: the International Socio-Economic Index of Occupational Status (ISEI), the highest level of education of the student's parents, the PISA index of family wealth (which measures the material wealth of the family), the PISA index of home educational resources; and the PISA index of possessions related to "classical" culture in the family home (mainly about books in the household) (OECD, 2002). The variable, aside from capturing all relevant dimensions of SES, such as education, occupation, and material resources, takes care of transforming all mentioned variables into comparable metrics across waves.

The ESCS index was derived from a principal component analysis of standardized variables, taking the factor scores for the first principal component as measures of the PISA index of economic, social and cultural status. All countries and economies (both OECD and partner countries/economies) were assigned the same weight in the principal component analysis, while in previous cycles, the principal component analysis was based only on OECD countries. However, for the purpose of reporting, the ESCS scale has been transformed with zero being the score of an average OECD student and one being the standard deviation across equally weighted OECD countries (OECD, 2016).

To the best of my knowledge this is the first piece of research that uses the newly-released ESCS index (OECD, 2016), which was rescaled so that all ESCS indexes are suitable for over-time analysis [6]. In other words, the ESCS index does not need any transformation or coding updates as it is ready for comparison over time.

Aside from SES, the other relevant variables are test scores for mathematics and literacy [7]. PISA does not provide a single test result for each respondent. Instead, it provides a *series* of 'plausible values' that the child could actually score. As explained in the PISA manual (OECD, 2012), these are imputed values that resemble individual test scores and have approximately the same distribution as the latent trait being measured (the true distribution of the possible scores a student can achieve) [8].

A more intuitive explanation is this: suppose we have μ_i, the average student test score in mathematics for student i. Instead of estimating μ_i alone, plausible values estimate a distribution of possible μ's for student i, together with the likelihood of each μ_i based on the respondents answers on the test. This is defined as the posterior distributions of μ's for student i. The reason why PISA uses this procedure is because estimating a single number μ_i is plagued with measurement error, among other types of bias (see Wu, 2005). The number of plausible values for PISA waves are usually five (although ten for PISA 2015) random draws from this distribution. In practice, each student has 5 scores for each test, which resembles their distribution. Those values are continuous, ranging from 0 to 500, with a mean of 250. However, PISA test scores were scaled to have a mean of 500 and a standard deviation of 100 over students in all OECD countries in the first year of focal testing (e.g. 2000 for mathematics and reading).

3.4.2 Coding and methodology

The aim of this chapter is to identify, disaggregate and explain country trends in the achievement gap for several countries. To represent the SES gap, most of the literature on achievement gaps has concentrated on indicators such as parental education, parental

[6]These rescaled indexes can be found at http://www.oecd.org/pisa/data/2015database/ under *Rescaled indices for Trend Analyses*.

[7]The analysis in this paper is mainly concentrated on Mathematics to be able to compare some of the findings with the existent literature which has predominantly focused on this subject. Literacy is used as a second test to check if the results hold. PISA also tests students in Science but since very little research has been done on this subject related to achievement gaps, it was not included in the analysis

[8]It should be noted that PISA has rotating modules for the main subject of that year. This means that the quality of data might be different for the same subject over time

occupational status, income achievement gaps and actual SES achievement gaps (Fryer and Levitt, 2004; Hanushek et al., 2006; Saw, 2016; Bradbury et al., 2015; Byun and Kim, 2010). The actual calculation of the achievement gap varies substantially and different strategies have been implemented. For example, Micklewright and Schnepf (2007) calculate the difference in achievement by crudely subtracting the difference between the 95th and 5th percentile of the mathematics distribution. Although in principle you should be able to capture some type of SES effect like this, theoretically, it should be much more accurate to difference out the mean score of, for example, parental education or some other SES proxy. Saw (2016), for instance, used parental education as a proxy of SES, whereas Byun and Kim (2010) use a similar SES index as the one used in this chapter, but created by them.

Reardon and Portilla (2015), Chmielewski and Reardon (2016) and Chmielewski (2016) used a different method developed by Reardon (2011), which I partially adopt in this chapter. SES achievement gaps are measured as the difference in standardized achievement between the average test scores for students at and above 90th and the average test score for students at or below the 10th percentiles of the chosen SES variable. The rule of thumb to choose the 90th, 50th and 10th percentile is arbitrary, as others have used, for example, the 95th, 50th and 5th (Micklewright and Schnepf, 2007). I use the conventional 90/10 cutoffs in the literature following the standard set by Reardon (2011) [9].

For each country in each wave, SES disparities in achievement are measured as the gap in standardized achievement between the 90th and 10th percentiles of each country's distribution of each SES variable, following the method for income achievement gaps by Reardon (2011). The *original* strategy of Reardon (2011) is as follows: first, achievement is standardized (see below for a statistical explanation of the standardization). I then use it to calculate the mean achievement (and standard error) for each category of the SES variable of interest (parent's education, income categories, etc..). "Category means are plotted at their percentile ranks and cubic models are fit through the points using weighted least squares." (Chmielewski, 2016) Finally, the average achievement is calculated for the

[9] It is important to highlight that I diverge from the analysis of Reardon (2011) because I calculate the gap between the average score of the students at and above the 90th percentile and the average score for students at and below the 10th percentile whereas Reardon (2011) calculates the gap between the 90th and 10th percentile

top and bottom 10%. The result is an SES gap from an ordinal variable of interest.

As mentioned before, PISA does not provide a single achievement indicator. Instead, I calculate the median of all plausible values for each student [10], resulting in one single score [11].

To *standardize* the test score I fit a linear model

$$y_i = \alpha + \beta_1 * AGE_i + \epsilon_i, \quad \epsilon_i \ N(0, \sigma^2) \tag{3.1}$$

for each wave, where y_i is the median student test score for student i and AGE_i is their age measured in months (following the same strategy as Reardon (2011) [12]) weighted by the student sample weights from PISA [13].

I then calculate $\hat{\gamma_i}$ by

$$\hat{\gamma_i} = \frac{\hat{\epsilon_i}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}} \tag{3.2}$$

where $\hat{\epsilon_i}$ is the residual for student i, $\hat{y_i}$ is the predicted test score for student i and the denominator is the root mean square error of the model.

This new standardized variable has a mean of zero. Standardizing the median test score solves the problem of comparability between different tests and across waves as the test scores have now the same metric across time. However, if the variance of the test scores changes drastically over time, then standardizing the overall score at each country-wave pair actually makes the transformation biased. That is, by standardizing test scores the variability is forced to be zero across all waves. But if the true deviations of the median

[10] Since each plausible value is a random draw from a theoretical latent normal distribution of possible student achievement scores, the median should be precise in getting a central measure of the latent distribution.

[11] The decision to calculate the median instead of using the traditional model that account for the uncertainty in the plausible values is because running these models is too computationally intensive. Moreover, the set of models available do not include multilevel models, something I use later on in the paper. For further information on this, see the documentation of the ‘intsvy‘ package from the R ecosystem

[12] This does not mess up the analysis by masking age-specific gaps as all students in the sample are 15 year olds. Controlling for age is simply to adjust for monthly differences in ages.

[13] I also tried to run the model for each country-wave separately but the results were very similar and it was more computationally expensive

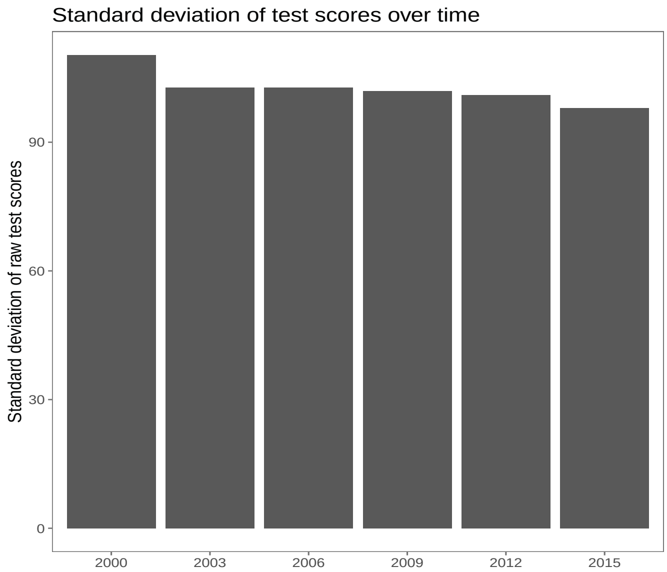


Figure 3.1: Standard deviation of test scores across all waves

academic achievement changes over time, then the estimated trend in the SES gaps will be underestimated, or viceversa.

I plot the standard deviation of the mathematics test score for all waves in figure 3.1. The plot suggests that it is something I should not be deeply concerned with, as the differences between countries are not very big with a not so drastic exception of the year 2000.

Another concern is whether test scores measured at different waves have different amounts of measurement error. If that is the case, then the amount of bias will not be the same in each measure of the gap. This can be misleading and suggest erroneous interpretations regarding trends of the gaps over time (Reardon, 2011). PISA has tried to make sure the tests are comparable across waves but it is still necessary to adjust for this imprecision (OECD, 2012). Accordingly, each PISA survey provides a reliability indicator for each of the tests which can be used to adjust for the reliability of the scores.

In order to correct for this I calculate λ_i which is just $\hat{\gamma_i}$ adjusted by the reliability indicator of each wave. More formally, I calculate it through

$$\hat{\lambda_i} = \hat{\gamma_i} * \frac{1}{\sqrt{r}} \tag{3.3}$$

where r is the reliability score of the test score in that PISA wave [14]. Note that I implement equation (3.3) separately by test scores and waves because there is a separate reliability indicator for each one. Once that is adjusted, the test scores should be roughly free of any bias in the trend that may arise from differential reliability of the tests.

In order to calculate the SES gaps it is necessary to estimate the thresholds for the 90th and 10th percentile. I calculate the thresholds using the SES index separately for each country-wave combination using the specific student sample weights of each one. I then generate a dummy of 1 for those above (including) the 90th percentile and 0 for those below (including) the 10th percentile for each country-wave pair. This means that all students that are below the 90th percentile and above the 10th percentile are excluded from the analysis.

I then fit a multilevel model:

$$\lambda_{ij} = \alpha_j + \beta_j * SES_i + \epsilon_{ij}, \text{ for } i = 1, 2, ..., n \text{ for each country } j \tag{3.4}$$

where SES_i is whether the student is at or above the 90th percentile (coded as 1) or whether it is at or below the 10th percentile (coded as 0). I allow α and β to vary by country j in order to obtain gaps for each country. I implement this model separately for each wave and weight by the wave-specific student sample weights. The previous model allows to calculate the achievement gap for each country by extracting the β's and α's for each country. I also calculate the standard error of this difference and generate uncertainty intervals.

I fit a multilevel rather than a linear model because by allowing the SES dummy to vary by countries, the gaps which have very little statistical power *borrow* strength from the other country samples by pooling information together. This is important because including the SES *dummy* reduces the sample size considerably given that only students above or below the 90th and 10th percentile are included in the analysis. Table 3.1 and

[14] Other procedures multiply each country by their own reliability measure for each year-subject pair (Chmielewski, 2016). The reliability estimates are calculated using Item Response Theory (IRT) analogues of traditional estimates of person separation reliability such as internal consistency. Unfortunately, PISA 2000 did not provide any reliability measure separately for each country and at the moment of the writing of this paper, PISA 2015 has yet to release their own. For these reasons, I implement the analysis following the original work of Reardon (2011)

table 3.2 in the appendix present the results of this model for mathematics and literacy. Moreover, figure 3.1 and figure 3.2 in the appendix plots the distribution of the random slopes of the SES dummy for both mathematics and literacy. By and large, these plots show the the slopes vary normally across all years.

Once the achievement gaps are calculated for every country in the six waves, the final dataset contains 183 observations as discussed in the 'Data' section. Until this point, the methodology described above is common to all research questions as this only standardizes and calculates achievement gaps. Research question 1 and 2 are addressed descriptively and thus do not require a description of a particular model or methodology; it is the calculation of trends and changes between countries. For research questions 3 and 4, the exact methodology is described below.

Considering the shortcoming of low sample size, modeling the differences in the achievement gap between countries might lack enough statistical power to generate stable and unbiased estimations. Once we consider the risks of modeling under such low statistical power, the best approach is to use a fully bayesian hierarchical linear model with informative priors. The benefits of this methodology are twofold: first, it is more intuitive when assessing uncertainty intervals as they truly represent the probabilities of the estimate being contained in the uncertainty interval 95% of the time. Moreover, it allows to specify prior information based on theoretical and empirical knowledge to counteract measurement error and uncertainty in such low sample size settings.

The empirical literature on tracking has concentrated on a very narrow definition of tracking by focusing only on the age of selection (Hanushek et al., 2006). I use a more fine grained definition of tracking, which is possible through the work of Bol and Van de Werfhorst (2013). Aside from the age at first selection into tracks, I also use the number of tracks in the country, the percentage of the entire curriculum that is tracked and a vocational index [15].

Below is an explanation adapted from Bol and Van de Werfhorst (2013) where they describe each indicator separately.

[15] This index is a factor loading from a principal factor analysis of the percentage of students in upper secondary vocational education (taken from two sources, to reduce measurement error). I take the data from Bol and Van de Werfhorst (2013).

Age of selection: These were taken from Economic Cooperation and Development (2005) and Economic Cooperation and Development (2006b) and simply reflect the age at which children separate into different tracks. The data are from years 2005.

Percentage of curriculum that is tracked: It indicates the length of the tracked curriculum and shows the share of primary and secondary education that takes place in tracked form. The data for this indicator are gathered for the year 2002 and are derived from Brunello and Checchi (2007).

Number of distinct tracks: The final indicator that Bol and Van de Werfhorst (2013) use is the number of distinct tracks available for 15-year-old students. Because differentiation takes place mainly in secondary education, the number of tracks that are available for 15-year-olds indicates tracking better than any other age. Data for the final indicator are provided by Economic Cooperation and Development (2005) and Economic Cooperation and Development (2006b). These are all from year 2006.

Finally, Bol and Van de Werfhorst (2013) created a tracking indicator. This index is constructed by performing a principal factor analysis on three country-level variables: (1) the age of first selection, (2) the percentage of the total curriculum that is tracked, and (3) the number of tracks that are available for 15-year-olds. All three indicators emphasize different aspects of tracking that are relevant, and it thus makes sense to use all three of them for the construction of the indicator of tracking.

Vocational orientation is calculated using two different but similar variables: the vocational enrollment (the same indicator but measured by two different sources) and the specificity of the vocational education. Data on the enrollment in upper secondary education are gathered by the Economic Cooperation and Development (2006a) and UNESCO for year 2006. The strength of the dual system and specificity of skills is measured by a single indicator: the percentage of upper secondary vocational education that takes place in a dual system. This indicator was taken from Economic Cooperation and Development (2007).

Both indexes have a mean of 0 and a standard deviation of 1 and the correlation between tracking and vocational indicators was 0.40 and 0.48, which shows that they are measuring different things (Bol and Van de Werfhorst, 2013). Moreover, the correlation between both vocational indicators is 0.54 (Bol and Van de Werfhorst, 2013).

For research question 3, I regress the achievement gap of a country on a dummy variable where 1 equals only 1 track vs more than 1 track, whether the age of first selection is 15 or more versus below 15, a dummy stating whether the percentage of tracked curriculum is above 0 (that is, any curricular tracking), the standardized vocational index and N - 1 year dummies to capture any yearly trend effects. [16]. This model allows the intercept to vary by country because there is not enough power to allow the coefficients to vary by year. For that reason, I include the yearly dummies, to adjust for between country differences. Note that each observation i is a year-country achievement gap. I am reluctant to include more variables in the model, first to preserve parsimony, and secondly to prevent overfitting due to the sample size. To be clear, this model allows to explain differences in achievement between countries and not modeling the evolution over time directly because the tracking and vocational indicators do not vary over time.

The final model can be expressed as

$$y_{ji} = \alpha + X_{ij}\beta + \epsilon_{ij} + \mu_j, \text{for } i = 2000,\ 2003,\ ...,\ 2015 \text{ years nested within each country } j \tag{3.5}$$

where y_{ji} is the achievement gap for country j and year i while $X_{ij}\beta$ is the matrix of $\beta's$. α and μ_j constitute the random intercept, where the group-level error parameter can be defined as

$$\alpha_j = \mu_\alpha + \mu_j \tag{3.6}$$

where μ_α is the mean intercept for all countries and μ_j is the country-level deviations.

The priors for all country intercepts, all of the $\beta's$ in the matrix X_i and the group-level variance parameter are assigned a t distribution with 3 degrees of freedom and scale parameter of 10, expressed as

[16] I create dummies to avoid multicollinearity. The final model has a maximum VIF of 2.2.

$$\alpha_j \sim \text{Student t}(3, 10)$$
$$\beta \sim \text{Student t}(3, 10)$$
$$\eta_j \sim \text{Student t}(3, 10)$$

These priors are chosen because they are weak enough to allow variation under the present uncertainty and allow the estimated quantities to be driven by the strength of the data. However, they force the estimates under plausible values closer to zero rather than infinity.

A more summarized specification is to actually use the index of the tracking indicators, something that Bol and Van de Werfhorst (2013) calculate for their data. Using this standardized tracking variable and the same bayesian multilevel model, I model the tracking and vocational indexes but add an interaction between the two, given that the second hypothesis looks to test whether these two features explain the differences between countries.

This model has the same priors for the varying intercept and the group-level error term but since these two indexes are a composition of the previous variables, I slightly adapt the prior to be normally distributed with varying means for the three $\beta's$ and standard deviation of 0.1, such that:

$$\text{Tracking index} \sim \text{Normal}(0.3, 0.1)$$
$$\text{Vocational index} \sim \text{Normal}(-0.2, 0.1)$$
$$\text{Tracking index * Vocational index} \sim \text{Normal}(-0.15, 0.1)$$

For the tracking index, past research has shown it to be positive (increase inequality) (Hanushek et al., 2006), so I provide a mean for the distribution of 0.3 with a standard deviation of 0.1. This allows for small positive effect and a zero effect about 10% of the time. The vocational index is less clear, so I provide a mean of -0.2. That is, vocational enrollment would decrease inequality as it helps children from the lower tracks achieve other types of jobs otherwise unavailable. For the interaction term I provide a mean of -0.15, an even smaller estimate to be conservative and let the data drive the interaction. This last distribution allows for positive values in the interaction term in about 10% of the time.

It is important to discuss the implications of the standardization both of the test scores and of the SES gaps. The first model presented here standardizes test scores for all countries within each wave, so that the mean of each wave comes from all countries pooled. However, the estimation of the SES gaps is done through a random slope for each country and not through a pooled SES mean. The test scores were standardized across a pooled country mean because test scores were indeed in the same metric across countries and are not susceptible to other influences such as economic inequality. However, the within country standardization was done so given that it reflects reality much more than calculating it within a single pooled SES mean, which would suggest that, for example, high SES groups from Sweden and Denmark should be pooled together with high SES groups from Turkey.

Given these two different transformation, the interpretation of results is slightly different as the position of each SES group is relative to the position of the other SES group within that country and not relative to the pooled SES mean across all countries. This has implications in the interpretation given that for example, variation in the average scores of the low SES group in Greece could be reflecting differences in the economic and social inequality between SES groups in that given country. On the other hand, the inequality between SES groups in Spain could be different from gap in Sweden not because they score differently but because of the inherent initial differences between SES groups in that country. This means that when comparing point estimates between SES groups across countries, we should keep in mind that this could reflect inherent differences between where these SES groups are in the first place. An SES gap of a particular size in standard deviations may have a different meaning depending on the initial achievement of that country.

Finally, until now the methodology has outlined how we can explain the differences in the 90/10 achievement gap between countries where each observation inside the country records the achievement gap in the 15 year-span under study. This approach was adopted given that the main independent variables, curricular tracking and vocational indexes, do not change over time, making it impossible to model the change over time directly. For this reason, the previous models explained differences between countries rather country-specific evolutions over time. However, there is another way of estimating the evolution of the achievement gap, which I develop next. The paragraphs below explain the methodology for research question 4.

Instead of explaining the evolution over time directly, the third model calculates the difference between the last year and the first year for each country, summarizing the 6 year-time span of achievement gaps into one cumulative inequality gap. That is, each country has one final gap which is calculated by the difference between the years 2015 and 2000 in the average student test score for those above (and at) the 90th percentile and below (and at) the 10th percentile in the SES index. Given that the variable is continuous, I fit a linear model to test whether the independent variables are linearly related to the 'cumulative inequality gap'. This linear model has standard weakly informative priors as the ones used in section 3.4.2 given that this specific question has not been researched before.

This approach dramatically reduces the total sample size to 32, effectively the number of total countries available [17]. It is for this reason that I only include at most 2 covariates in the model, the tracking and vocational index, which summarizes the previous models and keeps it parsimonious. The results are presented for the 90/10, 80/20, and 70/30 achievement gaps to test for robustness. In the next section I present the results.

3.5 Results

3.5.1 Evolution of the achievement gap

Table 3.1 shows a description of the sample size, mean score of the top and bottom SES groups and SES gap for only the first and last time point. One main concern from the planned analysis is that using the top 90th percentile and bottom 10th percentile would result in a small sample size for some countries. This table suggests that the data has a reasonable number of respondents to actually estimate gaps accurately. Moreover, we can see that in all instances the bottom SES group has a lower score than all top SES groups.

[17] I have been presenting a selected number of countries given the lack of space, but for every model/estimation I include all countries available.

		Low SES			High SES			
Year	Countries	N	Avg score	S.E	N	Avg score	S.E	SES gap
2000	Australia	294	0.22	0.07	270	1.22	0.09	1.00
2000	Austria	294	-0.04	0.11	235	1.14	0.15	1.19
2000	Canada	1970	0.30	0.05	1357	0.90	0.07	0.60
2000	Denmark	230	-0.01	0.13	228	1.16	0.17	1.18
2000	Finland	261	0.42	0.12	267	0.89	0.15	0.47
2000	France	253	0.03	0.04	258	1.20	0.05	1.17
2000	Germany	250	-0.64	0.04	303	1.80	0.05	2.44
2000	Hungary	291	-0.46	0.09	254	1.65	0.12	2.11
2000	Italy	278	-0.31	0.05	278	0.80	0.06	1.12
2000	Netherlands	129	0.46	0.08	158	1.24	0.11	0.78
2000	Poland	209	-0.31	0.04	166	1.25	0.06	1.56
2000	Spain	318	-0.28	0.05	340	1.09	0.07	1.37
2000	Sweden	241	0.12	0.10	246	1.06	0.13	0.94
2000	United Kingdom	531	0.15	0.04	415	1.32	0.06	1.17
2000	United States	222	-0.43	0.02	164	1.74	0.03	2.16
2015	Australia	1694	-0.15	0.04	1235	1.28	0.05	1.43
2015	Austria	676	-0.30	0.07	705	1.53	0.10	1.83
2015	Canada	2215	0.20	0.03	1863	0.99	0.05	0.79
2015	Denmark	1013	0.09	0.08	595	1.08	0.10	0.99
2015	Finland	575	0.07	0.08	584	1.16	0.11	1.09
2015	France	570	-0.37	0.02	615	1.66	0.03	2.03
2015	Germany	545	-0.09	0.02	582	1.45	0.03	1.54
2015	Hungary	466	-0.72	0.07	589	1.79	0.09	2.51
2015	Italy	947	-0.20	0.03	1043	1.23	0.04	1.43
2015	Netherlands	521	0.13	0.04	525	1.24	0.06	1.11
2015	Poland	446	0.10	0.03	448	1.24	0.05	1.14
2015	Spain	608	-0.25	0.03	703	1.24	0.04	1.49
2015	Sweden	527	-0.20	0.06	542	1.32	0.09	1.52
2015	United Kingdom	1387	-0.03	0.03	1388	1.15	0.04	1.18
2015	United States	585	-0.39	0.01	544	1.22	0.01	1.62

Table 3.1: SES sample size and ISCED composition

In table 3.4 in the appendix, I also show the same descriptives for all countries. The composition of the sample is mainly from industrialized countries, with the majority being European and Anglo-Saxon countries. There is only one country coming from developing regions, Chile. Aside from these countries there is also Japan, Israel and New Zealand, which are neither Anglo-Saxon nor European. This description then confirms that the analysis must be interpreted exclusively in the developed parts of the world. All of the interpretations should only be extrapolated to industrialized nations.

Given that the SES gaps were calculated within countries, the interpretation is relative to the country mean rather than the pooled mean for all countries. For this reason, we will compare the SES gap included in the last column of the table. The Germany and the U.S seem to have the biggest gap in 2000, with an average inequality of 2.44 and 2.16 standard deviations respectively and Canada seems to have the a very small gap with a standard deviation of 0.6. In contrast, we see that Denmark and Finland boast small gaps across the time span with gaps of 1.18 and 0.47 standard deviations respectively for the year 2000 and 0.99 and 1.09 for the year 2015. Sweden on the other hand, increased it's gap from 0.94 to 1.52 to similar levels to Spain.

In order to make sure the composition of the top SES group and bottom SES group is as expected, I present table 3.3 in the appendix. This table confirms that both groups are extremes in terms of education, with the top SES having highly educated parents while the bottom group has lower educated groups. More concretely, 96% and 93% of the top SES group have bachelor's degrees or above for 2000 and 2015 while 0% and 1% of the respondents in the bottom SES group have parents with bachelor's degrees or above for the same years. In addition to this, these groups not only differ in education but on other dimensions such as immigration and human capital measured as books in the household. These results confirm that both groups measure what they are supposed to: both extremes of the SES gradient. This table is presented for all countries pooled for years 2000 and 2010 to confirm that the results from the modeling section are not due to changes in the composition of the groups. The results indeed show that the composition is roughly the same between the two time points.

In table 3.2 I present the data related to the tracking setup of a few selected countries. The data was obtained from Bol and Van de Werfhorst (2013).

Countries	# of tracks	Age of selection	% of curric tracked	Std. Voc	Std. tracking
Australia	1	16	0.15	0.97	-1.04
Austria	4	10	0.67	1.7	1.82
Canada	1	16	0	-1.72	-1.32
Denmark	1	16	0.25	0.45	-0.87
Finland	1	16	0.25	0.74	-0.87
France	2	15	0.25	0.39	-0.47
Germany	4	10	0.69	0.89	1.86
Hungary	3	11	0.67	-0.7	1.42
Italy	3	14	0.38	0.95	0.17
Netherlands	4	12	0.45	1.26	0.94
Poland	3	15	0.38	0.3	-0.08
Spain	1	16	0.17	0	-1.02
Sweden	1	16	0.25	0.69	-0.87
United Kingdom	1	16	0.15	0.47	-1.04
United States	1	16	0	-1.84	-1.32

Table 3.2: Curricular tracking statistics for selected countries

It is evident that there is ample variation in the tracking structure of countries. Germany and Austria have highly tracked curriculums, early age of selection and high number of tracks, whereas countries such as the U.S. and Canada have the opposite. The last two columns of the table show a standardized index created by Bol and Van de Werfhorst (2013). The first one measures the degree of vocational specificity of a country and the last one the degree of curricular tracking. Both have a mean of zero and positive values mean higher tracking and vocational enrollment while negative values mean the opposite. Looking at the results more closely we see similar patterns: European countries have higher vocational specificity and curricular tracking than Anglo-Saxon countries. Two curious cases are Hungary and the Netherlands. The first one shows low vocational, while the other has high levels of vocational enrollment. Yet Hungary also has high levels of curricular tracking. In summary, table 3.2 shows by and large that countries have developed different tracking structures and roughly speaking there is a large divide between European and Anglo-Saxon countries, something missing in Bradbury et al. (2015). In table table 3.5 in the appendix I provide data for all countries, which is the complete data used in the modeling section.

Next I look at the evolution of the achievement gap [18]. I start by looking at the achievement for a selected number of developed countries in figure 3.2. I plot only mathe-

[18] The achievement gap refers to the average score of the 90th percentile minus the average score for the 10th percentile

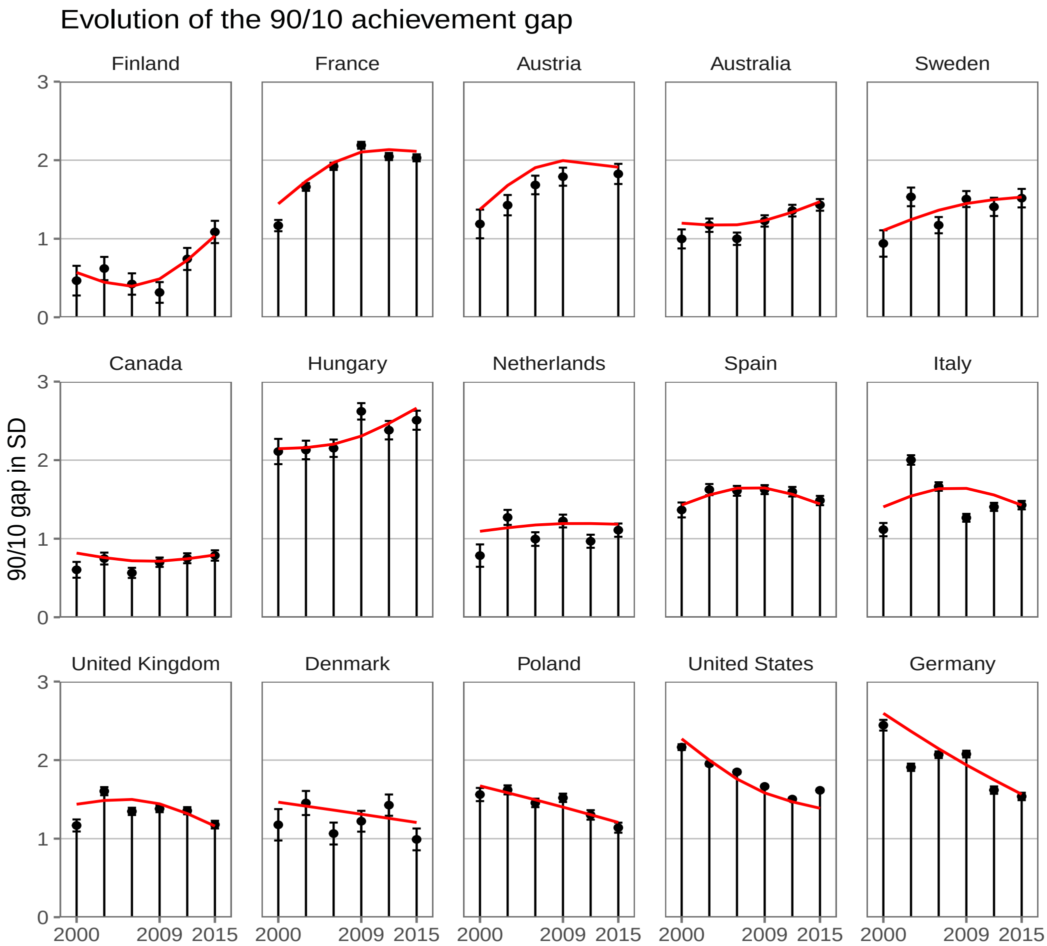


Figure 3.2: 90/10 achievement gaps in mathematics expressed as standard deviations for selected countries between 2000 and 2015

matics for each country (dark dots) but also a quadratic trend spline for *both mathematics and literacy* pooled. The pooled trend spline is calculated for both tests in order to test whether the results hold under stricter specifications. By comparing the dots (only mathematics) and the trend spline (mathematics and literacy) we can gauge how far reading is from the mathematics result. The trend spline is calculated by calculating the median of the two scores at each time point. The same plot is available for all countries for mathematics and reading in figure 3.6 and figure 3.7 in the appendix. In these appendix plots, the splines do not reflect the combination of both subjects but only the subject of their respective plots.

As we can see from the results, some countries have increased their achievement strongly. For example, France, Austria and surprisingly Sweden have very steep slopes. France experienced an increase in inequality by roughly 0.9 SD, Austria by 0.6 and Sweden by 0.6. This pattern happens similarly for literacy as can be seem from appendix table figure 3.7. For example, France has an increase of 0.6. For such a short period of time, the magnitude of these increases are reasonably big.

Given that no one has estimated the evolution of the gap I cannot cross-check how other empirical estimations put France at. However, the work of Micklewright and Schnepf (2007) is the closest reference available which also finds that France was a low dispersion country in 2000; there is no evidence on what happened over time. Fortunately, the work of Bernardi and Ballarino (2016) did study social origin inequalities (broadly speaking, not in terms of achievement gaps) in France and found that they increased since the 2000's.

Other countries have reasonable increases such as Finland and Hungary, with increases of nearly 0.6 and 0.4 standard deviations respectively. Aside from these countries, there are other countries which experience no changes at all, specifically, Canada, Netherlands and Spain. Canada excels here not only because the gap has been stable over time, but because it has the smallest gap of all countries presented here. It is nearly 0.5 SD in 2000 and it increased only by 0.2 in 2015.

On the other hand, there are other countries which experience a decrease in the SES achievement gap. Poland decreased by about -0.4 and Denmark by -0.2. However, the most notable cases are the U.S. and Germany. These two countries show high levels of dispersion in the year 2000 with SES gaps of over 2 SD. But in the 15-year time trend both countries reduced the gaps by -0.6 and -0.9 respectively. Their distinctively large gaps in 2000 also show up in the work of Micklewright and Schnepf (2007). This finding is similar to the one in Reardon and Portilla (2015), in which they found a decreasing gap for kindergartners. It is important to highlight that the cohorts in their analysis are different from the ones in this study but also reassures that evidence close to the cohorts in this study also found a decline. The decline for 15 year olds found in the plots from above could suggest that it might be more of an institutional change rather than a specific grade-level policy. Do note that the trend line is for mathematics and literacy pooled. If it were only for mathematics, then the trend line would be even more pronounced. Despite this, pooling the two subjects gives more strength to the analysis by showing that both tests are in the same direction. I also plot the 80/20 and 70/30 SES gaps and find similar patterns (see figure 3.10 in the appendix).

Analyzing figure 3.2 the reader may get the impression that these trends are not very steep and they should not be relevant in practical terms. However, note that the Y axis is measured in standard deviations. Small changes are actually large in practical terms. For example, evidence from PIRLS shows that the predicted growth of a student for a year of school is of around 0.30 standard deviations (Beaton et al., 1996). PISA has also documented this type of metric in their annual reports (OECD., 2009). Take the case of Sweden. The slope does not look that steep but in reality it increased the gap from 1 SD in 2000 to around 1.5 SD in 2015. With that information in mind, the trends of Poland, United States, France and Germany are gain practical relevance.

I find that the initial gap for the U.S. in 2000 is 2.16 standard deviations of magnitude varying between 2.13 and 2.2. This gap is much higher than found in previous studies because the 90/10 gaps defined here are more like the 95/5 gaps found in Reardon (2011). This is the case because Reardon (2011) used the percentiles for the 90/10 gaps whereas I defined the mean for students at and above the 90th and at and below the 10th. There is also two differences from previous research. First, past research has never really concentrated on the 15-year old achievement gap but rather on student achievement gaps while adjusting for age which forces the age effect to disappear. Secondly, past research has concentrated only on the SES income gap whereas here I use a a different index which uses parent's education, an occupational SES index, among other things. These reasons, either working together or in isolation could explain the difference in the magnitude of the U.S. gap. However, the methodological approach used here is very similar to Reardon (2011), with the main caveat that the author adjusts for age.

Figure 3.3 takes a more direct approach and looks at the percentage change from the first and last time point available. Each data point has been computed together with its 50% and 95% uncertainty interval [19].

About 10 countries increased their achievement gap over time (see the two first row of subplots in figure 3.6 in the appendix). France had an average increase of about 80% since 2000 varying down to 40%, whereas Germany had a similar figure but decreasing. Many of the countries that did not show a positive slope, such as Hungary or Australia, had in fact increases of about 40% in their gap. In contrast, the U.S. and Poland had

[19] Each of these uncertainty intervals were computed using a 500-replicate bootstrap

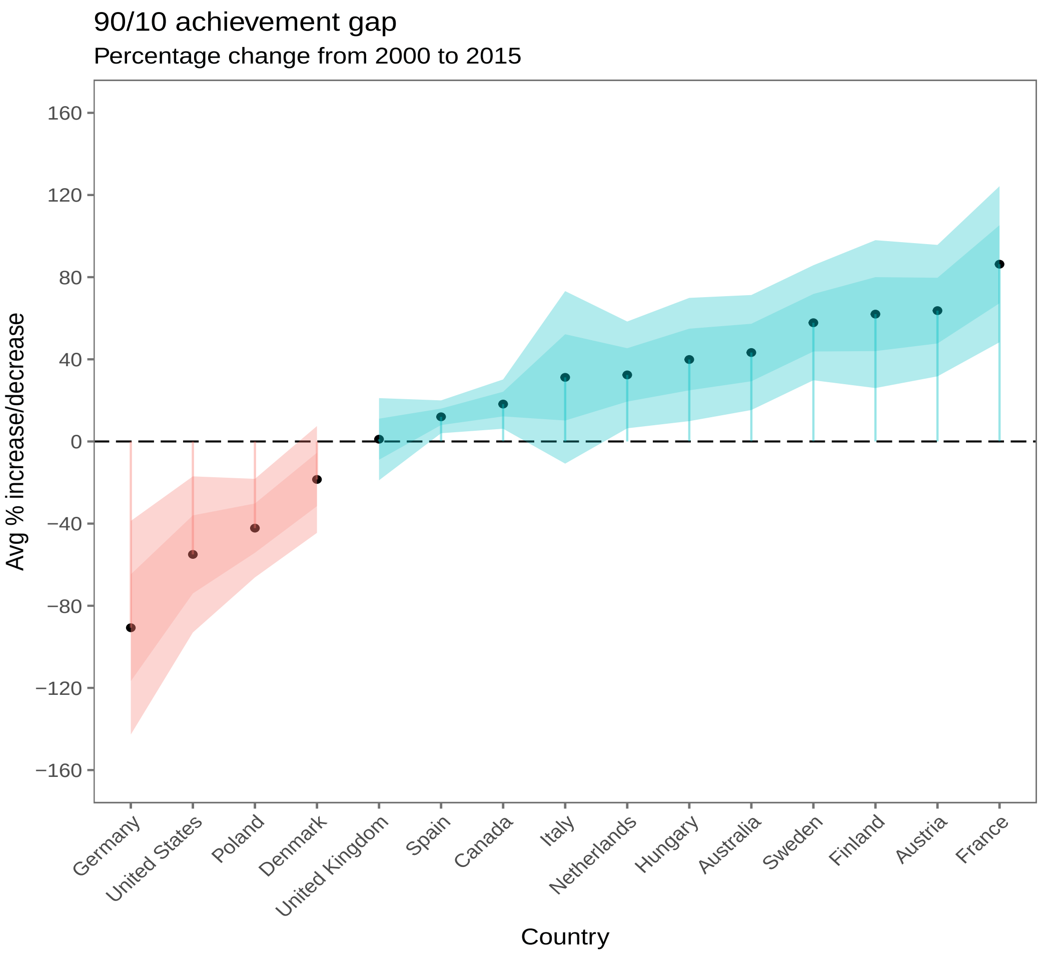


Figure 3.3: Percentage change in mathematics between years 2015 and 2000 for selected countries. Red regions represent decreases while blue regions represent increases in the achievement gap

also significant decreases of about 40%. The benefit of presenting these estimates this way is that the reader can actually assess the uncertainty of each calculation, and there is evidence that they have wide variability. Despite this, most countries show a clear sign of either decreasing or increasing. I plot the same graph for the 80/20 (figure 3.11) and 70/30 (figure 3.12) SES gaps in the appendix and the results hold for these achievement gaps as well.

Something reassuring is that mathematics and reading (figure 3.7 and figure 3.8 in the appendix) follow basically the same trend across the presented countries. This means that the result is not an artifact of chance alone.

The first question in the research/hypothesis section asked how does the achievement gap behave in terms of its evolution over time and between countries. The results show that the gap is very different between countries. There is ample variability across countries with some countries that have different educational systems experiencing similar trends in achievement gaps. Countries as disparate as the U.S. and Germany follow similar patterns while other such as France and Austria experience steep increases. Equally important, there are strong and visible changes in the evolution of the gap. This highlights the fact that achievement gaps are very variable and very context-specific to their educational systems. Moreover, the results of Reardon (2011) are different from the ones presented here. The main difference could come from the fact that Reardon (2011) has surveys for many different age groups and adjusts for age, eliminating the age-specific effects. But more over, the results here come from an SES index which uses information on the parent's educational level and other SES indicators such as an occupational hierarchy. These results are the first to document such starking contrasts for the achievement gap of 15 year-olds.

However, it is important to disentangle where changes in the gaps are originating from. Is this because the top are improving while the bottom decreases? Or is it that the bottom is catching up? These results are important because they help pinpoint whether countries experiencing similar changes are indeed originating from the same source. Next, I plot the same graph but show the divergent patterns between high/low SES origins.

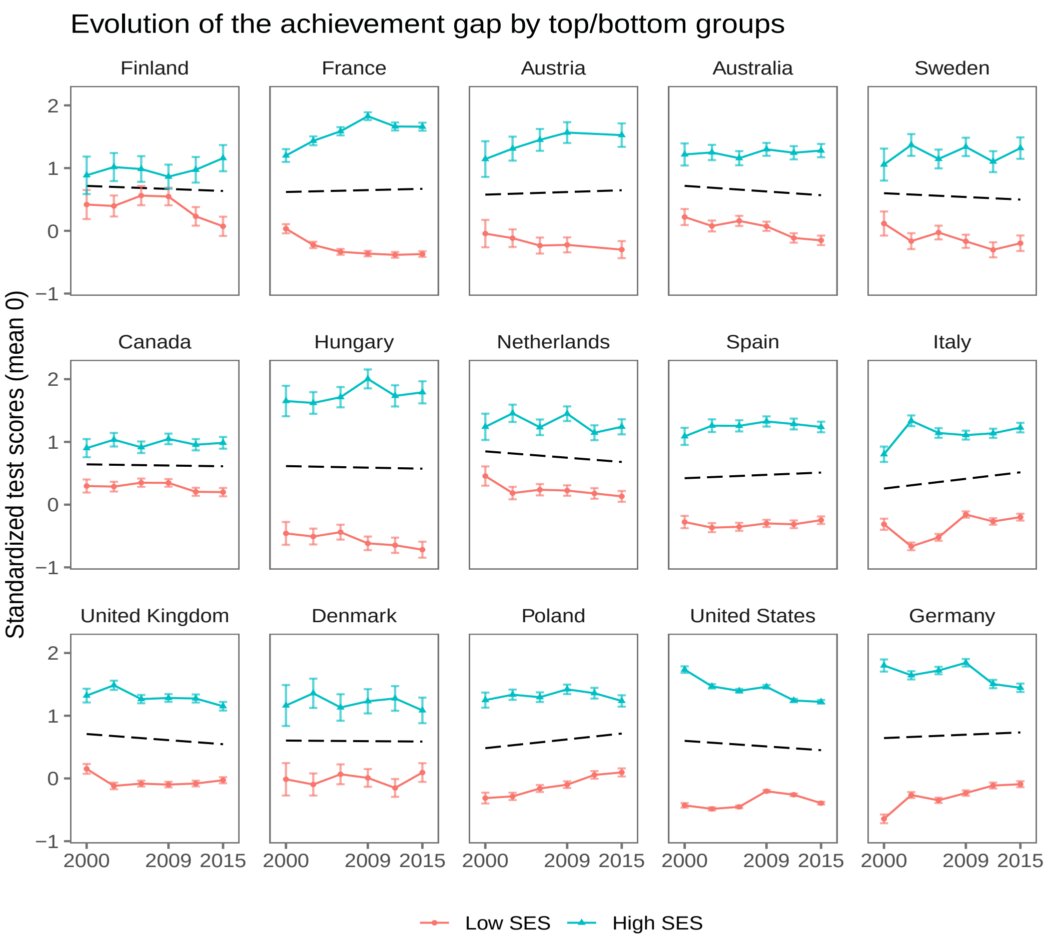


Figure 3.4: Evolution of the achievement gap for separate SES groups

3.5.2 Evolution and differences in achievement between top/bottom groups

Figure 3.4 shows the evolution for low and high SES groups separately. The middle line is a linear trend, calculated by averaging out between the top and bottom point for each year, showing whether one of the groups is growing/narrowing faster than the other. For example, in the U.S. the top seems to be declining much faster than the bottom is increasing. The UK seems to be following the same path as the U.S. as well. On the other hand, in Poland the gap between the two groups seems to be closing as the bottom SES group is catching up much faster. The Netherlands shows a similar pattern as the U.S. and the UK patterns but upon closer inspection the explanation is different. The slope is negative (like in the previous two countries) but that is because the low SES group is decreasing at a greater rate than the high SES is decreasing.

Moving on to the flat summary lines, these should be interpreted with caution as it does not mean that the gap is not changing. Denmark, Belgium, Netherlands, Norway and Spain show a flat line as the gap is growing little, if at all. On the other hand, Sweden, France, Finland, New Zealand, Austria, among other countries, show a flat line as groups are distancing themselves at a similar rate. These results highlight the importance of not only summarizing average achievement gaps: the source of the gap varies greatly between countries and it is easy to see how each of these patterns contributes to the overall inequality of a country.

After analyzing the trend of each country and where it is coming from, interesting results start to emerge. The U.S. is closing the achievement gap at a rapid rate, but it is because the top SES group is going down faster than the bottom is increasing. In contrast, countries such as Germany and Poland show a closing gap as well but it is because the bottom group is catching up faster than the high SES group is coming down. In fact, the Polish case is the most egalitarian of all: the bottom group is catching up at a rapid pace and the top group is maintaining their high levels of performance, fostering a truly equalizing effect.

In countries where the gap is widening, both groups are distancing themselves at a similar pace. Finland, France and Austria, show this exact pattern. The results seem to show a solid pattern: when the gap widens it is because both SES groups are distancing

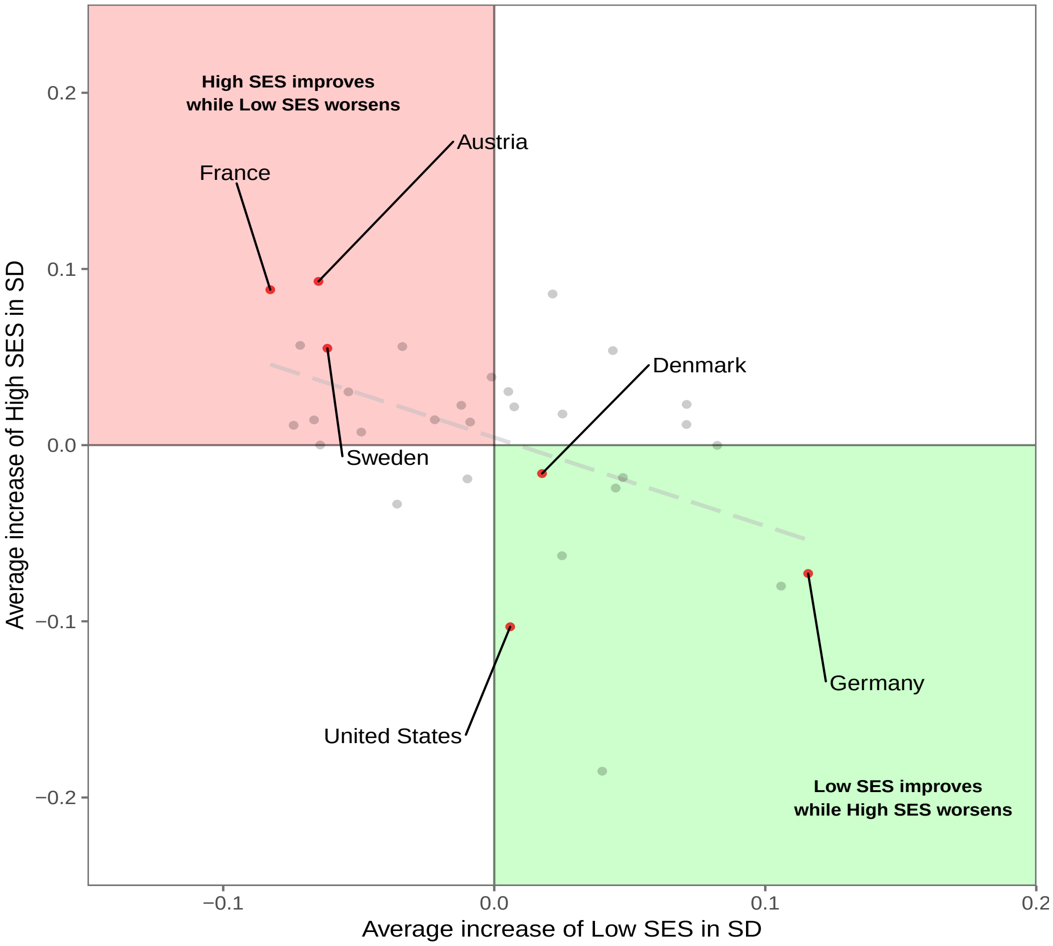


Figure 3.5: Rate at which top/bottom SES groups are changing for the mathematics 90/10 SES gap

from each other. But when the gap is being reduced the source can vary between both groups narrowing or just one of the two leading the decrease.

In a more convenient way, we can inspect whether there is a lot of variance in the rate at which top/bottom groups are changing in figure 3.5. For this plot I calculate the average increase for the high SES and bottom SES groups for all 32 countries across all years. For each SES group separately, this is done by calculating the lagged difference between years and calculating the average of these differences. For example, for the high SES group, if the average is positive, it means that the average of the year after year differences suggest that the high SES group is increasing. Similarly for the bottom SES group.

Moving on to the interpretation, figure 3.5 shows that there is a significant number of countries where high SES students are increasing their scores more than the low SES group is increasing (top-left panel). However, there is also a fair share of countries where the low SES group is increasing more than the high SES group, such as the U.S. and Germany (bottom-right panel). The plot reveals an overall negative relationship which is puzzling as it suggests that countries seem to be polarized on the rate at which their SES groups improve their achievement. To be clear, the plot shows the average increase/decrease at which the top SES group is changing in their average scores across in standard deviations on the Y axis and the rate at which the bottom SES group is increasing their average scores in standard deviations on the X axis.

This is interesting as there seems to be a tradeoff between improving either the low SES group or the high SES group but not both at the same time. The previous evidences a clear lack of improvement in achievement across all SES groups; most improvements for one SES group seem to prevent to be associated with the other SES group not improving. Figure 3.3 in the appendix shows exactly the same plot but highlights the name of all countries so that is clearer which countries are in which quadrant.

When we contrast all country names from Figure 3.3 in the appendix and the tracking profile for each country in table 3.5 in the appendix, countries within the same quadrant share very distinct tracking profiles. For example, for the upper-left quadrant Sweden has very little tracking compared to France or Austria, which is a high profile tracker with over 3 tracks and separating students as early as 10 years of age. A similar example is Czech Republic, Ireland and Finland. Czech Republic has 5 tracks and an early selection at age 11, one of the youngest early selection ages in Europe. Ireland on the other hand, has 4 tracks but an age of selection at 15, an age considered to be of late tracking. Finally, Finland is a complete opposite to the two previous countries with only 1 track and an age of selection of 16 years of age. Similarly, countries in the bottom right panel have also very different tracking setups. For example, Germany and the U.S, where the first one is an early tracker and rigid tracking system while the U.S has no formal tracking setup. Moreover, there are other countries such as Poland, with 3 tracks and age selection of 15 years, which share the quadrant with Denmark (1 track, age selection 16), Turkey (3 tracks, age selection 11), Slovakia (5 tracks, age selection 11) and Bulgaria (2 tracks, age selection 14). All of these results highlight that there is indeed a high degree of tracking variability within opposite quadrants. The relationship between the evolution of the achievement gap and tracking seems to be mixed as there isn't a clear pattern between increase/decrease of the gaps and the tracking profiles of each country.

The plot also shows very few countries where low SES are decreasing and high SES is also decreasing (bottom left panel). This can be interpreted as either one of the two SES groups increases while the other decreases. Interestingly enough, there seems to be also very few countries where both SES groups are increasing (top right panel), suggesting that on very few cases do SES groups increase their performance on par.

The second research question was interested in finding out whether the achievement gap was changing because the top was gaining ground, the top was falling behind or because of a dynamic interaction between the two. The hypothesis stated that in countries where there is high degree of tracking we should expect the gap to be increasing because both groups are distancing from each other. Conversely, in countries with little tracking we should expect for the gap to be narrowing. The results presented here shows mixed evidence in favor of the hypothesis. There does not seem to be a clear cut pattern between the tracking profiles of each country and the evolution of their achievement gaps. In other words, the results shown here confirm that the curricular tracking setup of a country does not imminently mean a certain degree of equality/inequality.

More concretely, the evidence suggests that countries where the top SES group is increasing and the bottom is decreasing (top-left panel) have different curricular tracking setups. Austria is an early tracker with several tracks, while Sweden is a late tracker with only 1 track. On the opposite panel (bottom-right), we can see countries where the bottom group is catching up while the top is coming down. The results are also similar as we see countries such as Denmark and Germany which are at opposite poles in terms of tracking, while the U.S. is distinctive in that it has little formal curricular stratification. The results do not show definitive evidence of top-to-bottom equality in all of these countries (except for Poland). However, an interesting finding clearly arises from these high-level plots: regardless of the different tracking setups, there is seems to be a trade off between SES groups, either the bottom comes up and the top comes down or the top goes even further up and the bottom further down.

To the best of my knowledge this is the first time this evidence has come up in the literature. The reason why this pattern seems to be present in over 30 countries is unknown and should be followed in future research.

3.5.3 The role of curricular tracking in explaining the achievement gap

We now move to the modeling section of the chapter, where I test a possible explanation to differences in the achievement gap between countries. The reason why countries differ in their evolution of inequality is still unknown. There is ample evidence showing the inequality between countries, to a certain extent, can be explained by the degree of tracking (Hanushek et al., 2006). But I am also interested in what explains the between country evolution of inequality, that is, the explanation as to why in certain countries it is increasing more than others over time.

Using the first bayesian model explained in the methodology section, I present the results of the first model in table 3.3.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Only 1 track		-0.53 (-0.81, -0.23)	-0.32 (-0.63, 0.01)	-0.34 (-0.67, 0.02)	-0.21 (-0.53, 0.12)	-0.21 (-0.55, 0.14)
Age selection >= 15			-0.4 (-0.72, -0.1)	-0.39 (-0.71, -0.07)	-0.58 (-0.9, -0.26)	-0.59 (-0.91, -0.25)
% of curric tracked				-0.03 (-0.59, 0.56)	0.58 (-0.13, 1.28)	0.58 (-0.12, 1.31)
Vocational Index					-0.26 (-0.46, -0.06)	-0.27 (-0.47, -0.06)
Year 2003						0.2 (0.11, 0.29)
Year 2006						0.1 (0, 0.19)
Year 2009						0.11 (0.02, 0.2)
Year 2012						0.08 (-0.01, 0.17)
Year 2015						0.06 (-0.03, 0.15)
Between-group variance:	0.27	0.2	0.18	0.19	0.16	0.16
Sample size:	183	183	183	183	183	183
Number of groups:	32	32	32	32	32	32
Intercept	1.59 (1.43, 1.76)	1.76 (1.6, 1.92)	1.94 (1.75, 2.14)	1.96 (1.32, 2.57)	1.56 (0.93, 2.22)	1.48 (0.81, 2.14)

Table 3.3: Explaining 90/10 achievement gap - Multilevel model with intercept varying by country

I present uncertainty intervals next to each coefficient. The first model includes an empty model to record the initial variance between countries, which is 0.27. The second model only includes the number of tracks in the country. The variable shows that countries that have one track have on average -0.52 SD of less inequality than other countries. This covariate alone explains about 26% of the variance between countries of the 90/10 achievement gap as the variance decreased from 0.27 to 0.20. The third model includes the dummy for whether the country has an age selection of 15 or more and the results show that countries with late selection into curricular tracking have about -.39 less SD in inequality. Also note that the coefficient for the number of tracks was reduced, suggesting these two variables are explaining similar things. This is expected given that late tracking is usually associated with fewer tracks, such as in the Scandinavian countries. The fourth model now includes whether the country has any degree of curricular tracking, meaning whether the country has above 0% of the curriculum tracked. The coding of the variable comes from visual inspection of the distribution, where some countries, such as the U.S. and Canada, have untracked curriculums. For this last variable there does not seem to be a relationship, but after including the vocational index in the fifth model, the coefficient is completely reversed with countries with tracked curriculums having about 0.59 more SD in inequality than countries with no curricular tracking. The last model includes the yearly dummies to adjust for any trends over time and the results seem unaffected by this.

It is important to highlight the fact that it is not only the case that "late tracking is usually associated with fewer tracks, such as in the Scandinavian countries" but that by definition countries with only one track must be classed as having late tracking. In Model 3, compared to the baseline of countries with ¿ 1 track and selection age ¡15, the combined effect for all countries with one track is -.32 - .39 = -.71 SD. The difference between Models 2 and 3 is being driven mostly by allowing a separate effect for countries with more than one track and a selection age of exactly 15.

Also note that the vocational index is associated with less inequality. Not surprisingly, the age of selection coefficient also increased due to the correlation between age of selection and vocational index. Countries with high levels of vocational index (vocational enrollment and specificity) are also those with early age selection, such as Germany and Austria. These two variables should be inspected further, perhaps with an interaction, given that the coefficients are sizable and correlated. This final model explains about 40% of the variance between countries in the 90/10 achievement gap as the total variance was reduced from 0.27 to 0.16. This is not a trivial percentage and suggests that these institutional factors are playing an important role in defining country differences in the achievement gap.

It is also important to highlight that as the model increases complexity (addition of variables), the estimated coefficients become more uncertain. It is true that all the coefficients point in the expected direction, even with a weakly informative prior, which suggests that the data has reasonable strength in influencing the estimation. However, the uncertainty intervals suggest that they vary considerably. For example, the coefficient for the percentage of curriculum that is tracked varies all the way from -0.10 to 1.27. This suggests that the coefficient is most likely to be positive as most of the region lies inside the positive spectrum yet there are possibilities that it varies down to -0.10. In contrast, the coefficient for the age of selection varies between -0.92 and -0.26. This coefficient shows a wide interval but only concentrated on the negative region, suggesting that it is highly probable that it is negative and strong.

It is important to highlight that we should not dismiss this finding as irrelevant simply because any of the estimates includes zero, as the uncertainty intervals need to be interpreted as a measure of how much we can trust our results. The explanatory measures presented here such as the variance between countries should be also interpreted with caution. The country-level variables presented here are only one set of possible explanations. Other explanatory variables might, and should, be correlated with the curricular tracking variables, making these estimates endogenous to other explanations.

All of these countries have a different curricular tracking structure, with some setups being more egalitarian than others. Using the previous model I make a simulation and predict the level of the 90/10 achievement gap for all logically possible curricular tracking setups. In figure 3.6 I plot all these combinations except the scenario '1 track and '<15 age' because it is logically impossible for a tracking setup to have more than 1 track and an age of selection of the less than 15 years of age; these two setups are mutually exclusive. Moving on to the plot, the X axis shows the number of tracks that a country has, the Y axis has the predicted 90/10 achievement gap and below each country name it is specified whether the country has an age selection of 15 or more or below 15. In addition, to confirm how accurate the model is, I also plot the actual level of inequality in the actual curricular tracking setup of the country (the red star). Let us take Germany for example.

The star point for Germany is in the '>1 track' and '<15 age' setup, meaning that Germany has a tracking system that has more than 1 track and an age of selection of below 15. We can see that the model predicts the level of inequality for that set up accurately (height of the bar), relative to the actual level (the star point). If the German tracking setup would be in the ideal '1 track' and age of selection above or equal to 15, then inequality would be lower.

First, this plot highlights that the model seems to fit the data reasonably well as the estimated prediction stars line up very close to the actual achievement gap of the country. All in all, the model does a reasonable job at capturing the achievement gap of each country. On the substantive side, this graph reveals some interesting patterns. Generally speaking, we can see that the simulation predicts that virtually all countries which are not in this 'ideal' tracking, that would switch to the 'ideal' setup, would experience a reduction of the achievement gap. The opposite is also true. Countries which are in the 'ideal' setup that would change to the 'worst' setup would be associated with a widening of the achievement gap. On average, I find that if all countries switched to the 'ideal' setup, the switch would be associated with a reduction of 11% of the achievement gap with some countries experiencing even a 30% reduction. This is an important estimate considering that Reardon (2011) found that the US gap increased by about 40-50% in 30 years. Had all countries in the 'ideal' tracking switched to the 'worst' curricular tracking, the model predicts an average widening of nearly 51%. For those countries which are not in the 'ideal' tracking setup, I compute the percentage change from switching to the ideal curricular tracking in the appendix in table 3.6. The associated reduction in the achievement gap only for these countries is an even stronger 31%. Note that these percentages have to be interpreted as associations and not causally as tracking has the possibility of being endogenous.

Table 3.4 shows the results for the second model outlined in the methodology section. This table is different from the previous as it summarizes the tracking index and adds an interaction between the tracking index and the vocational index.

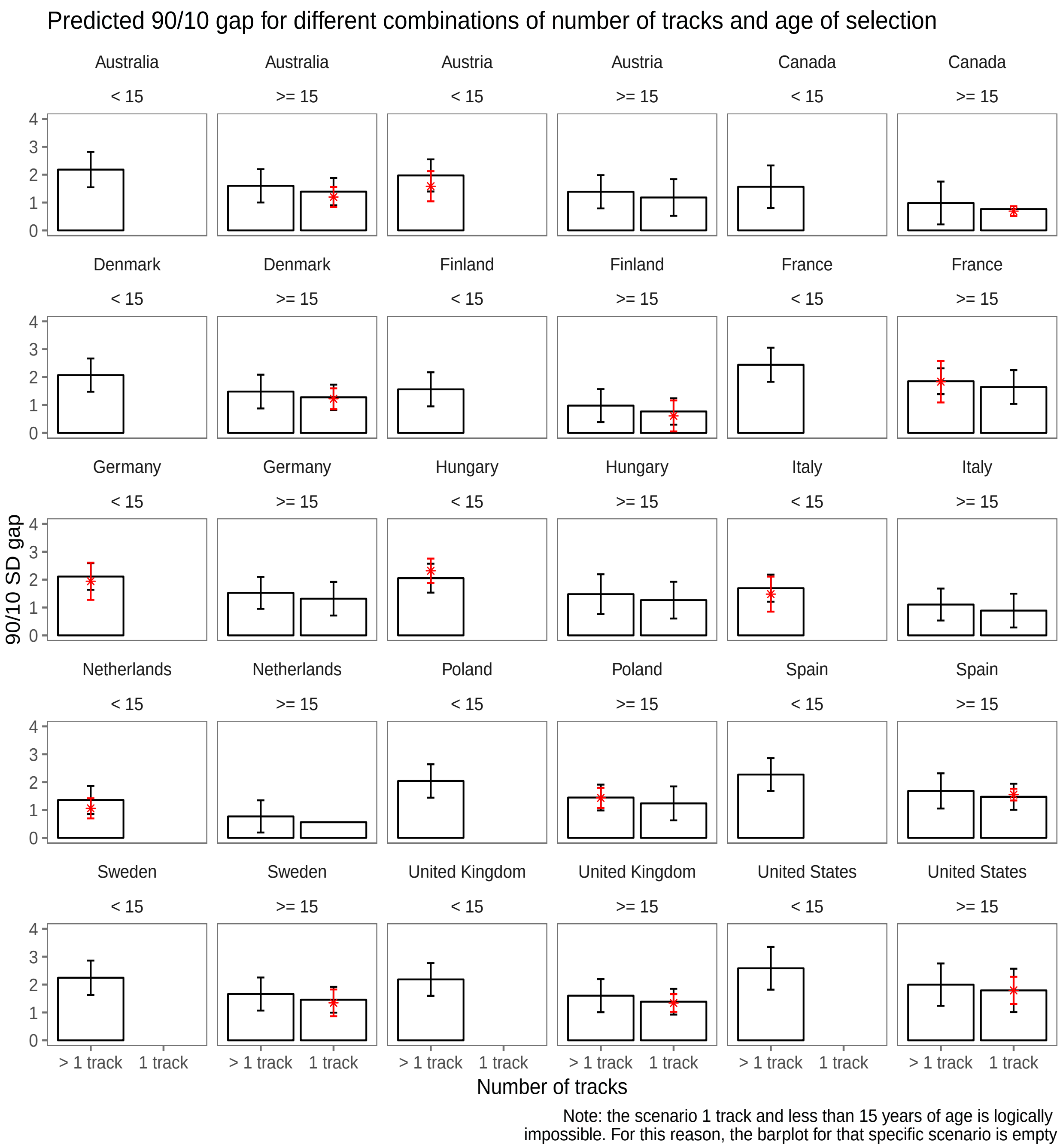


Figure 3.6: Simulation: predicted 90/10 achievement gap for different tracking setups

	Model 1	Model 2	Model 3	Model 4	Model 5
Tracking Index		0.27 (0.16, 0.37)	0.32 (0.22, 0.43)	0.36 (0.25, 0.47)	0.36 (0.25, 0.48)
Vocational Index			-0.19 (-0.3, -0.07)	-0.2 (-0.31, -0.09)	-0.19 (-0.3, -0.08)
Tracking * Vocational Index				-0.12 (-0.21, -0.02)	-0.12 (-0.22, -0.02)
Year 2003					0.2 (0.11, 0.29)
Year 2006					0.1 (0.01, 0.19)
Year 2009					0.11 (0.03, 0.2)
Year 2012					0.08 (-0.01, 0.17)
Year 2015					0.06 (-0.03, 0.15)
Between-group variance:	0.26	0.2	0.18	0.17	0.18
Sample size:	183	183	183	183	183
Number of groups:	32	32	32	32	32
Intercept	1.59 (1.44, 1.75)	1.59 (1.46, 1.72)	1.66 (1.53, 1.79)	1.71 (1.57, 1.85)	1.61 (1.46, 1.76)

Table 3.4: Explaining the 90/10 achievement gap - Tracking and Vocational interaction with intercept varying by country

The results from the previous models are confirmed, where more curricular tracking is associated with an achievement gap of about 0.36 SD wider and more vocational enrollment is associated with a reduction of about -0.2 SD, half the slope of curricular tracking. Could it be that the curricular tracking influence might be offsetting the effect of the vocational index given that it is half the size of the slope? Both these coefficients are actually significantly different from each other, so it is possible. Before we explore it visually, note that the interaction between the two is not either big or small in terms of its uncertainty. However, it shows to be mostly negative [20].

Figure 3.7 plots the interaction for different quantiles of the vocational index. As discussed in the 'Coding and methodology' section, the vocational index is composed by the vocational enrollment (measured separately by the OECD and UNESCO, both in 2006) and the specificity of the vocational track, measured as the percentage of upper secondary vocational education that takes place in a dual system (taken from the OECD in 2007). This indicator has a mean of 0 and a standard deviation of 1, as it is usual when applying a standardization procedure. However, this standardization was applied to all the countries analyzed by Bol and Van de Werfhorst (2013). Given that this analysis focuses on only the 32 countries available from the PISA data, it is more appropriate to include the percentiles based on this subset of countries rather than on the whole sample of countries, which some are not included in the analysis.

Table 3.5 in the appendix shows the tracking and vocational profile of the 32 countries under study. More specifically, it shows the values of the vocational index for all countries. Using this column, I've calculated summary statistics, in which -1.84 is the minimum (so low vocational enrollment and vocational specificity) and the maximum is 1.74 (high vocational enrollment and specificity). I've calculate the values for the 25th, 50th and 75th percentile of this distribution, which are the values -0.2, 0.46 and 0.95. Figure 3.7 plots these three cutoffs in the interaction.

[20] Given that the interaction term might be overly driven by the prior distribution, I ran the *same* model but assigning a uniform prior for the interaction and the coefficient is 0.10 (-0.25, 0.04). As it is evident, the prior distribution just helps to nudge the coefficient slightly because of the low sample size but the results are very much alike.

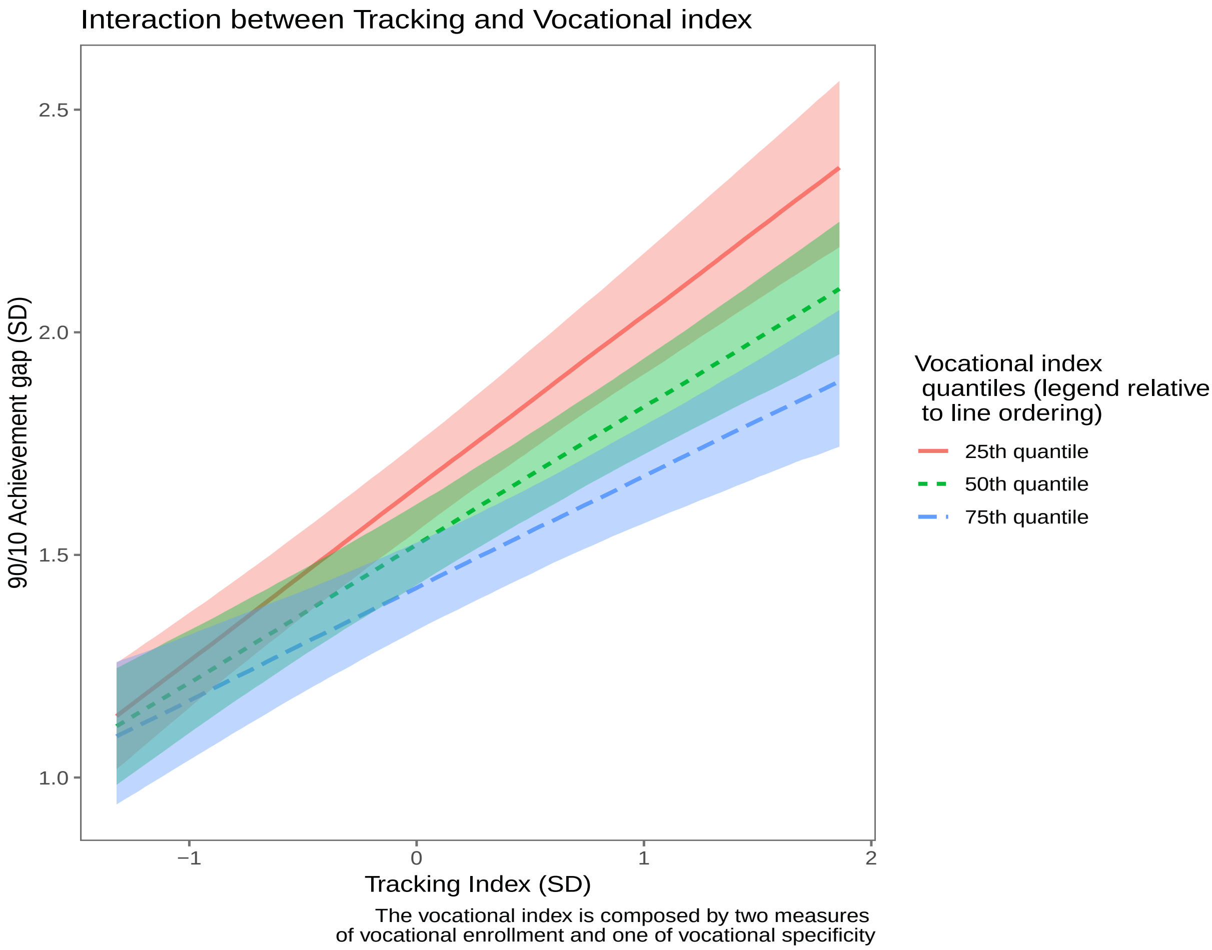


Figure 3.7: Interaction between tracking and vocational values. Legend ordered relative to the lines

The results are interesting. We can see that for lower levels of curricular tracking (left hand side of the X axis) there is a low achievement gap, regardless of whether the country has high or low vocational enrollment (all lines overlap). However, once curricular tracking enters the picture, vocational enrollment can be a strong equalizer (right hand side of the X axis). We see that the bottom 25th percentile of vocational enrollment (top line) has a higher achievement gap by about 0.7 SD than the 75th percentile for vocational enrollment (bottom line). Moreover, the earlier interpretation (that curricular tracking is more important than vocational enrollment for reducing the achievement gap) is evident here. Despite the equalizing power of vocational enrollment, a high level of curricular tracking with the highest level of vocational enrollment still leaves a country with an achievement gap of over 1.9 SD. However, a country with low levels of curricular tracking and *any* level of vocational enrollment is nearly 1.3 SD at best.

This interaction presents high levels of uncertainty considering that the predicted values have wide intervals. This is expected as estimating an interaction term requires more than twice the number of observations to accurately capture a simple main effect (Gelman, 2015). However, even with such a small sample, the interaction shows a clear cut pattern of a relationship.

To test whether these results hold for other samples, I compute the same models as before but also for the 80/20 and 70/30 SES gap. Figure 3.9 plots these coefficients for the three SES gaps in the appendix. All models present virtually the same size in the coefficients which shows that the estimates are robust as they replicate under other gaps.

The third research question looked to understand why some countries differed in their achievement gaps and test whether different indicators of tracking and vocational enrollment seemed to explain this cross-country variability. More concretely, the hypothesis suggested that tracking would explain why some countries showed big differences in achievement gaps while others showed small achievement gaps. The results from table 3.3 and table 3.4 seem to be in line with these claims. Tracking alone seems to explain about 25% of the between country variability while including vocational enrollment raises it nearly 40%. Moreover, Figure 3.7 shows a surprising result: curricular tracking seems to be a possible explanation of the 90/10 achievement gap, however it should not be studied in isolation from other features such as the degree of vocational enrollment.

The exact mechanisms through which the tracking setup is explaining these inequalities is still speculative. It is possible that curricular tracking, although not changing, is currently exacerbating inequality as time passes by. Another explanation, which this chapter finds evidence for, is that the curricular tracking setup interacts closely with the vocational setup of a country. This means that tracking alone seems to increase the achievement gap possibly as students in the higher tracks benefit much more than the ones in the bottom groups. However, this seems to be slightly equalized as students enroll in vocational tracks. It is important to highlight that from a labor market point of view, children enroll in vocational enrollment in the study are not yet in the labor market. This means that finding an association between high vocational enrollment and lower educational gaps is probably reflecting an association between an aggregate pattern of more equality produced by vocational enrollment in the population rather than on these specific students. These explanations are speculative and further research should carry the task of looking for a more concrete explanation and attempting to replicate these findings.

Finally, the last model described in the methodology section looked to model the evolution over time and test tracking and vocational enrollment as possible explanations. Table 3.5 presents the results.

	90th/10th	80th/20th	70th/30th
Track Index	1.65 (0.84, 2.49)	1.12 (0.44, 1.79)	0.99 (0.31, 1.66)
Vocational Index	-1.03 (-1.94, -0.11)	-0.9 (-1.64, -0.15)	-0.86 (-1.6, -0.11)
Intercept	9.36 (8.62, 10.11)	4.58 (3.99, 5.19)	3.88 (3.27, 4.46)
R-squared:	30%	25%	23%
Sample size:	32	32	32

Table 3.5: Linear model regressed on the cumulative achievement gap - Models for three different achievement gaps

We see that a 1 SD increase in the tracking index widens the 90/10 SES gap by 1.66 SD over time, and the vocational index reduces it by 1 SD. These two covariates explain nearly one third of the cumulative gap. As the achievement gap decreases (80/20, 70/30), these coefficients become smaller as expected because the gaps are smaller, but keep their strength and their predictive power. Comparing this model to the previous models, the evidence suggests a similar outlook. That is, tracking and vocational enrollment seem to explain differences in achievement gaps between countries and the preliminary results suggest that it also explains the cumulative yearly gaps indirectly. Moreover, these results are more reassuring of the overall results of the chapter given that this new specification is tackling the evolution of the achievement gap rather than the differences between countries. These results seem to align with the previous findings and make the claim more credible.

Finally, the fourth question and third hypothesis looked to understand not only whether tracking and vocational enrollment explained between-country differences but whether tracking can be also relevant as explaining the evolution of the achievement gap. Although tracking does not change much over time, there are traits that depend on it that do change over time. For example, the quality of instruction for the different tracks might change. This might be more pronounced in countries where there are several tracks and they are very differentiated.

The results from table 3.5 confirm this hypothesis. The results show that tracking seems to explain the cumulative evolution of the achievement gap for the 90/10 gap in Mathematics. The hard numbers point out that tracking and vocational enrollment explain about 30% of the variability on the evolution of the achievement gap. Moreover, the results are replicated for the 80/20 and 70/30 gaps with the results lining up as expected given that as the gaps become broader, the explanatory power decreases.

The results presented in this chapter seem to all converge towards one clear direction: the 90/10 student achievement gaps seem to be very different between countries and institutional features such as curricular tracking and vocational enrollment seem to be plausible explanations for these gaps. Some countries have gaps which are of over 2 standard deviations while others, like Canada, have gaps of 75% less magnitude. However, when all countries are pooled into a common model, tracking and its various features together with vocational enrollment seem to explain about 40% of the variation in the achievement gaps between countries. These results are not perfectly clear because there is a great deal of uncertainty in the estimation procedure. However, the same estimates replicate almost perfectly with smaller achievement gaps such as the 80/20 and 70/30 gaps.

Finally, as an indirect measure, the results also show that the tracking structure and vocational index seem to indirectly explain about 30% of the cumulative achievement gap in 15 years for the 32 countries under study. This test is not a magic bullet because it is very difficult to calculate a cumulative achievement gap if gaps increase/decrease every year with a constant positive or negative slope. However, for countries which constantly increase or decrease, then this is a valid model. As the evidence in the visualization of the achievement gaps suggests, there are a handful of countries, if not the majority, which show either increasing or decreasing patterns over time. All in all, the take away from the combination of all the previous results is that despite the uncertainty in all previous models, they all point in the same direction.

3.6 Limitations

The results presented here seem to converge to similar results and prove to be robust throughout different specifications. Despite this, there are several important limitations in the design of the study.

The main limitation of the study is the fact that once the achievement gap is summarized for each country-year combination, the resulting sample size is small and contains an accumulated degree of measurement error as calculating the achievement gap carries forward a number of errors in measurement. I adopt one of the most widely chosen strategies: bayesian estimation. This allows to counteract the low statistical power and measurement error by providing likely prior distributions. The prior distributions used here were chosen among experimentation, theoretical considerations and previous research. But of course, prior distributions carry some degree of subjectivity. This is a limitation that is important. Further research should attempt to study the findings presented here under a different scenario or under different priors (already tried in the exploratory attempts before the chapter was written down) and present the results.

The paper has evaluated the trends in achievement gap using visual representations of the trends as well as showing the point estimates in differences between the first and last year of the trends. Having said that, there are still concerns on whether these trends are reflecting the true cross-country trends or they are also reflecting sampling variation. To understand these short comings, future research should model the trends statistically and test for whether they are substantially different in their magnitudes. I have avoided this type of analysis until this point because there are only 6 time points per country which is not enough to apply any proper and reliably time-series analysis. The paper attempts to begin to understand the phenomena by providing descriptives of the trends but it should be acknowledged that future work should prioritize modelling the trends statistically.

When calculating achievement gaps I use the cognitive scores provided by PISA and follow estimations which have been replicated and validated in other studies. However, there have been major criticisms as to what exactly is the PISA test measuring. Perhaps these are tests that confound true ability with teaching-to-the-test. The previous phenomena is not quite clear for PISA given that the test is not announced with enough time for teachers to prepare. However, it is true that these types of tests have received extensive criticism [21] and further research should keep that in mind. These tests contain measurement error and are not bullet-proof in terms of the underlying latent variable being measured.

3.7 Conclusion

The literature on achievement inequality has recently started studying the evolution of the achievement gap and it has uncovered a wide variety of results. The main finding is that the gap of the U.S. has been widening over time by about 40-50% from 1970 until the 2000. Others find that this recent trend has reversed, at least for an age-specific group since the 2000's. From an international perspective, recent studies have shown that there is wide variability in this trend, with some countries showing decreases while others increases. This chapter looks to investigate the magnitude of these gaps across selected countries, and attempts to explain why they might be happening.

The chapter looks at the evolution of the achievement gap and finds starking variation for the 15-year old achievement gap across countries. The U.S. has been closing its 90/10 gap by about 50% with a wide interval going down to nearly 35%. This reduction is also present in Germany, shockingly a country with an institutionalized curricular tracking system. Other countries such as France and Austria experience the contrary with persistent increases in the achievement gap. Moreover, once I disaggregate these patterns into a more fine-grained analysis, I discover that the dynamics are different. The gap of the U.S. is shrinking because the top SES group is coming down at a faster rate than the bottom group is catching up. Germany, which has a similar reduction, portrays a different explanation. Both the top and bottom groups are closing together at a similar rate. The results from figure 3.5 once aggregated to all countries show a general pattern across all 32 countries: either the bottom catches up and the top comes down or viceversa. There seems to be a trade off between these forces. This pattern is interesting and there is still is not an explanation as to why it happens. I leave up to future research to test the mechanisms and explanations through which this trade off occurs.

[21] For example, see the article by the guardian `https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics` where over 100 academics signed a petition for the OECD to alter the logistics of PISA

I propose to study tracking and vocational enrollment as possible explanations to the evolution of the achievement gap and find that the degree of curricular tracking of a country seems to explain this phenomena rather strongly. Tracking alone explains nearly 30% of between-country variability in achievement gaps. However, the results suggest that it should not be studied without factoring in the degree of vocational enrollment. I find that if a country has a low level of curricular tracking, vocational enrollment will not help decrease the achievement gap. Conversely, once curricular tracking enters the picture, vocational enrollment can help ease the burden of inequality and reduce the achievement gap significantly. In a similar vein, I use this model to perform a simulation for each country: if every country switched to little or no curricular tracking, what would be the resulting reduction in the 90/10 achievement gap? I find that if all countries switched, the average gap would shrink by about 10%, with outlier countries like the Netherlands and Hungary experiencing reductions of 48% and 46% respectively. I contrast the former simulation with another counterfactual scenario: what would be the average gap increase if all countries switched to a system where curricular tracking predominates? I record an average widening of 51%, something that suggests that increasing curricular tracking is associated with a worsening of the achievement gap by a big margin.

I conclude by showing that curricular tracking and vocational enrollment are playing an important role in explaining the evolution of the achievement gap. I find that these results are suggestive that institutional features, more specifically, stratified curriculums, might be playing a sizable role in the promotion of inequality. After having ran a battery of tests and different empirical strategies, the results are consistent and point out that we should pay attention to the source of the achievement gaps and study whether particular SES groups are contributing more to the gap than others.

Future research should attempt to replicate these results under further empirical tests. If other studies manage to replicate and corroborate these results, then policymakers should begin to consider vocational enrollment not only as a 'labor market' solution to youth unemployment but also as a compensatory mechanism for countries which have highly tracked curriculums.

Bibliography

Bauchmuller, Robert, Mette Gortz, and Astrid Wurtz Rasmussen (2014). “Long-run benefits from universal high-quality preschooling.” In: *Early Childhood Research Quarterly* 29.4, pp. 457–470.

Beaton, Albert E et al. (1996). *Mathematics Achievement in the Middle School Years. IEA's Third International Mathematics and Science Study (TIMSS).* ERIC.

Bernardi, Fabrizio and Gabriele Ballarino (2016). *Education, occupation and social origin: a comparative analysis of the transmission of socio-economic inequalities*. Edward Elgar Publishing.

Bjorklund, Anders and Markus Jantti (2009). “Intergenerational income mobility and the role of family background.” In: *Oxford Handbook of Economic Inequality, Oxford University Press, Oxford*, pp. 491–521.

Bol, Thijs and Herman G Van de Werfhorst (2013). “Educational systems and the trade-off between labor market allocation and equality of educational opportunity.” In: *Comparative Education Review* 57.2, pp. 285–308.

Bradbury, Bruce, Miles Corak, Jane Waldfogel, and Elizabeth Washbrook (2015). *Too many children left behind: The US achievement gap in comparative perspective*. Russell Sage Foundation.

Breen, Richard and John Goldthorpe (1997). “Explaining educational differentials: Towards a formal rational action theory.” In: *Rationality and society* 9.3, pp. 275–305.

Breen, Richard and Jan O Jonsson (2007). “Explaining Change in Social Fluidity: Educational Equalization and Educational Expansion in Twentieth-Century Sweden 1.” In: *American Journal of Sociology* 112.6, pp. 1775–1810.

Byun, Soo-yong and Kyung-keun Kim (2010). “Educational inequality in South Korea: The widening socioeconomic gap in student achievement.” In: *Globalization, changing* *demographics, and educational challenges in East Asia.* Emerald Group Publishing Limited, pp. 155–182.

Campbell, Frances A, Craig T Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson (2002). "Early childhood education: Young adult outcomes from the Abecedarian Project." In: *Applied developmental science* 6.1, pp. 42–57.

Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang (2016). *The Fading American Dream: Trends in Absolute Income Mobility Since 1940.* Tech. rep. National Bureau of Economic Research.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez (2014). *Where is the land of opportunity? The geography of intergenerational mobility in the United States.* Tech. rep. National Bureau of Economic Research.

Chmielewski, Anna K (2016). "Changes in Socioeconomic Achievement Gaps in International Comparison, 1964-2012." In: *2016 APPAM International Conference.* Appam.

Chmielewski, Anna K and Sean F Reardon (2016). "Patterns of Cross-National Variation in the Association Between Income and Academic Achievement." In: *AERA Open* 2.3, p. 2332858416649593.

Coleman, James S et al. (1966). "Equality of educational opportunity." In:

Cunha, Flavio, James Heckman, Lance Lochner, and Dimitriy V Masterov (2006). "Interpreting the evidence on life cycle skill formation." In: *Handbook of the Economics of Education* 1, pp. 697–812.

Duncan, Greg J and Katherine Magnuson (2011). "The nature and impact of early achievement skills, attention skills, and behavior problems." In: *Whither opportunity*, pp. 47–70.

Dupriez, Vincent and Xavier Dumay (2006). "Inequalities in school systems: effect of school structure or of society structure?" In: *Comparative education* 42.02, pp. 243–260.

Duru-Bellat, Marie and Bruno Suchaut (2005). "Organisation and Context, Efficiency and Equity of Educational Systems: what PISA tells us." In: *European Educational Research Journal* 4.3, pp. 181–194.

Duyme, Michel, Annick-Camille Dumaret, and Stanislaw Tomkiewicz (1999). "How can we boost IQs of "dull children"?: A late adoption study." In: *Proceedings of the National Academy of Sciences* 96.15, pp. 8790–8794.

Economic Cooperation, Organization for and Development (2005). *Education at a Glance.* Tech. rep. OECD, Paris.

— (2006a). *Education at a Glance*. Tech. rep. OECD, Paris.

— (2006b). *PISA 2006: Volume 2 Data*. Tech. rep. OECD, Paris.

— (2007). *Education at a Glance,* tech. rep. OECD, Paris.

Elango, Sneha, Jorge Luis Garcia, James Heckman, and Andres Hojman (Nov. 2015). *Early Childhood Education*. Working Paper 21766. National Bureau of Economic Research. DOI: 10.3386/w21766. URL: http://www.nber.org/papers/w21766.

Erikson, Robert, John Goldthorpe, and Lucienne Portocarero (1979). "Intergenerational class mobility in three Western European societies: England, France and Sweden." In: *The British Journal of Sociology* 30.4, pp. 415–441.

Esping-Andersen, Gøsta, Irwin Garfinkel, Wen-Jui Han, Katherine Magnuson, Sander Wagner, and Jane Waldfogel (2012). "Child care and school performance in Denmark and the United States." In: *Children and Youth Services Review* 34.3, pp. 576–589.

Fryer J, Roland G and Steven D Levitt (2004). "Understanding the black-white test score gap in the first two years of school." In: *Review of Economics and Statistics* 86.2, pp. 447–464.

Gamoran, Adam (2001). "American schooling and educational inequality: A forecast for the 21st century." In: *Sociology of education*, pp. 135–153.

Ganzeboom, Harry BG (2010). "A new International Socio-Economic Index (ISEI) of occupational status for the International Standard Classification of Occupation 2008 (ISCO-08) constructed with data from the ISSP 2002-2007." In: *Annual Conference of International Social Survey Programme, Lisbon*. Vol. 1.

Ganzeboom, Harry BG and Donald J Treiman (1996). "Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations." In: *Social science research* 25.3, pp. 201–239.

Gelman, Andrew (2015). *The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective.*

Hanushek, Eric et al. (2006). "Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries." In: *The Economic Journal* 116.510, pp. C63–C76.

Heckman, James (2006). "Skill formation and the economics of investing in disadvantaged children." In: *Science* 312.5782, pp. 1900–1902.

Jaeger, Mads Meier and Anders Holm (2007). "Does parents' economic, cultural, and social capital explain the social class effect on educational attainment in the Scandinavian mobility regime?" In: *Social Science Research* 36.2, pp. 719–744.

Jakubowski, Maciej (2010). "Institutional tracking and achievement growth: Exploring difference-in-differences approach to PIRLS, TIMSS, and PISA data." In: *Quality and inequality of education*. Springer, pp. 41–81.

Jakubowski, Maciej, Emilio Ernesto Porta, Jerzy Wisniewski, and Harry Anthony Patrinos (2010). *The impact of the 1999 education reform in Poland*. The World Bank.

Kautz, Tim, James Heckman, Ron Diris, Bas ter Weel, and Lex Borghans (Dec. 2014). *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*. Working Paper 20749. National Bureau of Economic Research. DOI: 10.3386/w20749. URL: http://www.nber.org/papers/w20749.

Magnuson, Katherine and Jane Waldfogel (2008). *Steady gains and stalled progress: Inequality and the Black-White test score gap*. Russell Sage Foundation.

Marks, Gary N, John Cresswell, and John Ainley (2006). "Explaining socioeconomic inequalities in student achievement: The role of home and school factors." In: *Educational research and Evaluation* 12.02, pp. 105–128.

Micklewright, J and SV Schnepf (2007). *Inequality and Poverty Re-Examined*. Oxford Univ. Press. Chap. Inequality of learning in industrialized countries.

OECD (2002). *Education At A Glance: OECD Indicators 2002*. OECD Publishing. ISBN: 9264198903.

— (Mar. 2012). *PISA 2009 Technical Report*. OECD Publishing.

— (Dec. 2016). *PISA 2015 Results (Volume I)*. OECD Publishing.

OECD. (2009). *PISA data analysis manual: SAS 2nd edition*.

Piketty, Thomas and Emmanuel Saez (2003). "Income inequality in the United States, 1913–1998." In: *The Quarterly journal of economics* 118.1, pp. 1–41.

Reardon, Sean F (2011). "The widening academic achievement gap between the rich and the poor: New evidence and possible explanations." In: *Whither opportunity*, pp. 91–116.

Reardon, Sean F and Ximena A Portilla (2015). "Recent trends in socioeconomic and racial school readiness gaps at kindergarten entry." In: *Stanford, CA: Center for Education Policy Analysis. Retrieved December* 20, p. 2015.

Reardon, Sean F, Joseph P Robinson, and Ericka S Weathers (2008). “Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps.” In: *Handbook of research in education finance and policy*, pp. 497–516.

Saw, Guan Kung (2016). “Patterns and trends in achievement gaps in Malaysian Secondary Schools (1999–2011): gender, ethnicity, and socioeconomic status.” In: *Educational Research for Policy and Practice* 15.1, pp. 41–54.

Shavit, Yossi and Hans-Peter Blossfeld (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Social Inequality Series.* ERIC.

Van de Werfhorst, Herman G and Jonathan JB Mijs (2010b). “Achievement inequality and the institutional structure of educational systems: A comparative perspective.” In: *Annual review of sociology* 36, pp. 407–428.

Waldfogel, Jane (2006). “What do children need?” In: *Juncture* 13.1, pp. 26–34.

Waldinger, Fabian (2006). “Does tracking affect the importance of family background on students’ test scores?” In: *Unpublished manuscript, LSE.*

Wu, Margaret (2005). “The role of plausible values in large-scale surveys.” In: *Studies in Educational Evaluation* 31.2-3, pp. 114–128.

3.8 Appendix

	Year 2000	Year 2003	Year 2006	Year 2009	Year 2012	Year 2015
Intercept	-0.4 (0.1)	-0.34 (0.08)	-0.42 (0.08)	-0.47 (0.07)	-0.41 (0.07)	-0.49 (0.07)
SES Dummy	1.23 (0.05)	1.33 (0.04)	1.29 (0.04)	1.34 (0.04)	1.23 (0.04)	1.25 (0.04)
Standard Dev. Intercept	0.61	0.49	0.61	0.56	0.53	0.59
Standard Dev. SES Dummy	0.29	0.25	0.25	0.28	0.31	0.28
Correlation Intercept * SES Dummy	-0.37	-0.19	-0.39	-0.05	0.01	0.08

Table 3.1: Random slope/intercept models for estimating mathematics SES gaps between top and bottom performers across all years

	Year 2000	Year 2003	Year 2006	Year 2009	Year 2012	Year 2015
Intercept	-0.49 (0.09)	-0.44 (0.07)	-0.4 (0.08)	-0.55 (0.06)	-0.48 (0.07)	-0.6 (0.07)
SES Dummy	1.3 (0.05)	1.37 (0.05)	1.23 (0.04)	1.34 (0.04)	1.21 (0.05)	1.27 (0.04)
Standard Dev. Intercept	0.55	0.42	0.55	0.49	0.52	0.57
Standard Dev. SES Dummy	0.31	0.28	0.27	0.28	0.33	0.31
Correlation Intercept * SES Dummy	-0.36	-0.25	-0.42	-0.23	-0.22	-0.15

Table 3.2: Random slope/intercept models for estimating literacy SES gaps between top and bottom performers across all years

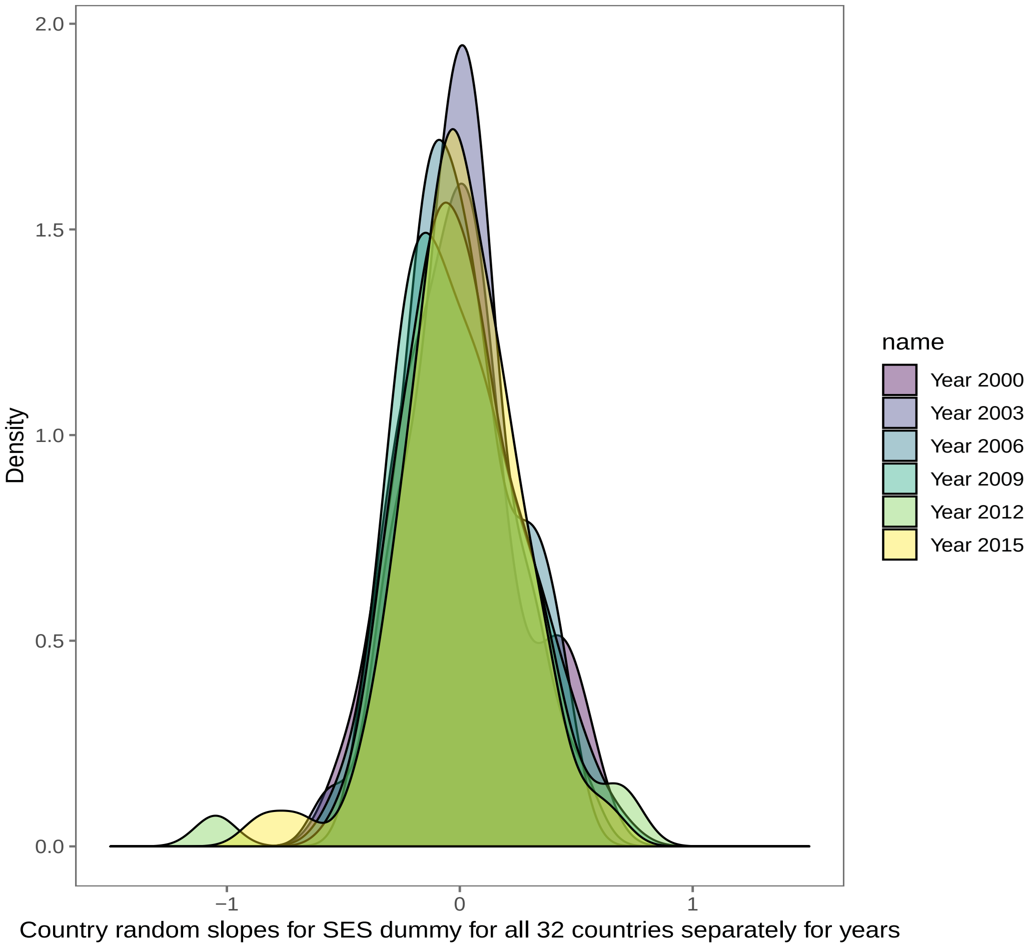


Figure 3.1: Mathematics country random slopes for SES dummy for all 32 countries separately for all years

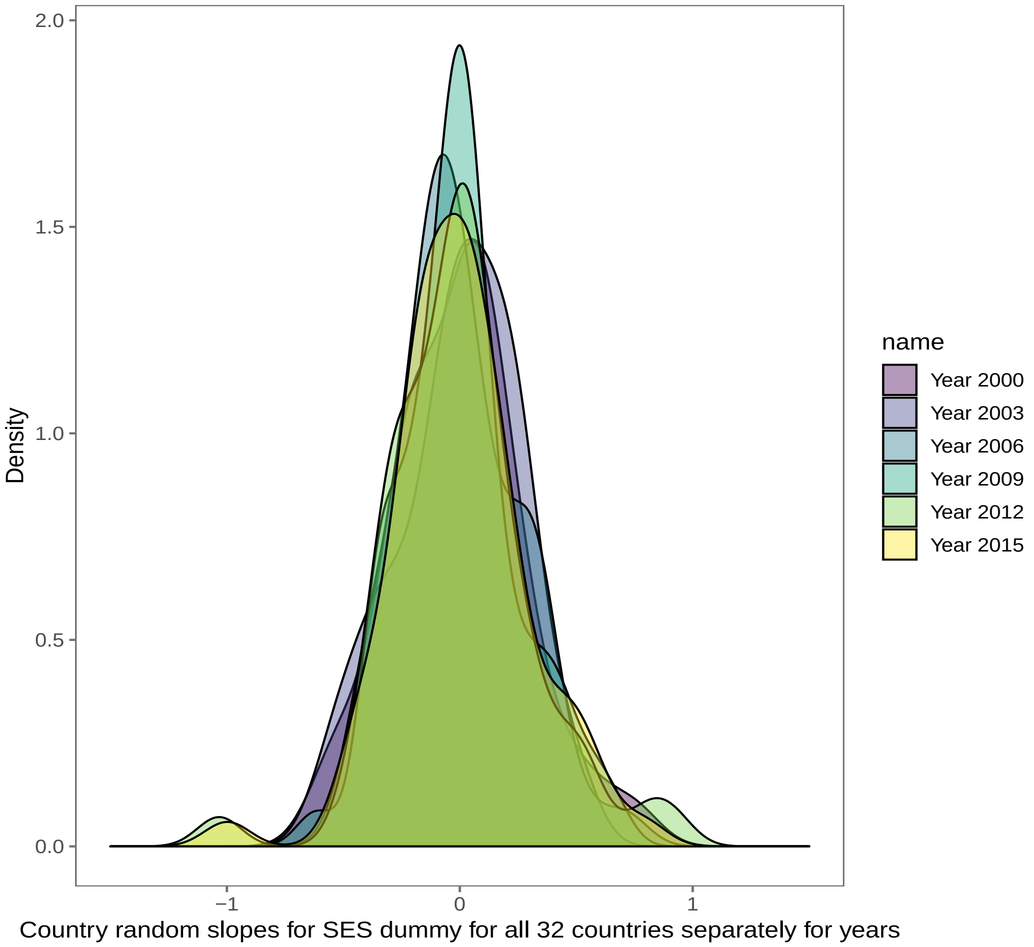


Figure 3.2: Literacy country random slopes for SES dummy for all 32 countries separately for all years

	2000				2015			
	Low SES		High SES		Low SES		High SES	
	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage
Gender								
- Female	6314	53.0%	5597	49.0%	25394	52.0%	25980	50.0%
- Male	5605	47.0%	5817	51.0%	23447	48.0%	26087	50.0%
- NA	21	0.0%	16	0.0%	0	0.0%	0	0.0%
Highest edu in HH								
- No schooling	784	7.0%	0	0.0%	4692	10.0%	0	0.0%
- Primary education	3480	29.0%	0	0.0%	9555	20.0%	2	0.0%
- Lower secondary education	3184	27.0%	5	0.0%	12447	25.0%	14	0.0%
- Upper secondary education	946	8.0%	83	1.0%	4844	10.0%	58	0.0%
- Post-secondary non-tertiary education	1427	12.0%	102	1.0%	10264	21.0%	681	1.0%
- Bachelors or above	25	0.0%	10949	96.0%	521	1.0%	48210	93.0%
- NA	2094	18.0%	291	3.0%	6518	13.0%	3102	6.0%
Father born in country of test								
- Yes	9163	77.0%	9902	87.0%	36716	75.0%	44074	85.0%
- No	1920	16.0%	1205	11.0%	9992	20.0%	6834	13.0%
- NA	857	7.0%	323	3.0%	2133	4.0%	1159	2.0%
Number of books in the HH								
- 0-10	3934	33.0%	221	2.0%	21407	44.0%	1738	3.0%
- 11-100	5735	48.0%	1851	16.0%	22400	46.0%	14042	27.0%
- 101-250	1054	9.0%	2384	21.0%	2568	5.0%	10955	21.0%
- 251 or more	680	6.0%	6762	59.0%	1487	3.0%	25111	48.0%
- NA	537	4.0%	212	2.0%	979	2.0%	221	0.0%

Table 3.3: Descriptives of selected variable for 2000 and 2015 for all countries pooled

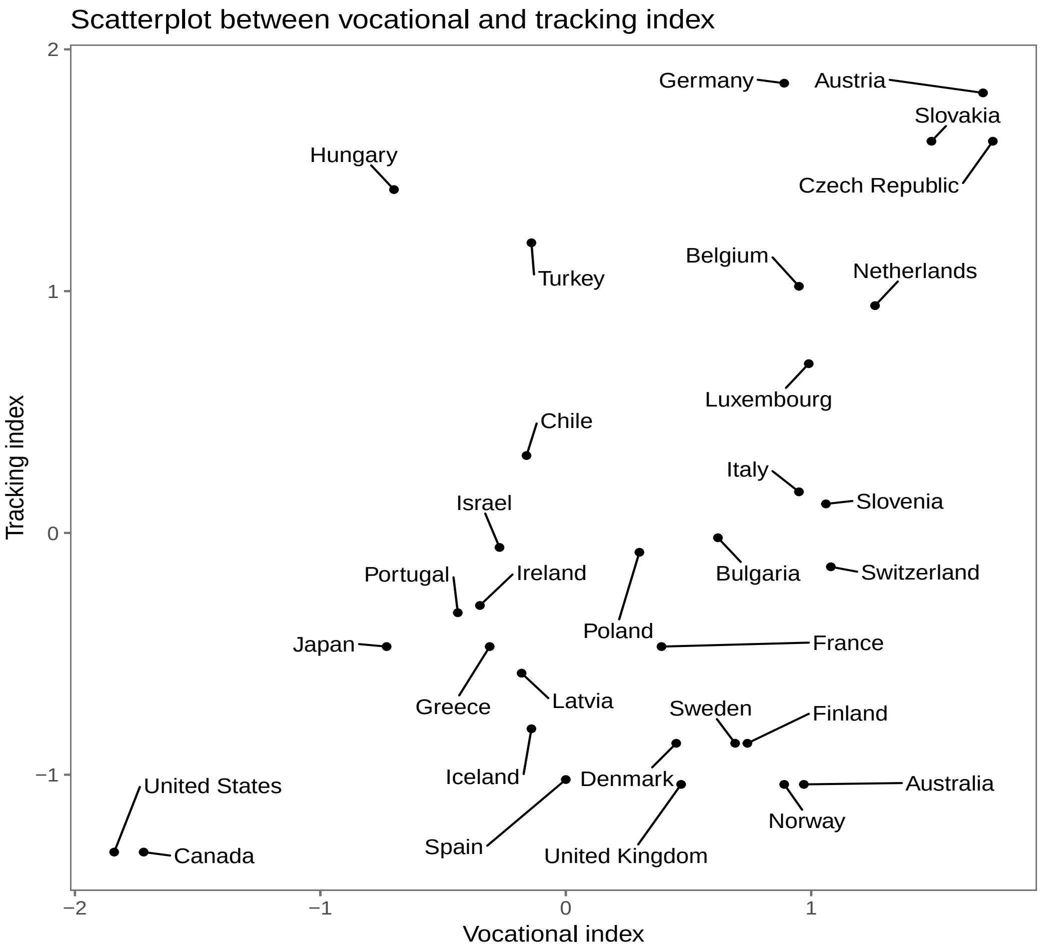


Figure 3.4: Scatterplot between tracking and vocational index with country names

Country	Number of tracks	Age of selection	Inequality	Predicted	% reduction
Austria	More than one track	Less than 15	1.58	1.18	26%
France	More than one track	15 or more	1.84	1.65	11%
Germany	More than one track	Less than 15	1.94	1.32	33%
Hungary	More than one track	Less than 15	2.32	1.26	46%
Italy	More than one track	Less than 15	1.48	0.89	40%
Netherlands	More than one track	Less than 15	1.06	0.56	48%
Poland	More than one track	15 or more	1.43	1.24	14%

Table 3.6: Simulation: reduction of achievement gap if countries switched to 'ideal' tracking

		Low SES			High SES			
Year	Countries	N	Avg score	S.E	N	Avg score	S.E	SES gap
2000	Australia	294	0.22	0.07	270	1.22	0.09	1.00
2000	Austria	294	-0.04	0.11	235	1.14	0.15	1.19
2000	Belgium	360	-0.19	0.09	375	1.44	0.12	1.64
2000	Bulgaria	263	-1.02	0.10	231	1.67	0.14	2.69
2000	Canada	1970	0.30	0.05	1357	0.90	0.07	0.60
2000	Chile	248	-1.23	0.07	288	1.37	0.09	2.60
2000	Czech Republic	306	-0.17	0.09	323	1.34	0.12	1.51
2000	Denmark	230	-0.01	0.13	228	1.16	0.17	1.18
2000	Finland	261	0.42	0.12	267	0.89	0.15	0.47
2000	France	253	0.03	0.04	258	1.20	0.05	1.17
2000	Germany	250	-0.64	0.04	303	1.80	0.05	2.44
2000	Greece	279	-0.56	0.08	250	1.31	0.11	1.87
2000	Hungary	291	-0.46	0.09	254	1.65	0.12	2.11
2000	Iceland	185	-0.04	0.35	186	1.14	0.27	1.18
2000	Ireland	207	0.05	0.12	213	1.09	0.16	1.04
2000	Israel	206	-1.03	0.12	241	1.46	0.15	2.49
2000	Italy	278	-0.31	0.05	278	0.80	0.06	1.12
2000	Latvia	182	-0.41	0.16	217	0.99	0.19	1.39
2000	Luxembourg	179	-0.56	0.34	190	1.24	0.26	1.80
2000	Netherlands	129	0.46	0.08	158	1.24	0.11	0.78
2000	Norway	227	0.04	0.13	234	0.98	0.16	0.94
2000	Poland	209	-0.31	0.04	166	1.25	0.06	1.56
2000	Portugal	243	-0.50	0.10	264	1.32	0.13	1.82
2000	Spain	318	-0.28	0.05	340	1.09	0.07	1.37
2000	Sweden	241	0.12	0.10	246	1.06	0.13	0.94
2000	Switzerland	323	0.01	0.11	291	1.34	0.14	1.33
2000	United Kingdom	531	0.15	0.04	415	1.32	0.06	1.17
2000	United States	222	-0.43	0.02	164	1.74	0.03	2.16
2015	Australia	1694	-0.15	0.04	1235	1.28	0.05	1.43
2015	Austria	676	-0.30	0.07	705	1.53	0.10	1.83
2015	Belgium	860	-0.20	0.06	983	1.63	0.08	1.83
2015	Bulgaria	544	-0.91	0.08	588	1.42	0.11	2.34
2015	Canada	2215	0.20	0.03	1863	0.99	0.05	0.79
2015	Chile	536	-1.06	0.04	1453	1.58	0.06	2.64
2015	Czech Republic	604	-0.34	0.07	821	1.61	0.09	1.95
2015	Denmark	1013	0.09	0.08	595	1.08	0.10	0.99
2015	Finland	575	0.07	0.08	584	1.16	0.11	1.09
2015	France	570	-0.37	0.02	615	1.66	0.03	2.03
2015	Germany	545	-0.09	0.02	582	1.45	0.03	1.54
2015	Greece	481	-0.60	0.06	606	1.21	0.08	1.81
2015	Hungary	466	-0.72	0.07	589	1.79	0.09	2.51
2015	Iceland	322	-0.30	0.23	334	1.17	0.23	1.47
2015	Ireland	574	-0.01	0.08	561	1.21	0.10	1.22
2015	Israel	639	-0.59	0.06	681	1.16	0.08	1.76
2015	Italy	947	-0.20	0.03	1043	1.23	0.04	1.43
2015	Japan	650	0.30	0.02	659	1.21	0.03	0.91
2015	Latvia	419	-0.29	0.14	470	1.09	0.17	1.37
2015	Luxembourg	517	-0.22	0.20	521	1.37	0.22	1.59
2015	Netherlands	521	0.13	0.04	525	1.24	0.06	1.11
2015	Norway	531	-0.00	0.08	530	1.04	0.11	1.04
2015	Poland	446	0.10	0.03	448	1.24	0.05	1.14
2015	Portugal	948	-0.14	0.06	571	1.36	0.08	1.50
2015	Slovakia	579	-0.71	0.08	630	1.63	0.11	2.34
2015	Slovenia	747	0.05	0.13	491	1.23	0.17	1.18
2015	Spain	608	-0.25	0.03	703	1.24	0.04	1.49
2015	Sweden	527	-0.20	0.06	542	1.32	0.09	1.52
2015	Switzerland	575	0.04	0.07	573	1.43	0.09	1.38
2015	Turkey	593	-0.80	0.02	559	1.12	0.03	1.92
2015	United Kingdom	1387	-0.03	0.03	1388	1.15	0.04	1.18
2015	United States	585	-0.39	0.01	544	1.22	0.01	1.62

Table 3.4: SES sample size, average score in Mathematics and ISCED composition for all countries

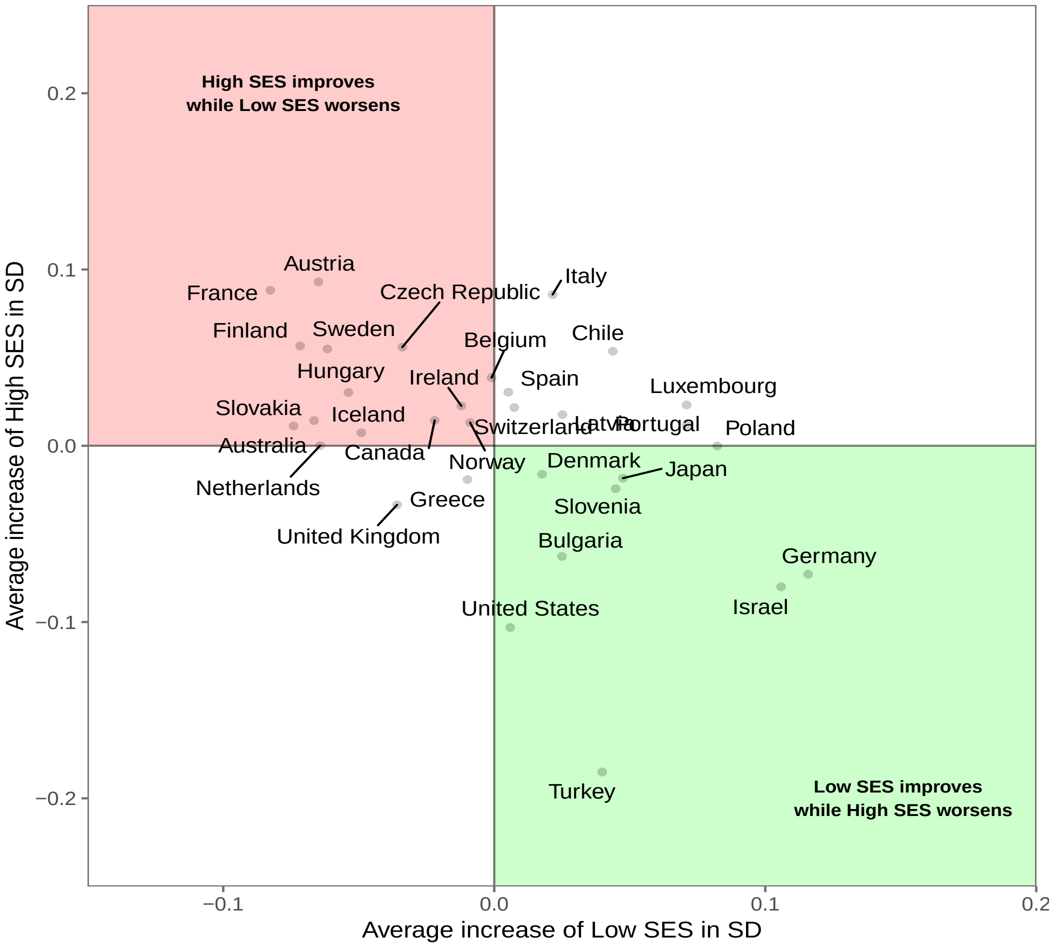


Figure 3.3: Rate at which top/bottom SES groups are changing for the mathematics 90/10 SES gap. All country names highlighted.

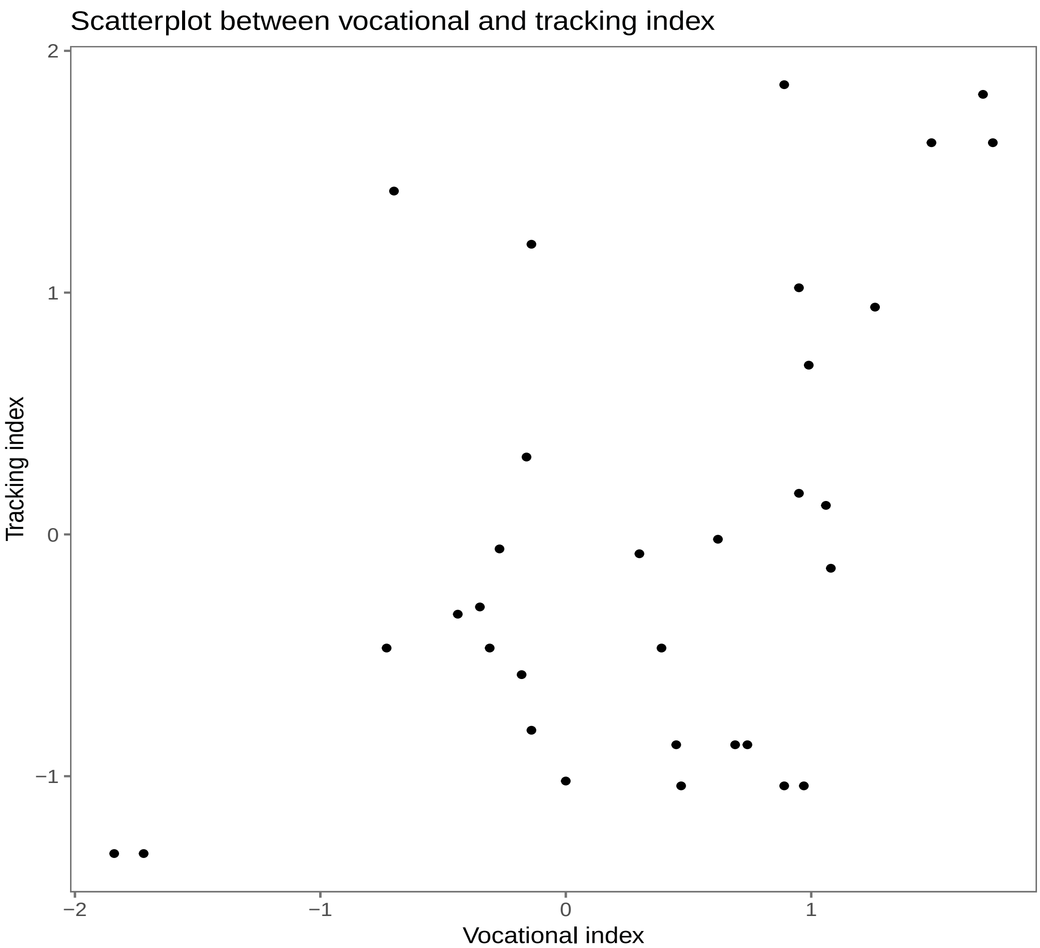


Figure 3.5: Scatterplot between tracking and vocational index

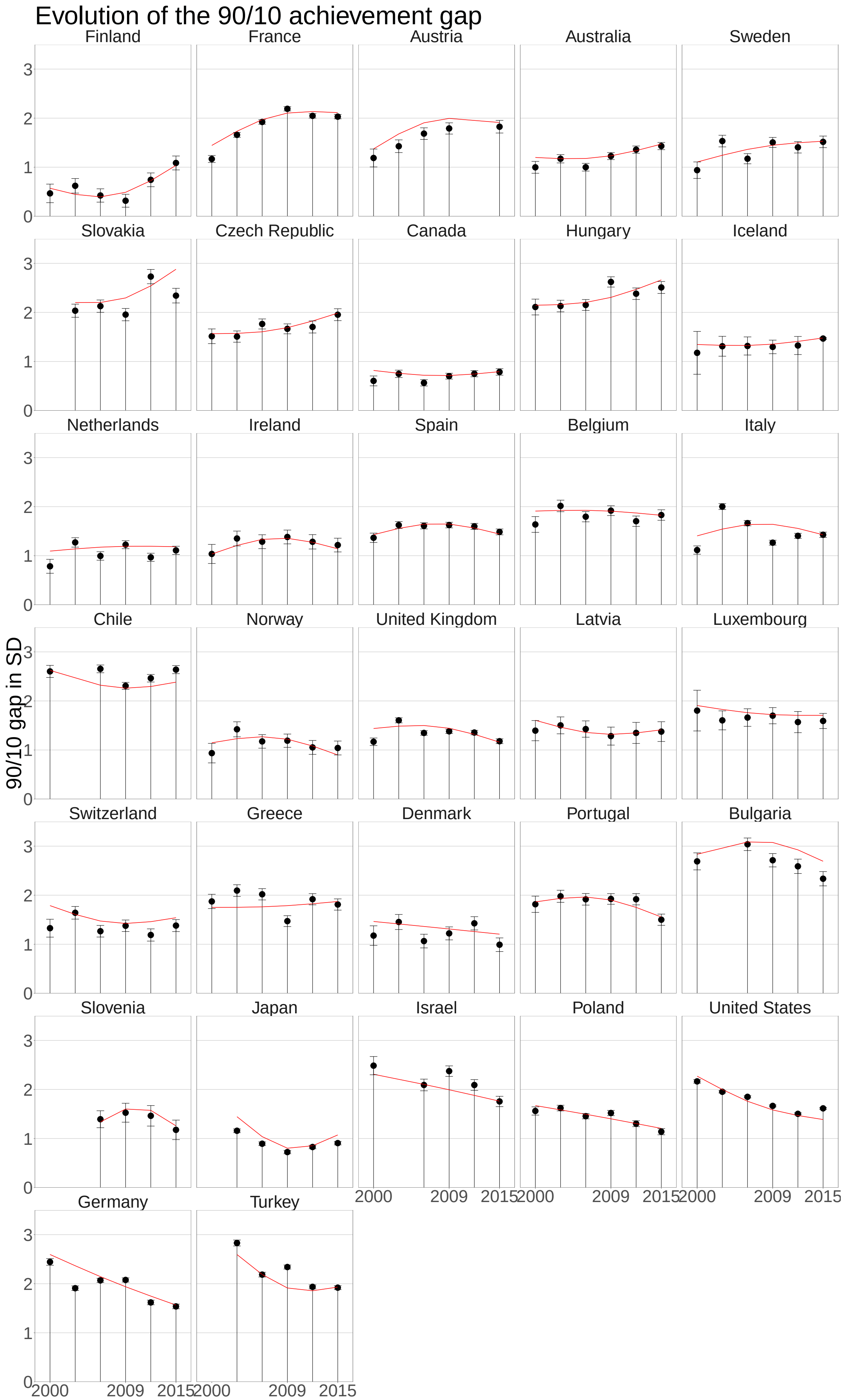


Figure 3.6: 90/10 achievement gaps in Mathematics expressed as standard deviations for all countries between 2000 and 2015

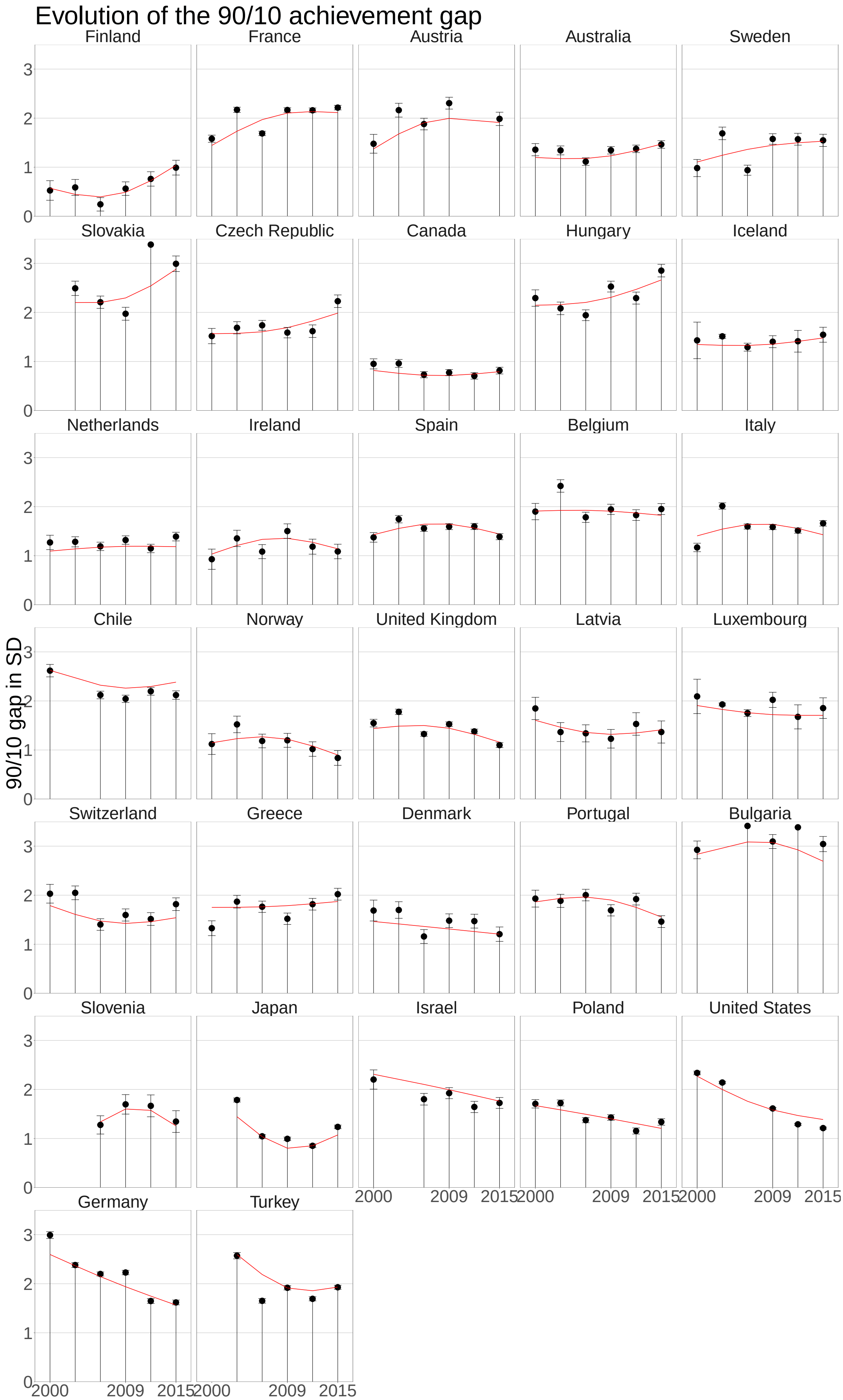


Figure 3.7: 90/10 achievement gaps in literacy expressed as standard deviations for selected countries between 2000 and 2015

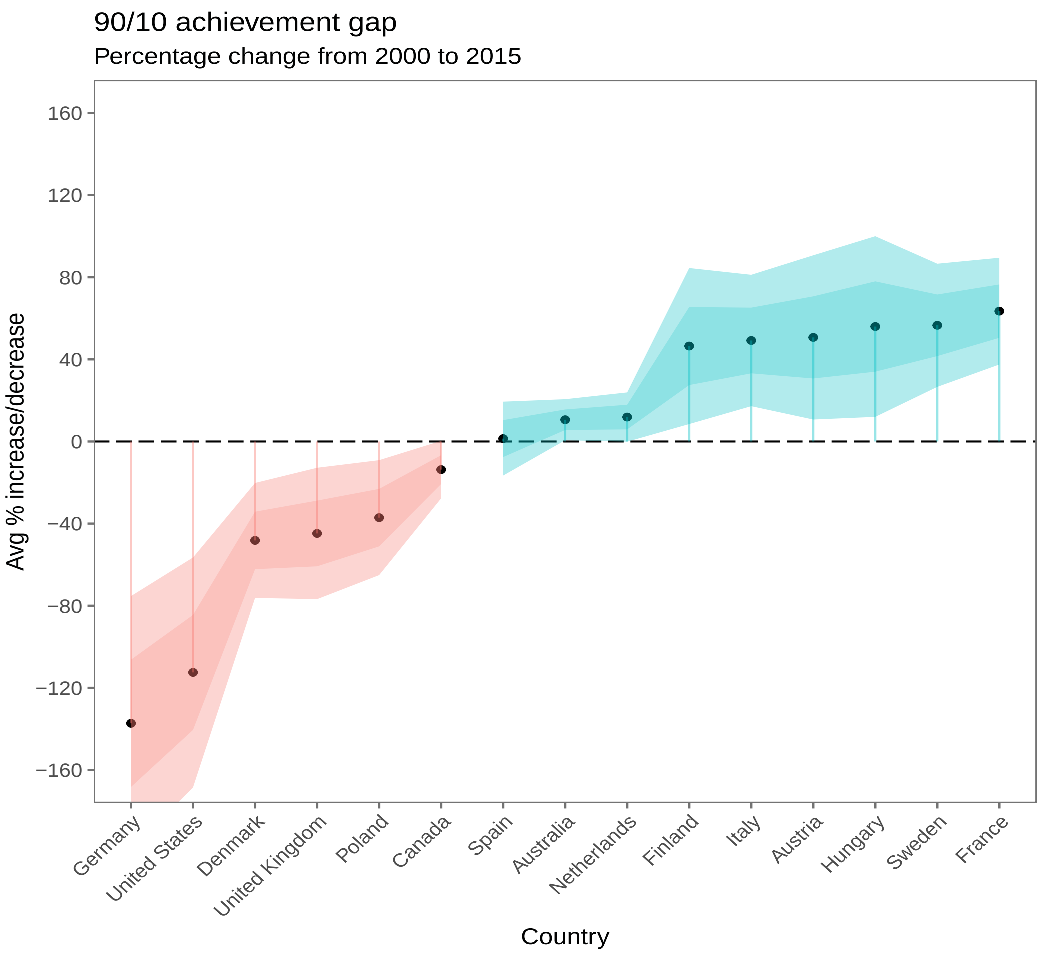


Figure 3.8: Percentage change in literacy between years 2015 and 2000 for selected countries. Red regions represent decreases while blue regions represent increases in the achievement gap

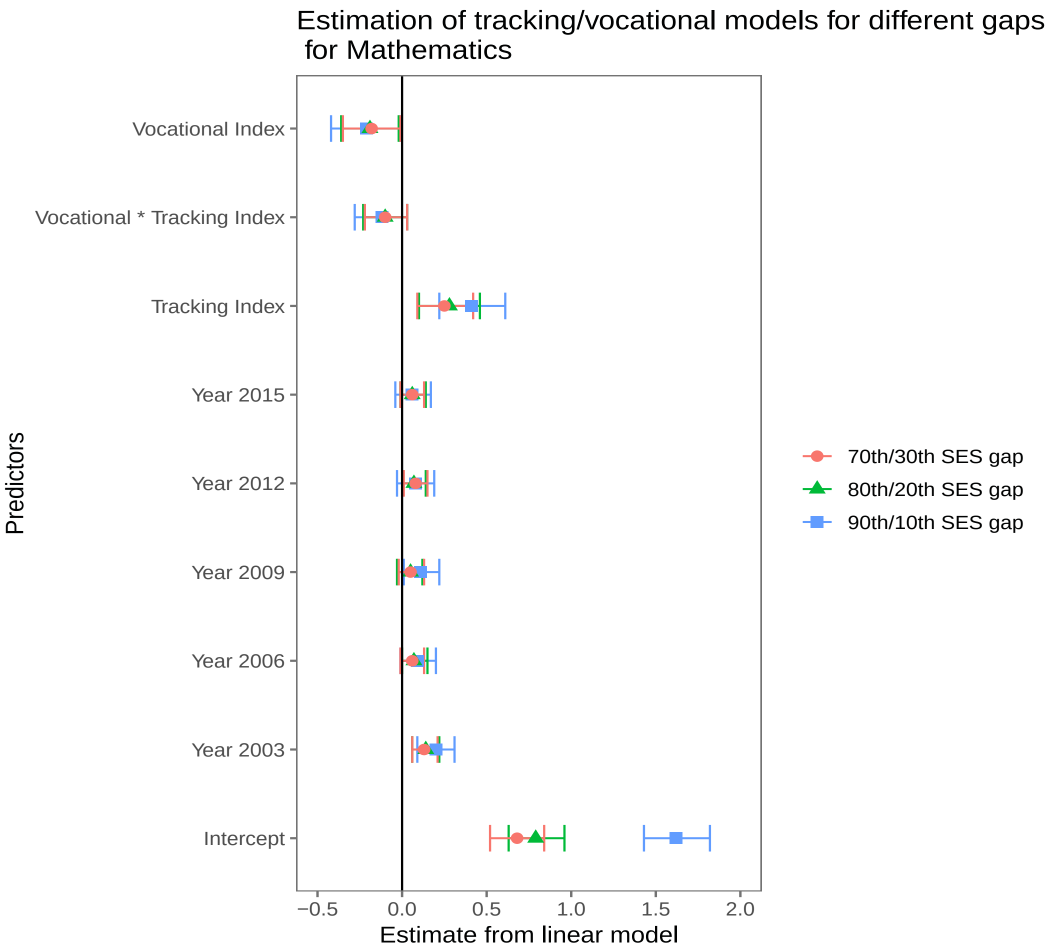


Figure 3.9: Explaining achievement gaps - Model comparisons for Mathematics achievement gaps

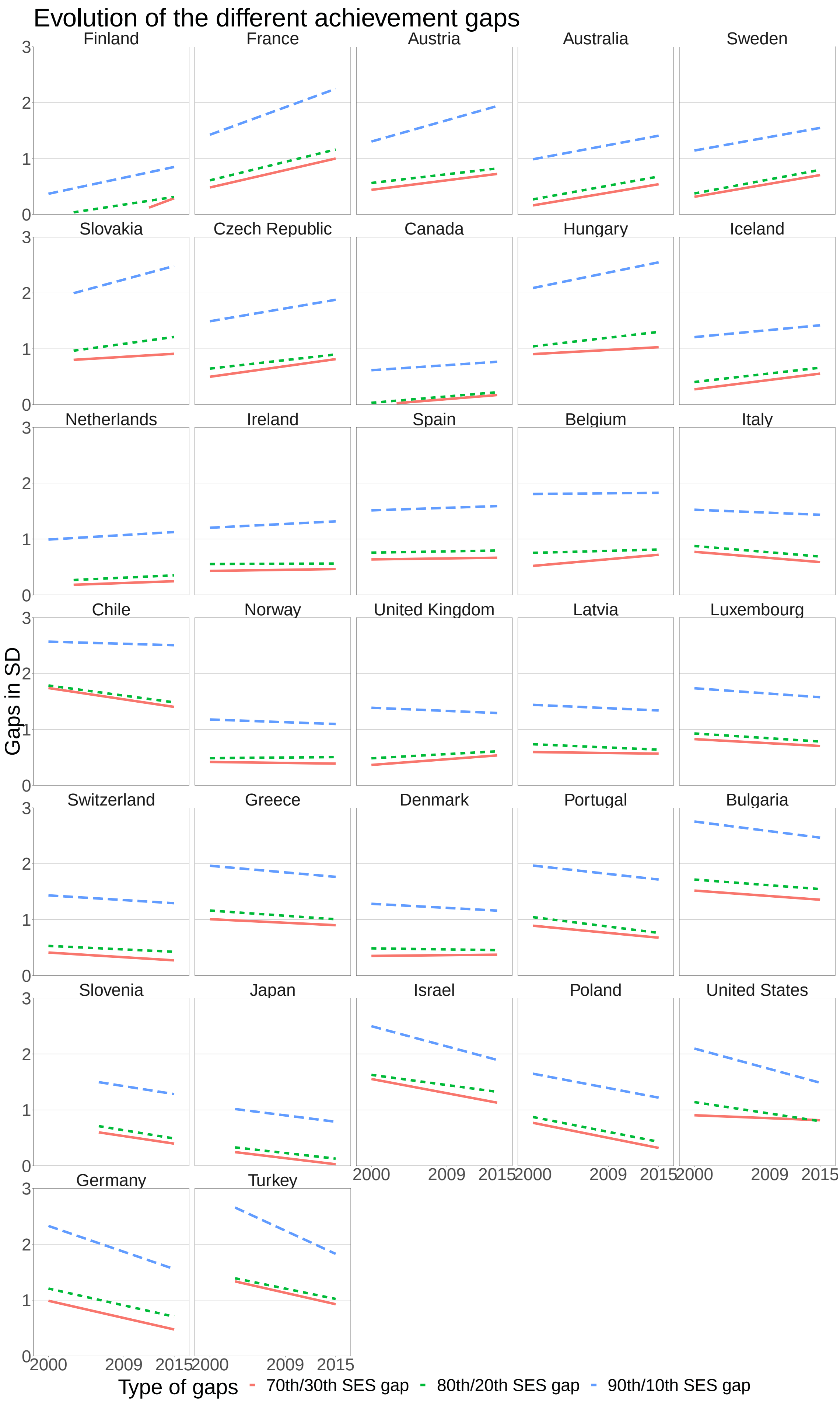


Figure 3.10: Evolution of the achievement gap for all countries in several gaps in Mathematics

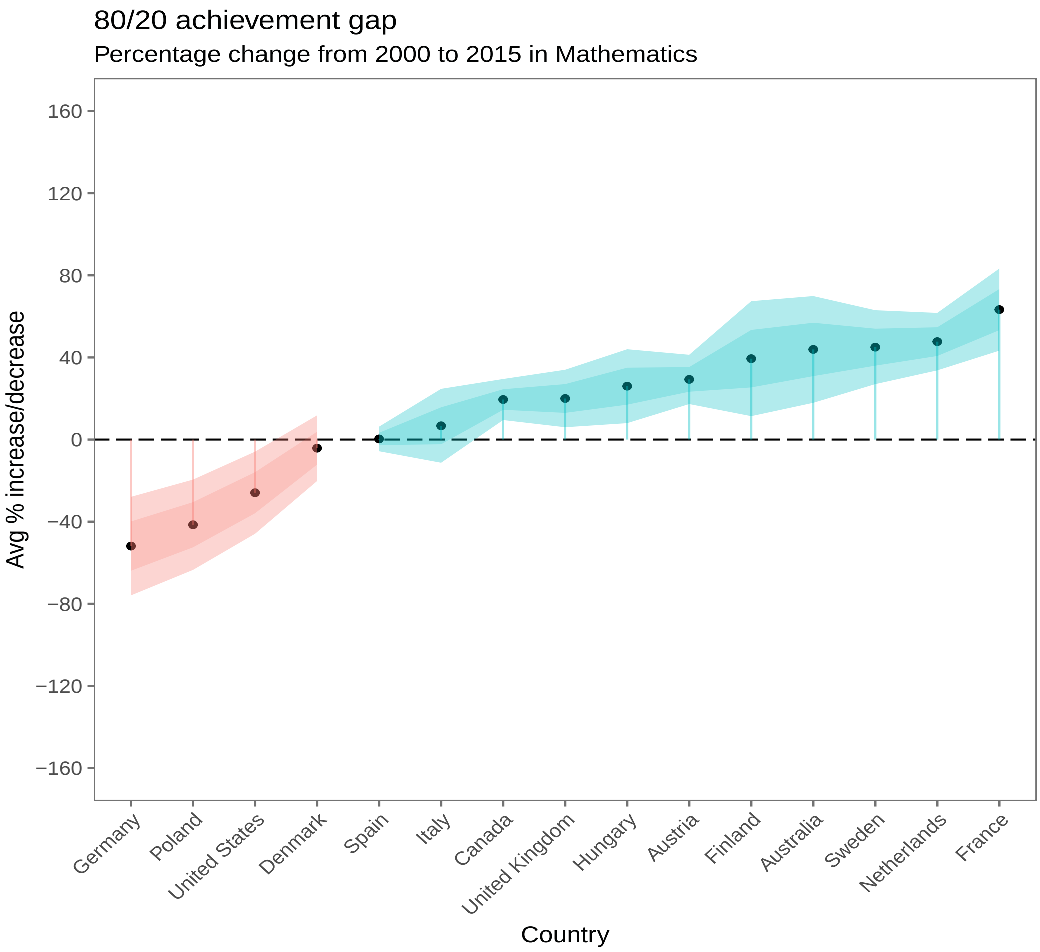


Figure 3.11: Percentage change in the 80/20 achievement gap from 2000 to 2015 in Mathematics

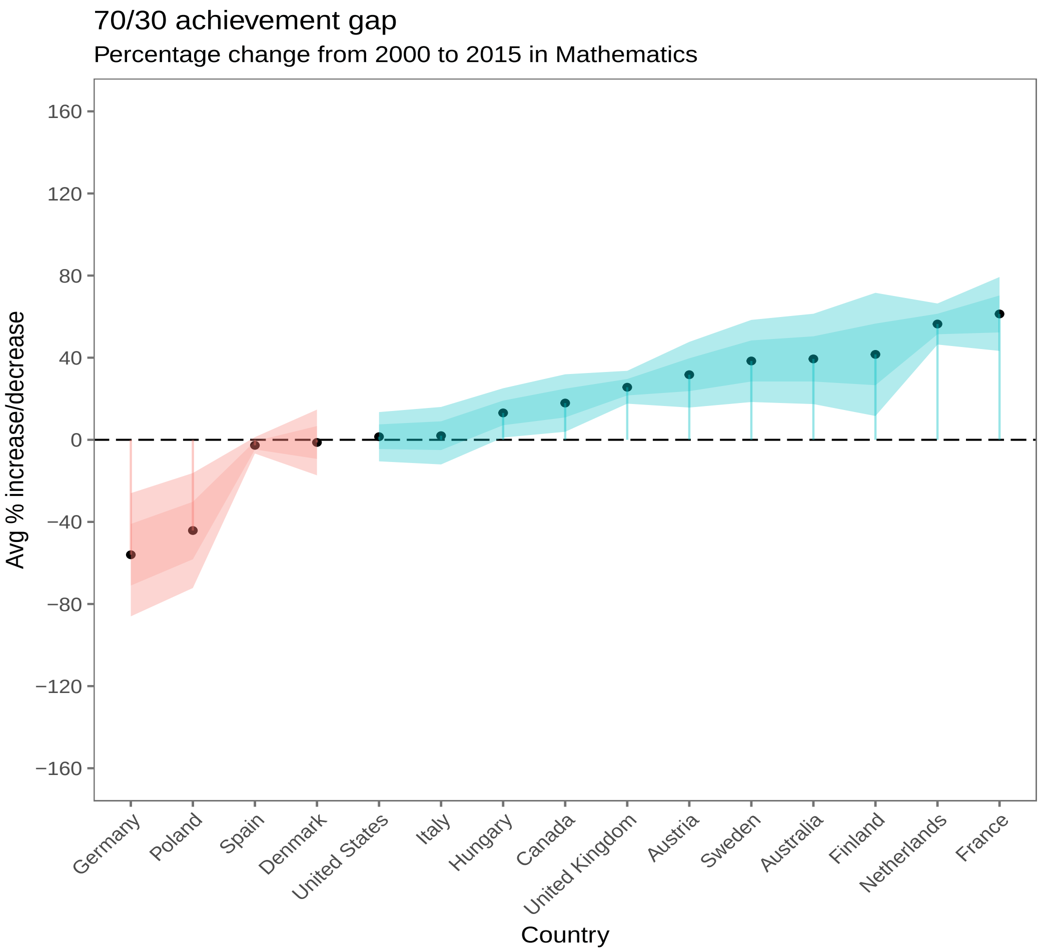


Figure 3.12: Percentage change in the 70/30 achievement gap from 2000 to 2015 in Mathematics

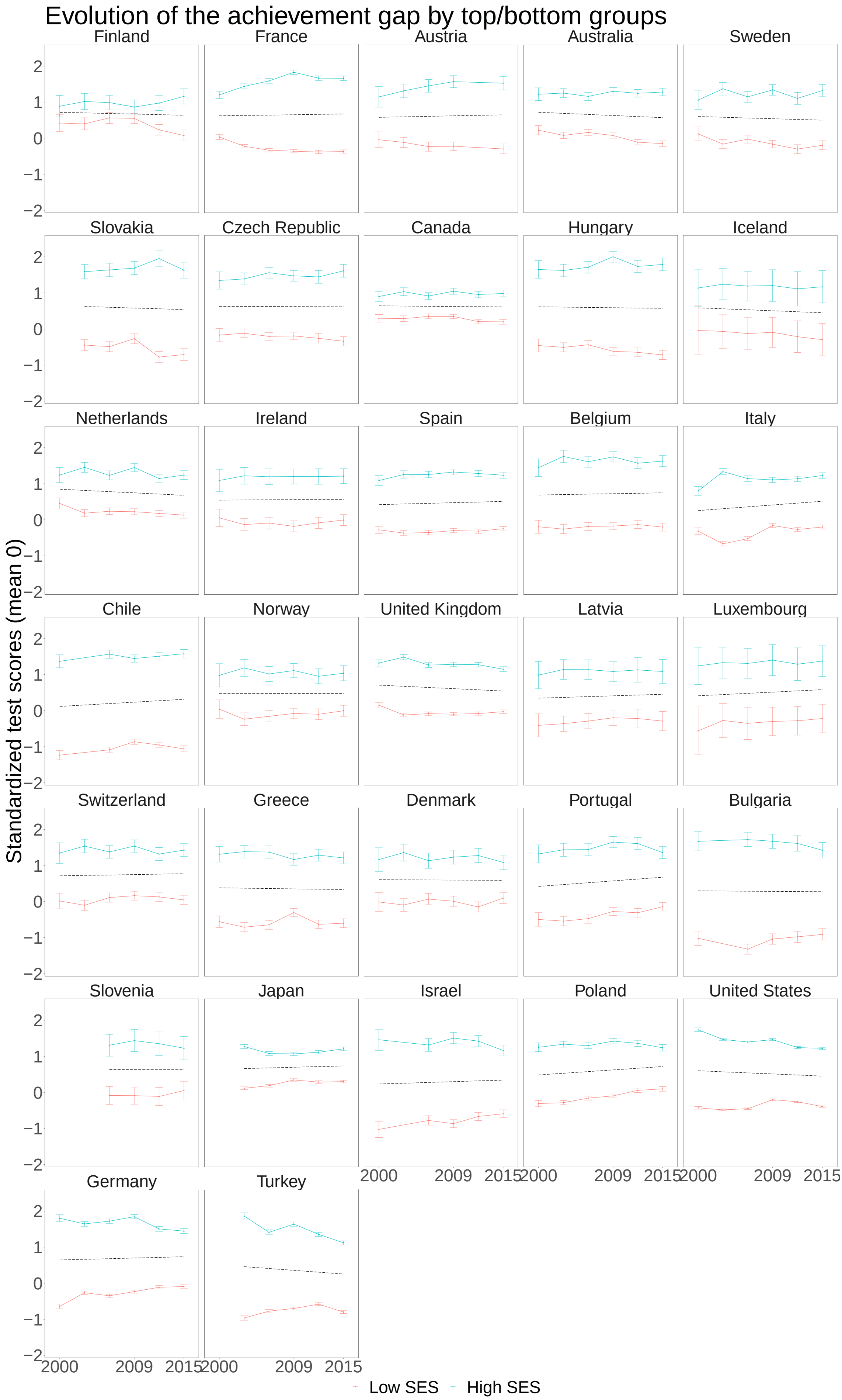


Figure 3.13: Evolution of the achievement gap for separate SES groups for all countries

Countries	# of tracks	Age of selection	% of curric tracked	Std. Voc	Std. tracking
Australia	1	16	0.15	0.97	-1.04
Austria	4	10	0.67	1.7	1.82
Belgium	4	12	0.5	0.95	1.02
Bulgaria	2	14	0.36	0.62	-0.02
Canada	1	16	0	-1.72	-1.32
Chile	2	13	0.42	-0.16	0.32
Czech Republic	5	11	0.62	1.74	1.62
Denmark	1	16	0.25	0.45	-0.87
Finland	1	16	0.25	0.74	-0.87
France	2	15	0.25	0.39	-0.47
Germany	4	10	0.69	0.89	1.86
Greece	2	15	0.25	-0.31	-0.47
Hungary	3	11	0.67	-0.7	1.42
Iceland	1	16	0.29	-0.14	-0.81
Ireland	4	15	0.18	-0.35	-0.3
Israel	2	15	0.48	-0.27	-0.06
Italy	3	14	0.38	0.95	0.17
Latvia	3	16	0.25	-0.18	-0.58
Luxembourg	4	13	0.46	0.99	0.7
Netherlands	4	12	0.45	1.26	0.94
Norway	1	16	0.15	0.89	-1.04
Poland	3	15	0.38	0.3	-0.08
Portugal	3	15	0.25	-0.44	-0.33
Spain	1	16	0.17	0	-1.02
Sweden	1	16	0.25	0.69	-0.87
Switzerland	4	15	0.27	1.08	-0.14
United Kingdom	1	16	0.15	0.47	-1.04
United States	1	16	0	-1.84	-1.32
Japan	2	15	0.25	-0.73	-0.47
Slovakia	5	11	0.62	1.49	1.62
Turkey	3	11	0.55	-0.14	1.2
Slovenia	5	15	0.33	1.06	0.12

Table 3.5: Curricular tracking statistics for all countries

Country	Available years
Slovenia	4
Austria	5
Bulgaria	5
Chile	5
Israel	5
Japan	5
Slovakia	5
Turkey	5
Australia	6
Belgium	6
Canada	6
Czech Republic	6
Denmark	6
Finland	6
France	6
Germany	6
Greece	6
Hungary	6
Iceland	6
Ireland	6
Italy	6
Latvia	6
Luxembourg	6
Netherlands	6
Norway	6
Poland	6
Portugal	6
Spain	6
Sweden	6
Switzerland	6
United Kingdom	6
United States	6

Table 3.7: Number of years in which countries have data (2000-2015, every three years)

CHAPTER 4

The long-term relationship between early education and educational trajectories: the case for GED recipients

Abstract: This chapter attempts to establish a relationship between early education and particular educational trajectories in young adulthood in the United States. Building on the literature on early education and future life outcomes, this chapter concentrates specifically on the odds of attaining a high school degree over a GED/No qualification, attaining a high school degree over a GED and attaining a GED over no qualifications. Exploiting the detailed care history of children in more than 17 years of data from the Panel Study of Income Dynamics (PSID), estimates show that being in preschool relative to being cared for by one's own parents is associated with 25% higher odds of attaining a high school degree relative to having a GED/No qualifications. These results are also replicated when comparing the odds against graduating high school vs GED and attaining a GED vs No qualifications with an odds increase of 11% and 60% respectively. Moreover, the preschool association holds as well when comparing against all other types of care together in all competing educational trajectories described above. An interaction between the care variable and the SES origin turns out to be insignificant although this just might be because of weak statistical power. Extensive simulations and sensitivity checks are implemented to confirm the strength and stableness of the results and all of them point in the same direction: the relationship is stable and moderate.

4.1 Introduction

Research on early education has accumulated convincing evidence in favor of helping disadvantaged children catch up with their better-off peers. Many countries have attempted to implement several early education interventions but the country that has experimented the most with these types of interventions is the United States. For example, Head Start, a nation-wide program that targets disadvantaged children with early education and parental assistance, is the oldest early intervention program still running today since 1965. As inequality of opportunity seems to have stagnated in the U.S since the early 80's (Chetty et al., 2016), the importance of these programs has reached mainstream attention, up to the point that the White House produced a report urging the administration to increase funding for early education programs (The White House, 2014).

However, early education programs cannot be seen as a silver bullet. One of the most debated criticisms of these programs is whether its effect holds as time passes by (J. Heckman, 2013). Even though there is still debate around the topic, the evidence shows that the early schooling effect does hold, even up to 40 years later (Schweinhart and Weikart, 1981). This has been corroborated by J. Heckman and Raut (2016) as they find that a subsidized early education program for children coming from low socio-economic backgrounds will reflect an increase in intergenerational income mobility as well as more children being better educated in the future.

In these long-term studies, sociologists and economists investigate whether quality early education is associated with increased cognitive abilities, decreased criminal activity, improved health outcomes, among other things. In properly randomized trials this has been causally linked rather than merely correlated. This chapter looks to extend this literature by exploring whether early education is associated with particular educational trajectories in early adulthood.

Educational trajectories have been studied before in the context of early education (J. Heckman and Raut, 2016), but to the best of my knowledge, nobody has concentrated on the GED diploma, a qualification that has been gaining ground in the United States for the past 20 years. The GED diploma is a series of tests that non high school graduates can take to achieve high school degrees later on in their life. This aims to lower the percentage of non graduates and increase job qualifications in the population to improve employment opportunities.

Specifically, this chapter attempts to test whether attending early education is associated with greater odds of graduating high school over attaining a GED diploma. In the same line, it also tests whether early education is associated with greater odds of achieving a GED diploma over dropping out of high school. Finally, the chapter explores whether the previous questions differ by the SES of the respondents.

Studies dealing with the long-term benefits of early education are usually constrained to study the topic with non-experimental data since very few experiments have been able to follow children until adulthood. In the cases where they have been able to, researchers are reluctant to release their data. For this reason most of the research on early education in the United States is done using surveys such as the National Longitudinal Survey of Youth 1997 (NLSY97) or the Early Childhood Longitudinal Study (ECLS). These two sources of information allow scientists to follow children who have not had any sort of treatment for a few years. However, this chapter uses a survey which has not been used very often in early education research although freely available: the Panel Study of Income Dynamics (PSID). PSID is the only survey of the sort which is still active today and has background information on families since 1968.

Using the Child Development Supplement (CDS) from PSID, which has 4 follow-up surveys establishing a time frame of nearly 20 years, I exploit rich information on the detailed care history of the child before entering high school together with her/his background information before he/she was born. The combination of these two data sources allows to build a detailed description of what type of care the child was in, who the child was cared by, the reason why the child was enrolled in that care, among other things. This information is then linked to the background information of their parents which has over 40 years of information from whenever they enrolled in the survey.

The results show that across all models being in preschool relative to being cared for by one's parents is associated with higher odds of finishing a high school degree relative to either GED or being a dropout. For example, in two different variants of the preschool category, children who participated in preschool relative to those cared for by their parents and preschoolers against all other types of care saw an increase in odds of 25% and 18%

respectively. These odds were specifically for attaining a high school degree rather than attaining a GED or being a dropout. These results are also replicated when comparing the odds against graduating high school vs GED and attaining a GED vs No qualifications with an odds increase of 11% and 60% respectively. Moreover, the preschool association holds as well when comparing against all other types of care together in all competing educational trajectories described above.

Unfortunately, all of these estimations run the risk of low statistical power given the low sample size due to the attrition in follow-up surveys. For that reason, I provide several tests that show that under hypothetical replications, these estimates would hold their magnitude as well as their direction.

The structure of the chapter is as follows. I first introduce the vast literature on early interventions and highlight the uniqueness of the American educational system. I then introduce the research question and the hypothesis. Next I describe the methodology and data while the results section shows the empirical results and the sensitivity simulations. To finish, I conclude with the limitations of the study and the discussion of the results.

4.2 Literature Review

4.2.1 Early interventions

The growing support for child-related policies in the U.S has been gaining ground steadily as the percentage of child poverty grows continuously year after year (Children Trends, 2016; Moore et al., 2002). At the same time, and most likely related, the gap in cognitive abilities has been growing consistently between social classes (Reardon, 2011). Darkening the landscape even further, the work of Raj Chetty and his colleagues (Chetty et al., 2014; Chetty et al., 2016) has shown that the land of opportunity – something which every American has cherished – is no longer true, if it ever was at some point. Social classes are becoming more distant and far from converging.

It is for these reasons that early interventions have been considered as the most cost-efficient solution to equalizing opportunities. Many reviews have demonstrated the efficiency of this policy, but specifically Anderson et al. (2003) find that most early interventions that were rigorously evaluated had positive effects on children. Most importantly, and something argued by Waldfogel (2006), is that in nearly all studies the costs of implementing the program and subsidizing the enrollment fees are outweighed by the decrease in future health costs, criminal activity and labor productivity. This is indeed an attractive solution, specially for policymakers. The research available thus far highlights how a child, which is deemed to be poor and excluded, tends to improve his or her life earnings, job opportunities and health conditions, among other areas.

But implementing universal early education programs is not a smoking gun. As the mounting evidence suggests, early interventions do not improve development as much as we'd like if they are not of a reasonable quality (Waldfogel, 2006, chap. 3). Numerous reviews have found that early intervention programs – these can vary from before birth to before kindergarten – by themselves do not produce lasting impacts (J. Heckman and Krueger, 2004). But the same programs, combined with quality characteristics such as highly educated teachers, full time weekly participation and small teacher-to-student ratios, produce sizable and significant improvements in outcomes measured as far as 40 years later (Elango et al., 2015).

The proliferation of these programs has allowed economists to estimate the dollar value benefit for the associated costs. Some cost-effective analysis suggests that for every $1 dollar spent on a child in their early years, there is a return ratio of nearly $7 to $12 dollars (J. Heckman et al., 2010; Elango et al., 2015). Despite the fact that most research suggests positive results, some evaluations show the contrary. A study by Currie and D. Thomas (1998) evaluated whether the impact of Head Start was cost effective but they did so disagreggating by race. They found that for African Americans the costs outweighed the benefits, but the results were the opposite for whites. A comprehensive review of all rigorous cost-benefit analysis can be found in Elango et al. (2015).

In the U.S the biggest early education program is Head Start. Head Start is a nationwide program that targets disadvantaged children by giving them the option to enroll in comprehensive early education, aiming to improve their learning environment, helping parents and providing subsidized breakfast and lunch (Garces et al., 2002). This last aspect varies widely between states. Head Start has received a lot of attention because some studies have shown that it does not have a big impact in the development of children

(McKey et al., 1985; Deming, 2009). On the other hand, different studies have concluded that there is indeed a visible effect, but given the erroneous research design of previous studies, others have not been able to find results (Elango et al., 2015).

Although they are most common in the U.S, these programs are also present in other places. Other countries have implemented such programs in small, as well as large scale formats. For example, Norway and Denmark are two countries that introduced large-scale universal subsidized childcare in the late 20th century. Havnes and Mogstad (2011) evaluate the Norwegian reform using regional as well as time variation to exploit the impact that the reform had on labor market indicators. They conclude that children who enrolled in the program had better labor market opportunities as well as higher educational attainment. In a similar vein, they found that these same children had lower odds of being welfare dependent. The Denmark success is very similar. Some authors find an increase in cognitive and non-cognitive abilities as far as 10 years after the early education period ended (Bauchmuller et al., 2014; Esping-Andersen et al., 2012). The success of these interventions is such that several states have implemented these preschool initiatives such as Quebec (Baker et al., 2015) and Georgia and Oklahoma (Cascio and Schanzenbach, 2013)[1].

Despite all of this evidence, some experts on the topic are still skeptical on how generalizable the results are. In some evaluations there is still inconclusive results and many skeptics raise the fact that most interventions have been effective simply because they are small-scale. Their criticism is rooted on the notion that when similar programs have been taken across-states, the results are positive, but they are far from being as effective as the small-scale pilots (Ludwig and Phillips, 2007). In fact, the biggest example they mention is Head Start. Another equally important criticism, and one which this chapter is concerned with demonstrating, is whether the impact of the program tends to fade after the first few years of the program [2].

In a similar line, there is also the need to establish the mechanisms through which these interventions produce their positive results. For example, educators and economists have joined efforts to try and understand which specific aspects of education are decisive

[1] For a detailed description of international programs, see Waldfogel (2006)

[2] Elango et al. (2015) and J. Heckman and Kautz (2013) explain in detail why these findings are erroneous.

but the results are not clear (J. Heckman, 2013). For example, among the main theoretical aspects of an early intervention are its length and intensity of care. While studying the Chicago Initiative Study, Reynolds et al. (2011) evaluated outcomes up to 28 years later and found that the length of the initial program is unrelated to any of the positive outcomes in later life. Others have considered the amount of weekly hours as an indicator of quality and find a strong relationship with cognitive performance (Dämmrich and Esping-Andersen, 2017; Esping-Andersen et al., 2012). In addition, there are other indicators which have been empirically linked to improvements in cognitive abilities such as teacher-to-student ratios and the educational level of the teacher. Despite this, they do not hold up to the test in other studies. This chapter proposes some possible determinants of educational trajectories, concentrating on the GED diploma.

4.2.2 Educational attainment and early interventions

This study is strictly concerned with the educational attainment dimension. Examinations of educational reforms and improvements in social mobility have confirmed that educational expansion leads to a higher educational attainment across the board (Esping-Andersen, 2014). As the higher classes usually have higher attainment before the educational expansion takes place, it is also true that most of the increase in attainment is driven by the working class. Countries where the educational expansion has involved universal education at the early stages are the ones which have seen the greatest improvement in vertical mobility (Havnes and Mogstad, 2011; Bauchmuller et al., 2014).

Early interventions have been found to increase educational attainment in almost all scenarios (Reynolds et al., 2011; Elango et al., 2015; J. Heckman et al., 2010). But the usual design of the studies has not allowed researchers to disaggregate this finding into detailed educational categories. Take the GED example. The General Educational Development (GED) tests are a group of exams which have been developed to assess whether someone has the necessary skills to attain high school-level knowledge. Students who could not graduate high school, but posses the necessary skills to achieve it can request the examination. This test was developed in the 1970's because the dropout rate in the United States was increasing, and the labor market qualifications required more and more that employees had high school diplomas (J. Heckman and Rubinstein, 2001).

The question of whether early development and high school graduation are related has been studied extensively in the literature (Reynolds et al., 2011; Ludwig and Phillips, 2007; Melhuish, 2011) but this has been the case relative to dropping out and not relative to GED attainment. For example, in the context of an educational intervention, Luster and McAdoo (1996) studied the relationship between the Perry Preschool program on educational attainment. However, they do so focusing on the general milestone of attainment without disaggregating. In many other contexts, the main focus is in understanding why dropouts increase. The work of Balfanz et al. (2007) shows that failing courses in middle and high school increased the odds of dropping out yet controlling for their early development care arrangements. On the other hand, Ou (2005) explored even further and found out that the increase in attainment is done through improvement in better support from the families, improved cognitive abilities in early development and other institutional factors such better support from teachers.

The work of Deming (2009) provides an attempt to study the relationship between early education and graduating via the GED test. When evaluating the effectiveness of Head Start, he finds that participating in the program improved high school graduation rates across all U.S states. He goes further and distinguishes whether the improvements were obtained via traditional high school graduation or through graduating via the *GED test*. He confirms the hypothesis that participating in Head Start helped children graduate high school and for those which did not, chances are they went to obtain a GED diploma rather than dropout. This gives starting evidence towards understanding the association between early education and following educational trajectories in young adulthood. However, his results are limited only to the Head Start population, a high selected sample.

This chapter looks to address the question of whether the care arrangement received in infancy is related to children obtaining a GED over dropping out. However, the paper follows the a similar strategy as Deming (2009) where we can compare first whether the care arrangement is related to graduating high school over any other qualification and then for those which did not finish high school, who went to on to obtain a GED diploma. This question is motivated by the fact that very little research has investigated whether early education helps some children deviate from a GED trajectory into finishing high school in the general population. And, as some authors argued in the past, it is really important to find out, considering that GED attainment and high school attainment, although theoretically equal, are rewarded unequally in the labor market (J. Heckman and Rubinstein, 2001).

It is important to highlight that students who drop out of high school have the opportunity to enroll in GED programs later on in their life. This means that whenever there is the interest in studying GED recipients versus dropouts, it should be explained that dropouts have the opportunity of applying for a GED qualification later on in their life. Throughout this paper I refer to the dropout and GED trajectories as mutually exclusive but in reality they are not. This is a decision that had to be made to be able to model different trajectories statistically with the present methodology. The work of J. Heckman and Raut (2016) exemplifies a different methodology that allows to assess competing trajectories using a far more complex methodology.

Educational systems are usually fragmented into curricular or streaming tracks. Some countries have marked tracking structures, where children go into specific educational trajectories and never meet other tracks (Van de Werfhorst and Mijs, 2010a). Countries with little tracking are called comprehensive school systems. In countries with comprehensive school systems the impact of early education is usually less important than in fragmented ones, simply because the stakes of going into an inferior track are null (Brunello and Checchi, 2007). Conversely, in highly tracked countries early education gains much more importance because it can be a definitive step in helping a child jump into the higher tracks (Brunello and Checchi, 2007).

It is paradoxical how the United States has no curricular tracking below tertiary education but has a highly segregated school system, which mimics in all other senses a traditionally tracked curriculum (Holtmann, 2016). The American curriculum is instead tracked by 'ability' groupings. Ability tracking allows children to be in the same school and grade as their peers, but well performing children have the advantage of enrolling in advanced classes.

The American system is of interest for two reasons. First, as outlined above, it has no official tracking below tertiary education, meaning that although early education should be beneficial, it should not be decisive in paving the educational trajectory of the student. In reality it is exactly the opposite. Given that each state is autonomous in most educational decisions, the quality between states does have reasonable variation. Within each state, there are major levels of segregation, with poor children in many instances not even receiving early education. When they reach secondary school they mostly go to segregated public schools which lack the quality standards that will allow them to achieve a 4 year college (Beeghley, 2015).

The second reason why the U.S is interesting is because once students graduate high school or obtain a GED diploma, higher education is extremely tracked, at least in practice. Colleges are divided into 4-year colleges, community colleges and vocational technical institutions. GED graduates theoretically can apply to all types, but in practice, they might not even enroll in higher education. If they do, it is almost always the case in community colleges. A seminal paper by J. Heckman and Rubinstein (2001) studied if obtaining a GED gave the same returns as a high school graduate. They found that simply having a GED rather than a high school diploma gave them lower starting salaries and less labor market opportunities. In a latter study, J. Heckman et al. (2014) actually found that GED recipients are essentially different from high school graduates in key aspects such as perseverance and grit. It is precisely for this reason that the argument for this chapter gains relevance: having quality early investments can enhance the chances of graduating high school rather than follow a GED path or have no qualifications.

The particular mechanism through which a person might follow a GED after dropping out warrants a discussion. The main argument for why a person might enroll in a GED program is because of the labor market opportunities that come associated with the qualification. As was discussed earlier in the paper, the primary objective behind the implementation of the GED qualification was to give 'valid high school diplomas' to people who did not graduate high school in the first place. This was done considering that the demand for companies to have high school qualified employees increased and the high school dropout rate was also increasing (J. Heckman and Rubinstein, 2001). There are of course other explanations such as that citizens want to obtain a degree for the simple fact of obtaining a qualification as a goal. However, the labor market mechanism gains credibility given that high school dropout episodes occur at an age where students are not particularly aware of the long-term consequences of dropping out and thus have other opportunities of getting a formal qualification.

4.3 Research questions and hypothesis

We know that several evaluations suggest that early education can help decrease the high school dropout rate and increase the years of attained education. But we do not know if that boosting effect helps prevent GED attainment in favor of graduating high school in the general population. It could be that early education increases higher education participation but it does so through the GED path (as we have noted before, GED recipients are increasing in the United States). Moreover, and equally important, we do not know if participating in early education is also associated with getting any qualification (GED) over staying a drop out.

More specifically, the questions can be outlined as:

- Is participating in quality early education associated to a student graduating high school rather than obtaining a GED or having no qualifications relative to other types of care?
- Is participating in quality early education associated to a student graduating high school rather than obtaining a GED diploma relative to other types of care?
- Is participating in quality early education associated to a student obtaining a GED diploma rather than having no qualifications relative to other types of care?
- Do children coming from low socio-economic families benefit the most from quality early education in terms of attaining a high school degree over a GED diploma or dropping out?

The first three questions tackle the general notion of whether early care arrangements help the child develop skills necessary to attain different educational diplomas. More specifically, question (1) evaluates whether early education is associated with the 'best' qualification, question (2) evaluates, for those who have any qualification, if early education is associated with the 'best' qualification. This second question is important given that in theory, high school graduation and GED graduation are theoretically equal. Question (3) evaluates whether early education is associated with getting *any* qualification over staying a dropout and finally, the question (4) is more concerned on who benefits more from this type of care.

Defining quality early education is difficult because no indicator is standard across all contexts. Moreover, there is mixed evidence on classical quality indicators, for example on student-teacher ratios (Bruns et al., 2011), so defining a clear benchmark is not the aim of this chapter. For simplicity, but keeping in line with empirical considerations, in the present chapter I define quality early education as children who participated in preschool education relative to children who were cared for in informal settings, were cared for by their parents or participated in Head Start. This means that the variable has four categories: preschool education, cared for in informal settings, cared at home and cared at Head Start. The idea behind this coding comes from the work of Bradbury et al. (2015) where non-preschool education is often associated with low socio-economic families and poor family environments. This is precisely the case in the United States, the country which is studied in this chapter. The chapter also tests the robustness of the result with the specification of Magnuson et al. (2007), which focuses on preschool education vs all other types of care. Although their particular question if interest is different from the one in the text, their coding for the care variable is very similar to the one used in this model specification. This gives both empirical and theoretical justification for the alternative specification of the model. Moreover, this robustness check resorts to a simpler specification where 1 equals preschool and 0 all other types of care. This variable is clearly more broad and captures the general differences between preschool and other types of care.

However, it is important to highlight that Head Start education is supposed to be, at least in principle, of high quality considering the vulnerable population that is enrolled in the program. We must keep in mind that although Head Start in many instances is pooled together with informal settings – which might contain care alternatives such as home care, informal one-to-one care with a non-relative, among other things – Head Start is more akin to preschool than the other types of cares. Yet, as having explained before, since the sample is so selected it is better to investigate the relationship separately for Head Start. Specially, since other research has also studied the category separately given their unique self-selection (Magnuson et al., 2007; Ludwig and Phillips, 2007; Deming, 2009). For all models in the main text, I run the same specification as Magnuson et al. (2007) but excluding Head start in the appendix for sensitivity purposes.

Having said that, the hypothesis that follow each question are:

- Conditional on achievement of a high school equivalent qualification, participating in preschool education relative to home care, informal care and Head Start increases high school attainment over GED attainment or being a dropout

- Conditional on achievement of a high school equivalent qualification, participating in preschool education relative to home care, informal care and Head Start increases high school attainment over GED attainment

- Conditional on achievement of a high school equivalent qualification, participating in preschool education relative to home care, informal care and Head Start increases GED attainment over being a dropout

- Children coming from low socio-economic backgrounds benefit the most from the preschool education

The first three hypothesis are formulated based on the persistent findings of preschool being associated with long term improvements. For example, there is international evidence showing that participating in early education improves cognitive abilities up to 10 years later (Bauchmuller et al., 2014) and labor market opportunities (Havnes and Mogstad, 2011). For the U.S, the thorough review by Elango et al. (2015) confirms that early education improves the chances of reducing welfare dependency and criminal activities. When pooled together, all of this information points out that preschool educated children will have greater odds of usually achieving the best qualification (that is, high school education vs all other types of educations and any qualification vs being a dropout).

The specific rationale through which this mechanism occurs is not easy to pinpoint given that it is endogenous to other likely mechanisms. For example, it is easy to confound the fact that children with more educated parents will more likely graduate high school on time but this might be due to their parents influence (through extra curricular activities which promote responsibility, cognitive development, etc..) and not through the role of preschool. However, the extensive work by James Heckman shows a particular boosting effect of preschool given that it is one of the main equalizers between social classes specially for children who had a particularly non-enriching childhood. One example is the work of

J. Heckman et al. (2014) where they dedicate a complete chapter 'Fostering and Measuring Skills: Interventions That Improve Character and Cognition' to researching the literature on early education and it's indirect association to graduating high school on time. In it, they discuss that GED enrollment is increasing, and even more worrying, it is because students are dropping out given that they can follow a GED later on. That is, students are beginning to see this as a second-chance that they can take and thus quit school earlier. However, they also discuss that one of the most promising alternatives to this is the early interventions because they usually equip children with soft skills such as grit. Again, this must be interpreted with caution as this is endogenous with other explanations, usually related to family resources.

One concrete example of this is the work of J. J. Heckman and Kautz (2012). They show that children who skip school tend to reflect the absence of skills such as responsibility and perseverance even at very early ages [3]. Once we control for the influence of their parent's education and cultural background of the household, then it becomes more plausible that preschool can provide a particular experience that might elicit children to be more proactive at school. This cannot be directly linked to graduating on time many years later but it can be argued that it is associated with a behavior that begins at preschool and early education given that it is one of the first experiences at a schooling institution.

Similarly, the work of J. Heckman and Kautz (2013) suggests that even though preschool education improves life-long opportunities, children coming from low socio-economic environments experience a particularly strong improvement. Building on that previous argument, then the formulation of the fourth hypothesis becomes more evident.

4.4 Methods

4.4.1 Data

The empirical analysis will be carried out using The Panel Study of Income Dynamics (PSID) dataset on children interviewed in 1997 together with their parents. The study

[3] Note that for young children, this does not make sense because they cannot choose to skip school. However, they find that net of parent's education, being in early education increases these soft skills as well

began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Information on these individuals and their descendants has been collected continuously, including data covering employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, education, and numerous other topics. I will be using the Child Development Supplement (CDS), which is one research component of the Panel Study of Income Dynamics (PSID).

While PSID has always collected information about children, in 1997 they supplemented their main data collection with additional information on 0-12 year-old children and their parents. The objective was to provide researchers with a comprehensive, nationally representative and longitudinal data base of children and their families with which to study the dynamic process of early human capital formation. The CDS-I successfully completed interviews with 2,394 families (88%) [4], providing information on 3,563 children. In 2002-2003, CDS re-contacted families in CDS-I who remained active in the PSID panel as of 2001 for CDS-II, and again in 2007-2008 for CDS-III. Another follow up was conducted in 2014-2015 (CDS-IV) for a total of 3 follow ups after the baseline survey. This unique aspect of PSID offers a rich opportunity to study life-trajectories as they have information on a child's family and on parenting practices before/after the child was born or before/after the child entered schooling. This chapter leverages this information to identify the type of care the child received before entering their first year of school while controlling for the *reason* why the child was enrolled in that care. This is a very important distinction given that this is a likely self-selection mechanism which has not been accounted for in the past. That is, if the reason why children enrolled in a particular kind of care is related, for example, to mother's not having enough time (such as going back to school), then this explains the reason for the choosing of a particular care rather than another.

The final analysis of the paper will use the CDS-I sample with questions both retrospectively (from the PSID panel) and from the CDS-II, CDS-III and CDS-IV samples. That is, the PSID website automatically allows to select questions from past/present/future samples (CDS-I, CDS-II, etc...) and carefully creates the dataset automatically with all students matched accordingly. Using the 3,563 students from CDS-I, I merge all types of care arrangements for children in different CDS samples and record their final educational

[4] 88% refers to 88% of the CDS sample rather than the total PSID sample

attainment in the last follow up (CDS-IV, years 2014-2015). This means that for children which were too young in CDS-I to be in pre-kindergarten (age 2), the care arrangement is actually taken from the CDS-II sample, in which they were already eligible to answer the question.

In contrast, for children who were aged, for example, 8 in the CDS-I sample (after pre-kindergarten education in the first sample), the care arrangement variable was asked retrospectively to parents, so they actually answered for when those children were in pre-kindergarten age. The first question asked whether the child participated in kindergarten. If children have not participated in kindergarten, then they are not asked the care arrangement variables (and they have missing values in the care arrangement variable for *that* specific year. In the analysis, their care arrangement is taken for when they already participated in kindergarten in CDS-II). If they have participated (or are participating) in kindergarten then the questionnaire asks for the complete history of care arrangements, specifically before kindergarten. The first question on the care arrangement history stats as:

> ”The next questions ask about the child care arrangements or programs that you have used for CHILD since (Month/Year). We want to start with the first arrangement you used for CHILD and then continue through any additional arrangements you may have used, in the order that you used them. We will end the history when CHILD started kindergarten.”

where it is clearly outlined that this will be the care history of arrangements which will end when CHILD started kindergarten. Once the respondent understand, then the care history begins:

> Starting in (month/year), what was the first type of childcare arrangement or program that you used on a regular basis for CHILD?

These questions continue until last care before kindergarten. The total sample size used in the study is of 1384 children. It is the sample of students from the initial CDS-I sample that contain all non-missing values for the independent and dependent variables.

4.4.2 Coding and methodology

The main dependent variable for this study is whether a child has a high school diploma, a GED equivalence diploma or is a high school dropout in 2014-2015. The educational attainment question asks the respondent for the highest degree they have completed coded like this:

- Less than high school diploma
- GED, no college
- High school graduate, no college
- GED plus some college
- High school graduate plus some college
- GED plus Associate's degree
- High school graduate plus Associate's degree
- GED plus Bachelor's degree
- High school graduate plus Bachelor's degree
- GED plus Master's degree
- High school graduate plus Master's degree
- GED plus Doctoral degree
- High school graduate plus Doctoral degree
- GED plus Medical degree
- High school graduate plus Medical degree
- GED plus Law degree
- High school graduate plus Law degree
- GED plus other degree
- High school graduate plus other degree

With this coding it is fairly easy to obtain the three trajectories mentioned above. Anyone who completed a GED is coded as the GED track, anyone who graduated high school or above can be coded as high school graduates and those with less than a high school diploma that do not have a GED can be coded as dropouts. Because every child who is at least 2 years old in 1997 should be 18 in 2015, anyone who has less than a high school diploma in 2014-2015 is not because they have not finished high school but because they dropped out. Figure 4.1 in the appendix shows the distribution of age in 1997. This suggests that most children were at age 2 or above in 1997 so indeed they should have, in theory, a high school degree. With this design it is clear how to pinpoint the education of each individual. Of course, there is the risk that people in this specific sample will eventually take the GED at some point.

The PSID team begun interviewing children in 1997 and followed up several times until 2014-2015 when the last available wave was conducted. This time frame means that parents with children who were 2 year old in 1997 would be 18-19 year olds in 2014-2015. Similarly, children who were 5 year olds in 1997 would be around 22 in 2014-2015. Children who were between 16 and 35 year olds in 2014-2015 have enough time to have registered for the GED tests as test takers can already take the test at age 16 with parental permission. Moreover, this becomes an even more plausible assumption for the specific sample in this paper considering that around 72% of GED test takers are between ages 16 and 29. More specifically, 22% of test takers are between ages 16 and 19. This figure has been fairly stable for the last decade as can be seen from the official statistics from the National Center for Education Statistics extracted from https://nces.ed.gov/programs/digest/d15/tables/dt15$_2$19.60.asp.

Using the three categories from above I develop three variables.

- High school graduate = 1 / GED attainment + High school dropout = 0: This dependent variable differs from the above by including the respondents who dropped out of high school in the 0 category. With this approach I attempt to gauge how much preschool is related to the traditional trajectory of graduating high school versus any of the other two trajectories. The aim is to assess the competing risk of GED alone and then both together.

- High school graduate = 1 / GED attainment = 0: This dummy excludes the respondents who dropped out of high school because I am interested in looking at the relationship from GED to high school. With this approach I study the chances of graduating high school relative to having dropped out of high school *and* attained a GED.

- GED = 1 / High school dropout = 0: This dummy excludes the respondents who graduated high school because I am interested in looking at the relationship from GED to dropping out. With this approach I study the chances of obtaining a GED relative to having dropped out of high school.

I will model the three variables above by using a logistic regression and estimate odd ratios instead of logistic odds [5]. Because the sample size is low for the GED and dropout categories, some of the estimates might be very uncertain. In section 4.6.1 which describes the sensitivity checks for the models, I explain the procedure that I follow to make robustness simulations and provide plots showing that the estimates do not vary greatly under different scenarios. Additionally, I include the Leave-One-Out (LOO) metric and the AIC to assess how much the models are generalizable to the general population. Moreover, for all the main tables in the results I provide equivalent tables in the appendix that remove sensitive categories to show the robustness of the relationship. Additionally, in all of the models I calculate robust standard errors to remedy the effect of heteroskedasticity.

The model specification is the same for all models using a matrix X of covariates. I explain the coding of these covariates below.

- Care arrangement: This is the main independent variable. This is the last care arrangement before they a child entered formal schooling. The coding here is (i) cared by their parents, (ii) Head Start, (iii) Informal care and (iv) Preschool. This variables can only contain only one care which is the last and ignores the previous arrangements in the care history. Another question which is also possible is whether children were enrolled in several types of care at the same time. However, the question asked by PSID specifically asks for the main type of care and thus children only have one type of care at a time.

- Home Index: This is a composite index created by the PSID team that looks to measure the cultural environment in the household. They formally define it as *a measure*

[5] I use a logistic regression rather than a multinomial regression because I'm interested in comparing one category relative to other categories combined – a comparison model between the three dependent variables to test if the inclusion/exclusion of some trajectories shifts the results

of cognitive stimulation and emotional support that parents provide to their children (Bradley et al., 1988). The PSID team also adapted this index for different age groups, so older as well as young children have been measured. This variable is important given that the odds of improving educational attainment could be explained by the fact that some children have different types of household environments. Presumably, children from low educated families have more negative environments in terms of child development.

- Highest education in the household: the highest education achieved in the household. If it is a lone mother household then her education is used. If both parent's are living in the household then the highest is used. This variable is coded as an ordered categorical variable with over 15 categories. After several modeling approaches, I use this independent variable as a continuous measure as there was not much difference in the model fit (and statistics) for using it as a categorical variable [6]. This variable is an important confounder given that different types of education can explain the income and opportunities of a given family. This variable combined with the Home Index controls for the degree to which the family's environment is influencing the child's opportunities. It also helps to disentangle the family influence from the care arrangement estimations.

- Reason why the child was enrolled in care: This variable describes the reason why the parent's enrolled children in the care presented in the 'Care arrangement' variable. The categories are (i) started/returned work, (ii) increased/changed work hours, (iii) started looking for work, (iv) started school, (v) started other activity, (vi) child needed playmates/activities and (vii) other reasons. I recoded these categories into (i) Child needed play, (ii) Started school, (iii) Started/returned to work and (iv) other reasons as a reference category. The rationale behind coding the reason for care variable was to distinguish between 'meaningful' reasons in the context of improving a child's development. The first category of interest that I wanted to separate was active searching/starting/return to some type of employment. This would capture the effect of time availability and income, depending on the type of

[6] It can be argued that this is wrongly specified as it is properly a categorical variable. However, because as each education category increases there is also an increase on the outcome of interest, there is a perfectly linear relationship between the two variables. To avoid using many degrees of freedom (15 to be exact) under such low sample size setting, I prefer to model it as a continuous variable for simplicity. If the relationship was not linear for all categories, then this would not be the best approach. Fortunately, this is not the case.

job and SES status of the family (which is always controlled for in all models).

A different type of category which would also separate the reason for care was for whether the parent's started school (in nearly all cases, the mother, which is the one who answered these questions). This captures a different effect from job related reasons given that even though they might also have less time availability, there is the likelihood that there is no income coming from that person (in any case this should differ by SES, which is controlled for in the estimations). The two categories outlined above were coded separately, yet they reflect some type of necessity given that the reason why they're enrolling their children at some type of care is because they *need* to do something else. For this reason, the other category which I define separately is for whether the child needed play/playmates. This category reflects somewhat a concern for more soft skills of the child and development rather than being a specific necessity such as finding a job, looking for jobs or started school. Finally, the last category is other types of care, which specifically encompasses the answers 'Started other activity' or 'Others'.

The category 'Other' in this particular case reflects other types of reasons which clearly are not capture by the work/school/play main reasons. This category captures all other reasons for enrollment, something that is difficult to ascertain what contains. For example, this could capture reasons related to the unavailability of grandparent's or other type of family members. In the same line, this category could also be interpreted as a dummy category that captures different types of self-selection; this gains particular relevance substantially as it might explain why they did not enroll their children in informal care (care from relatives or others) in the care arrangement variable, but also if the total sample size within that category is big which would make it somewhat costly to throw the information away. For example, people who answered 'Other' types of care might be endowed with other non-observable characteristics that make them different from other respondents; one example would be capturing more family-oriented households focusing on being cared for by relatives. In order not to dismiss all observations within that category (there are around 400), I have chosen to keep them in the models to adjust for any sort of self-selection within the category, although the specific reasons why they choose being there are not entirely clear.

This variable is one important addition to the literature as it allows to control for the possible self-selection of some children into particular care arrangements. The main problem with interpreting the relationship between early schooling and other outcomes later in life is the inherent self-selection of some children into some types of care/schooling. This is in fact one of the motivations for conducting Randomized Controlled Trials (RCT): to determine the impact of early education in later life outcomes. Most research on the topic, however, focuses on cross-sectional studies which are difficult to adjust for this self-selection, where parents coming from high SES might have some motives (economic and educational) to enroll children in some types of care over other parents. For example, I expect parents from high SES families to have on average greater motivation to enroll their children in high-end schools whereas low SES families to give less weight to this option. This is because of monetary reasons and motivation on the academic expectations of their children. By controlling for the reason why the parents enrolled their children in this type of care, the association controls for a (usual unobservable) source of information which might explain why some children are enrolled in better care (preschool) than other types of care such as cared for by their parents. More specifically, parents which choose to enroll the child in care because they thought the child needed play (while adjusting for SES in the regression models) might reflect other motives in comparison to parents that started school (while adjusting for SES in the models). For this reason, I believe this variable offers a new source of information which adjusts for some type of unobservable preference towards some type of care over another.

It should be noted that this does not mean that any of the associations in the regression models can be interpreted as causal or as the 'true' effect of early education. It merely suggests an association which adjusts for a possible source of self-selection which reflects preference or reasons why they enrolled in any type of care.

- Race: This variable indicates the race of the head of the household coded as (i) white, (ii) other race and (iii) black as the reference group. This variable is an important confounder in terms of the white/black divide in opportunities in adulthood, something present particularly in the American society (Fryer and Levitt, 2004).

- Number of siblings: This variable indicates the number of siblings that the child has. This variable allows to adjust for the fact that some children might live in families where the time allocation to the children is curtailed by having many people living in the household and decreases personalized development.

These variables constitute the right hand side of the equation for all the dependent variables presented above.

In a variation of this model I change the specification of the independent variables slightly. More specifically, I recode the care variable to be a dummy variable where 1 is preschool and 0 is all other types of care (Magnuson et al., 2007). Additionally, I also include a dummy for the number of hours a week that the child attended that care. 1 means full-time (35 hours or more) and 0 means means part-time (below 35 hours a week) [7]. These changes are implemented to test an alternative specification of high quality care without only relying on preschool education. Given than grouping all care arrangements together as 0 can drive the coefficients with small vulnerable groups, such as Head start participants, I also run the same models with the preschool category against all other categories excluding the Head start group in the appendix.

4.5 Descriptives

As described before, the total sample size of the study is 1384. It is the sample that contains all non-missing values for the independent and dependent variables. All models and tables should have the total sample size of 1384 except any the ones that compare high school graduates vs GED, GED vs high school dropouts and models that exclude categories to test for sensitivity in the appendix. For the first model the sample size should be 1292 and for the second model it should be 192. All of these sample sizes are simply the product of the total observations within the three categories of the dependent variables which are 1192 for high school graduates, 100 for GED holders and 92 for high school dropouts. Table 4.1 shows the percentage of respondents (and means, for continuous variables) for each dependent and independent variable.

The main dependent variable, educational attainment in 2014-2015, shows that 86% of the sample has a high schools degree or above, while only 7% has a GED and another

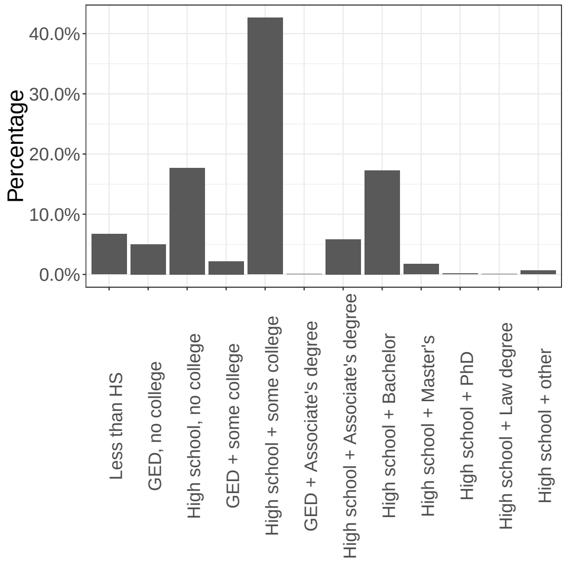
[7] I do not present the raw distributions here but there are two high peaks, one of 40 hours and the other at around 20 hours a week. Following this distribution I recoded the variable into the described dummy.

Variables	Count	Percentage
Educational attainment		
- >= High School	1192	86.0%
- GED	100	7.0%
- No HS	92	7.0%
Care arrangements		
- Cared by parent	420	30.0%
- Head Start	71	5.0%
- Informal care	495	36.0%
- Preschool	398	29.0%
Highest education in HH		
- > Bachelors	326	24.0%
- Bachelors degree	76	5.0%
- High school	404	29.0%
- Less than HS	288	21.0%
- Some college or 2 year college	290	21.0%
Care reason		
- Started/Return work	588	61.0%
- Child needed playmates	141	15.0%
- Other	158	16.0%
- Started school	77	8.0%
# of Siblings		
- 0	263	19.0%
- 1	653	47.0%
- 2	320	23.0%
- 3	93	7.0%
- 4	30	2.0%
- 5	18	1.0%
- 7	2	0.0%
- 10	5	0.0%

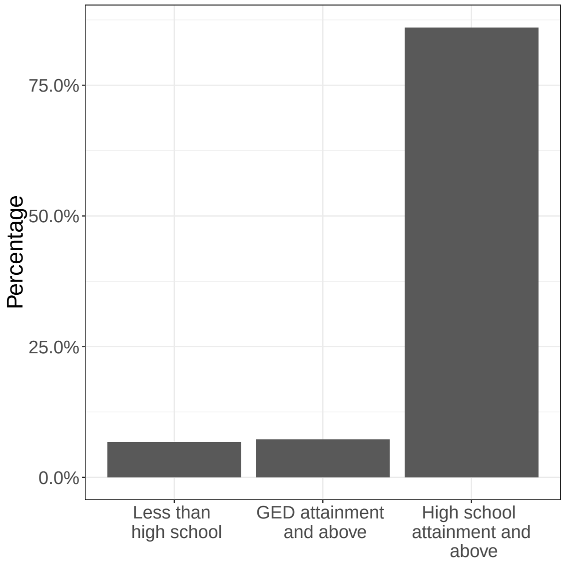
Table 4.1: Descriptives of all dependent variables and independent variables

7% has not any degree whatsoever. As per the other variables, we see that the majority of children were cared for by their parents or through informal care, where 30% was cared for by parents themselves and the other 36% was cared for through informal settings such as cared for by relatives or non-relatives in informal settings. Only about 29% were enrolled in some type of preschool and the remaining 5% participated in Headstart. As expected, the Headstart group is the smallest, being a targeted intervention towards very disadvantaged families. Moving on to the education of the household variable, the sample is mainly composed of families where the highest education in the household is at a high school degree or above (58%) and the remaining qualifications having 21% respectively. As per the race of the head of the family, nearly half are whites and 42% are blacks. As per the care reason, the majority of families did so because they 'Started/Returned to work' (61%) whereas 15% did it because the child needed play. This an important number and highlights the importance of the variable. In theory, we should expect that higher educated families choose reasons related to the development of the child, such as child needed play, instead of the reason of 'Started/Return to work'. Given that the majority of the sample has either a high school degree or above, this variable would help to disentangle within these education categories the reason or motivation why some students were self-selected into certain types of cares.

Moving on more specifically to the dependent variable, figure 4.1 shows the percentages of the educational attainment in 2014-2015 categories.



(a) Original educational attainment



(b) Recoded educational attainment

Figure 4.1: Raw and recoded educational attainment in 2014-2015

Figure 4.1a shows the percentages. The big peak lies on those who graduated high school and did some type of college. Equally likely we also see those with a high school degree and no college and those with a high school degree and a bachelor's degree. A summarized version of this variable is presented on figure 4.1b. Most of the sample graduated high school with a total of 86.13%, and around 7.23% are GED recipients and 6.65% did not graduate high school. In plain numbers, the N's are 1192, 100 and 92 respectively. In this new coding missing values are excluded.

Throughout this section I explore the variables of interest using visual representations rather than tables because they are more useful at spotting patterns we otherwise would not be able to see.

To confirm that excluding missing values from the educational attainment in 2014-2015 does not bias the estimates towards self-selected respondents, I present table 4.2 in the appendix which shows the composition of the main independent variables for missing and non-missing responses in educational attainment in 2014-2015. Overall, the percentages between both groups look fairly similar. For example, for the highest education in the household the non-missing categories have proportions of 24%, 6%, 30%, 19% and 21% respectively for 'More than bachelors', 'Bachelors', 'High school', 'Less than HS', 'Some college or 2 years college'. Similarly, the missing group has 23%, 5%, 30%, 21% and 22%, showing no significant differences between both groups.

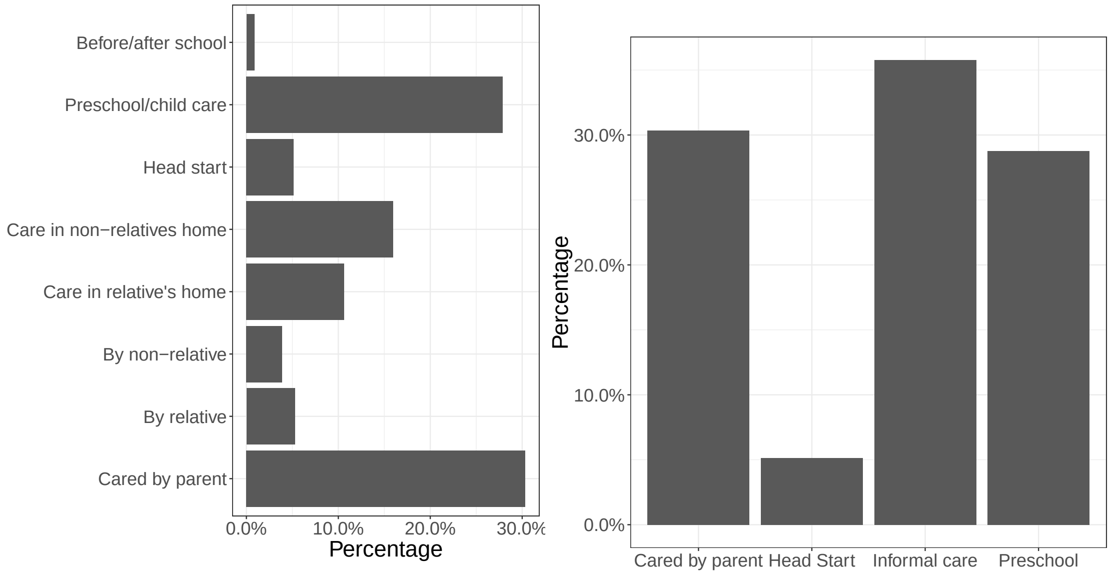
In most other covariates the percentages are similar with the exception of two cases. For the non-missing category, 23% of respondents were cared for by their parents whereas in the missing category this increases to 38%. For all other types of care arrangements both groups have similar percentages. The other case is for the reason why they enrolled in that care, where non-missing respondents had 7% less respondents in the 'Started/Returned work' category than in the missing category. Although roughly speaking both groups are very similar and do not show strong differences in the composition across all covariates, the results must be interpreted with caution when it comes to the care arrangement variable.

Considering that those who had a missing value in their educational attainment in 2014-2015 had greater percentage of parental care rather than formal care such as preschool, there is the risk that the sample who responded to the educational attainment in 2014-2015 are inherently self-selected into having better educational experiences than those who did not respond. Of course, this is only the case for this particular category (parental care) and the percentage is only 8% higher than the other group. However, the results could be potentially optimistic for the current sample with upwardly biased coefficients, all due to the fact that the composition that responded the educational attainment question had a previously higher level of family and social background.

As all other categories for each independent variables seemed to be balanced between both missing and non-missing groups (including typical proxies of SES background in the United States such as highest educational level in the household and the race of the head of the household), then it seems that both samples are fairly balanced. However, to be cautious, in any worst case scenario we must interpret the care arrangement variable with caution as it is the only variable with a category that differs in balance, which could be a visible mechanism to self-selection which I cannot adjust for.

The main independent variable is the type of care they received when they were in pre-K age (between 4-6 years old). A child can participate in several care arrangements before kindergarten; the possible types of care are presented in figure 4.2a.

As we can see in figure 4.2a, the vast majority of the sample was cared for by a parent or was in preschool before kindergarten. The other respondents are distributed unevenly



(a) Percentage of type of care received by each child

(b) Recoded type of care

Figure 4.2: Raw and recoded care arrangement before kindergarten

between informal types of care and Head Start.

In figure 4.2b I plot the distribution of the recoded care arrangements to capture more broad categories. The recoding equivalence I described in the coding section was: informal care is care either by relatives or non-relatives at home or outside home, cared by parent is only the cared by parent category, preschool is preschool/child care and before/after school and Head Start is only Head Start. Most respondents were cared for in informal settings or by their parents. This shows that generally speaking the sample is looks like is composed of children coming from low socio-economic background as this type of care is generally of low quality (Bradbury et al., 2015). In spite of this, there is a reasonable share of children who were enrolled in formal preschool education.

In addition to the care arrangement variable, the questionnaire contains information on the reason why the child enrolled in this care and how many days and hours per week the child visited the care arrangement.

In terms of interpretation, this means that the coefficients presented in the models could be susceptible to important changes if we were to control for this type of self-selection. More concretely, the difference between children cared for by their parents and children cared for in other settings could in fact disappear, once we control for this type of self-selection. However, until further research can reliably record such type of information, exploratory and descriptive studies such as this one are bound to investigate the relationship further.

Table 4.2 presents the relationship between the child's education in 2014-2015 and the child's care arrangements.

Care Arrangement	< High school	GED or above	High school or above
Cared by parent	40%	33%	29%
Head Start	4%	14%	4%
Informal care	37%	33%	36%
Preschool	19%	20%	30%
Total	100%	100%	100%

Table 4.2: Percentage of respondent's within the care arrangement they received in early childhood by the respondent's education in 2014-2015

There seems to be a relationship between the two variables. High school educated respondents seem to be the ones with more preschool education (30%) and the percentage goes down with each educational level (20% for GED and 19% for less than high school educated), as expected. 40% of respondents with less than high school educated were educated by their parents and this percentage decreases with higher educational credentials, also as expected. These results line up as expected, with a clear correlation between early education-education levels in 2014-2015. However, informal care seems to be slightly puzzling as 36% of high school educated children had this type of care and less than high school educated respondents had a similar percentage with 37%. This is something to consider (and to remember when interpreting the results), as most research highlights that this category is of lower quality and thus we should not expect similar proportions between opposite educational ladders. I have added table 4.1 with the raw sample size of each category in the appendix.

Another important independent variable is whether the child went to full time or part time care before kindergarten. This variable highlights the quality aspect of the education to a certain extent. In order to understand the correlation between the time spent in care and educational attainment in 2014-2015, I plot figure 4.3.

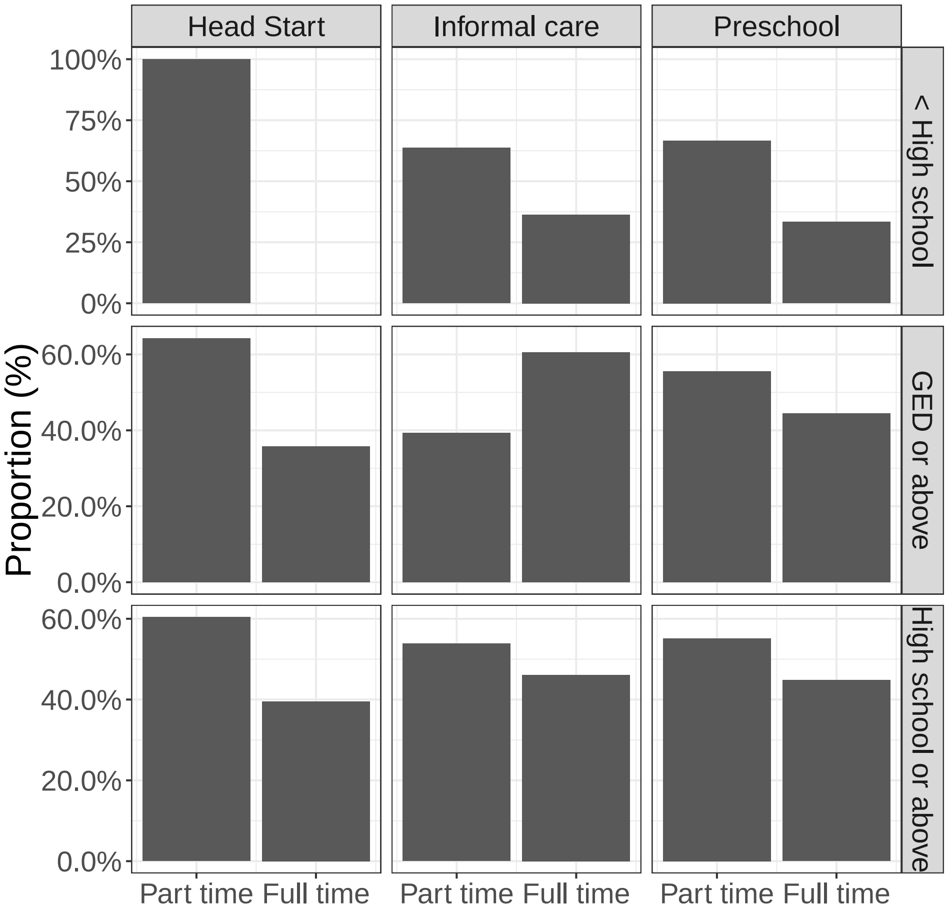


Figure 4.3: Type of care by intensity and respondent's education

An interesting pattern is evident. We can already see the differences in attainment by 'dosage' of schooling. First, respondents who attained high school degree or dropped out by 2014-2015 (right legend) were mostly in part-time care rather than full-time. Full time care is perhaps not that widespread given that children are very young, and in fact those who receive full time care are because of specific reasons (such as the mother working). On the other hand, GED holders do have a higher share of full time enrollment in both their care arrangements. This is surprising given that it is only GED holders and not also the respondents who dropped out of high school.

4.6 Results

Table 4.3 presents the first batch of models. The dependent variable for these models is whether the person graduated high school over attaining a GED or did not attain any qualification. The modelling was conducted using a logistic regression, with odd ratios instead of logit coefficients. The first model includes only the care arrangement the child received, which is the main independent variable, and the subsequent models include all the control variables outlined before. For the care arrangement variable the reference category is whether the child was cared by their parents.

Table 4.3: Logistic regression with dependent variable graduated high school (1) over attaining GED + high school dropout (0) expressed as odd ratios with robust standard errors

	Dependent variable:					
	1 = High school / 0 = GED attainment + No qualification					
	(1)	(2)	(3)	(4)	(5)	(6)
Care arrangement: Head Start	0.59* (0.18)	0.62 (0.21)	0.62 (0.22)	0.63 (0.23)	0.75 (0.28)	0.71 (0.27)
Care arrangement: Informal care	1.30 (0.24)	1.39* (0.27)	1.31 (0.26)	1.05 (0.21)	1.14 (0.24)	1.07 (0.23)
Care arrangement: Preschool	1.89*** (0.41)	1.86** (0.46)	1.68** (0.42)	1.16 (0.30)	1.33 (0.35)	1.25 (0.33)
Care reason: Child needed play		1.83 (0.68)	1.70 (0.65)	1.49 (0.59)	1.51 (0.61)	1.46 (0.59)
Care reason: Other (Ref: Started/Returned to work)		0.83 (0.21)	0.80 (0.21)	0.81 (0.22)	0.82 (0.22)	0.77 (0.21)
Care reason: Started school		0.50** (0.16)	0.55* (0.18)	0.53* (0.17)	0.58 (0.19)	0.54* (0.18)
Home index			1.11*** (0.02)	1.06** (0.03)	1.04* (0.03)	1.05* (0.03)
Highest education in HH				1.74*** (0.12)	1.74*** (0.13)	1.74*** (0.13)
Race: Other race (Ref: Black)					2.69*** (0.87)	2.73*** (0.90)
Race: White					1.40* (0.27)	1.32 (0.25)
Number of siblings						0.82*** (0.06)
Constant	5.00*** (0.65)	5.00*** (0.65)	0.72 (0.28)	0.62 (0.27)	0.58 (0.26)	0.73 (0.33)
Number of observations	1384	1384	1384	1384	1384	1384
Log Likelihood	-549.19	-544.3	-533.05	-500.14	-493.97	-490.41
AIC	1106.38	1102.6	1082.11	1018.28	1009.94	1004.81
LOO	0.8	0.8	0.78	0.73	0.73	0.72

Note: *p<0.1; **p<0.05; ***p<0.01

As we can see from the first model, the preschool category boosts positively the odds of achieving a high school degree with the estimate being 1.89 odds ratios. The informal care category also has a positive relationship, boosting the odds of achieving high school education by 1.30 odds. Finally, Head starters seem not be benefit from their education as the relationship shows that participating in head start decreases the odds of achieving a high school diploma by 41% (1 - 0.59). Once we incorporate the reason for care variable, the odds of preschool education are very similar yet the ones for informal care are increased by 9%. For Head starters, the odds of achieving a high school education increased only slightly to 39%. Model (3) and (4) include both the home index and the highest education in the household respectively and we see a decrease in the odds of preschool of education to 1.16 odds ratios. However, when race and number of siblings is accounted for (model (5) and (6)), the odds increase to about 1.33 and 1.25. In a similar line, the Head Start category kept increasing it's odds of achieving high school education, yet by model (6) the odds are still negative with an odds ratio of 0.71. On the other hand, the informal care category decreases it's positive odds to 1.07 in the last model, showing nearly no relationship whatsoever with achieving a high school education [8].

All in all these models so what is expected based the hypothesis. Preschool education has the strongest relationship to graduating high school, the type of care that is considered to have the best quality. Informal care, also shows a positive relationship but very weak as the point estimate is only 1.07. We can also see that the Leave-One-Out Cross Validation (LOOCV), AIC and Log Likelihood shows that the generalizable error of the model decreases for each model, confirming that these variables are indeed improving the model fit. Having said that, the estimations of the model are very uncertain. As can be seen from the preschool category, the standard error is even bigger that the preschool coefficient suggesting a lot of uncertainty. This applies equally to the informal care category and the head start category. We should not dismiss these results simply because they are very uncertain but judge their plausibility based on mechanisms and theory. Based on the description in the literature review and the hypothesis and research questions section, these point estimates lay in the direction as expected. Having said that, we should embrace uncertainty and just report them as expected but with a lot of uncertainty.

[8] In the section 4.6.1 I perform a simulation to show that these coefficient become significant by adding around 200 observations. This is done to highlight that the effect seems to be small but consistent and that significance criteria should not discard the results.

As described in the coding section, the preschool category is composed of both preschool educated children and those which went to some sort of after school care. To isolate even better the association with preschool, table 4.3 in the appendix presents exactly the same model but excludes the after school care from the model. We see virtually the same results, a sign of robustness.

There are some concerns related to the low sample size in the category of GED holders and those with no qualifications. Given the few respondents in their specific groups, it is reasonable to assume that most of the estimations should be volatile because of weak statistical power. In order to test for the reliability of the model I perform two simulations. In the first simulation I run the model 1000 times, each time randomly picking a new sample for the model (bootstrap with replacement). Since each of the 1000 models has a set of coefficients, I calculate the mean coefficient for each covariate and its variability. The final results is the mean coefficient for each of the covariates over 1000 samples, where each sample has a different sample composition. In the second simulation, I run 1000 models randomly picking *half* of the entire sample with replacement. Similar sensitivity checks are discussed and explored in statistical textbooks such as Gelman and J. Hill (2006) where they dedicate a complete chapter called 'Simulation for checking statistical procedures and model fits' where they introduce and explore ideas similar to this one.

The objective behind the first simulation is to test whether the model's coefficients are extremely sensitive to the exact composition of the sample (if the estimates were robust, they should not change). The second simulation tests if the model is highly sensitive to the sample size [9].

Figure 4.4 shows the mean coefficient for selected independent variables with their respective 95% uncertainty intervals. The Y axis refers to the variable names and the X axis refers to the coefficients. The legend refers to (1) fraction as the sample with half the observations, (2) random is the random draw of the same sample and (3) original is the original coefficients.

As we can see, the random and the original samples are virtually the same for all covariates, suggesting the coefficients do not vary when the composition of the data changes.

[9] All models are sensitive to sample size. However, if the model is unstable then halving the sample should be informative as to how much the coefficients differ

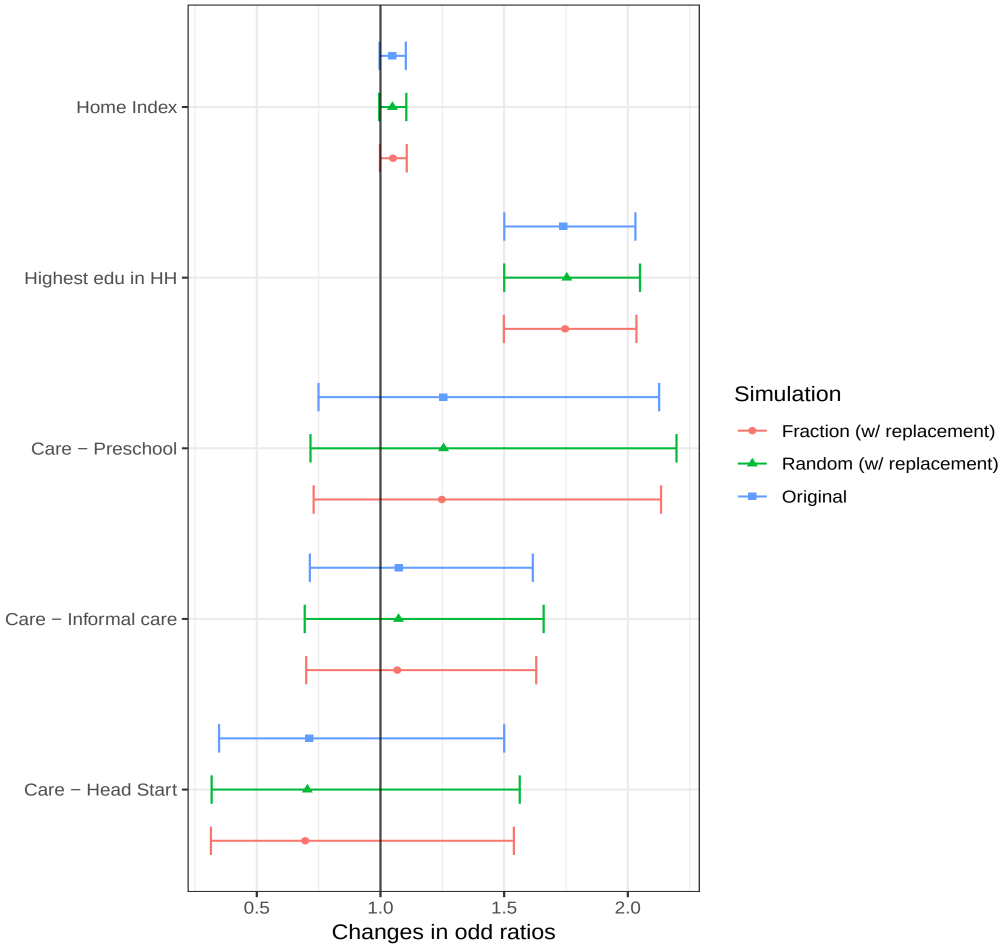


Figure 4.4: Mean coefficients after 1000 simulations for model of (1) high school vs (0) GED + No qualification

The fractional sample also shows consistency which is reassuring as the estimates do not seem to alter significantly when half the sample is removed. But do note that the the uncertainty intervals differ between samples, with the fraction sample being the most uncertain. This is important because it tells that the data seems to be sensitive to the sample size in some specific covariates like Head Start and informal care.Note as well that preschool and informal care are uncertain to the point that they include increasing as well as decreasing odds. The first hypothesis suggested that preschool would have a stronger relationship to educational attainment relative to other types of care. In other words, preschool helps increase the probabilities of graduating high school over going to a GED. The evidence here demonstrates that the relationship seems to be stable.

Table 4.4 shows the same model as before but for the dependent variable high school vs GED. This variable captures the relationship that conditional on having any qualification, the care arrangement is associated with the best qualification (high school degree).

Table 4.4: Logistic regression with dependent variable graduated high school (1) over attaining GED (0) expressed as odd ratios with robust standard errors

	Dependent variable:					
	1 = High school / 0 = GED attainment					
	(1)	(2)	(3)	(4)	(5)	(6)
Care arrangement: Head Start	0.36*** (0.13)	0.35*** (0.14)	0.35** (0.14)	0.37** (0.16)	0.40** (0.18)	0.39** (0.17)
Care arrangement: Informal care	1.23 (0.31)	1.32 (0.36)	1.28 (0.35)	1.06 (0.30)	1.15 (0.33)	1.12 (0.32)
Care arrangement: Preschool	1.70* (0.50)	1.57 (0.51)	1.49 (0.49)	1.04 (0.35)	1.15 (0.39)	1.11 (0.38)
Care reason: Child needed play		2.93* (1.63)	2.87* (1.61)	2.54 (1.46)	2.57 (1.49)	2.54 (1.47)
Care reason: Other (Ref: Started/Returned to work)		0.79 (0.25)	0.78 (0.25)	0.80 (0.26)	0.81 (0.27)	0.79 (0.26)
Care reason: Started school		0.52 (0.22)	0.54 (0.22)	0.52 (0.21)	0.54 (0.23)	0.52 (0.22)
Home index			1.05** (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)
Highest education in HH				1.68*** (0.14)	1.71*** (0.15)	1.71*** (0.16)
Race: Other race (Ref: Black)					1.97 (0.86)	2.01 (0.88)
Race: White					0.99 (0.24)	0.96 (0.24)
Number of siblings						0.89 (0.09)
Constant	10.61*** (1.93)	10.61*** (1.93)	4.02*** (2.03)	3.80** (2.11)	3.21** (1.80)	3.60** (2.03)
Number of observations	1292	1292	1292	1292	1292	1292
Log Likelihood	-344.01	-339.1	-337.65	-321	-319.3	-318.65
AIC	696.02	692.21	691.3	660.01	660.6	661.3
LOO	0.54	0.53	0.53	0.51	0.51	0.51

Note: *p<0.1; **p<0.05; ***p<0.01

At first sight the results look fairly similar to the previous model but there are indeed some differences. Note how preschool became weaker both in terms of coefficients and standard errors in model (6), meaning that an average child has 1.11 greater odds of completing high school relative to obtaining GED. In contrast to the previous model, where preschool had an increase in odds of about 1.25, the extra odds of 14% (1.25 - 1.11) can be attributed to the odds of going from no high school to high school. These models can be compared given that they hold exactly the same specification and sample while the only change is excluding those which did not graduate high school. On the contrary, informal care did see an increase in odds to 1.12, up 5% from the previous model. Finally, head starters did experience a worsening of the relationship with a much more precise estimation, as the standard error become much smaller. More specifically, head starters have an approximate 61% less odds of completing high school over a GED relative to those cared for by their parents. This also evidences that the increase in odds of graduating high school relative to the previous is attributable to those which did not graduate high school. Up until now, it seems that the relationship between care arrangement and achieving high school seems to be much stronger for the comparison between high school vs no high school than for the comparison high school vs GED. However, the results are still positive although more weak than the previous model.

It is important also to highlight that from the care reason variable, the category for child needed play is the one with the strongest relationship in the two previous models. This is important because as described in the coding section, this is the only category which actually denotes some sort of interest towards the development of soft skills rather that the reason being a necessity to find a job or go back to school. More concretely, the odds ratios for the first model was 1.46 whereas the odds ratio for the second model is 2.54. Although both carry great deal of uncertainty, they point in the expected direction in both models.

It is important to highlight that informal care and preschool seem to have similar associations in the models that refer to the competing odds between achieving a high school degree or attaining a GED. This is puzzling as it suggests that for those who have any degree, informal care is as strongly associated to achieving a high school degree as it is to being in preschool.

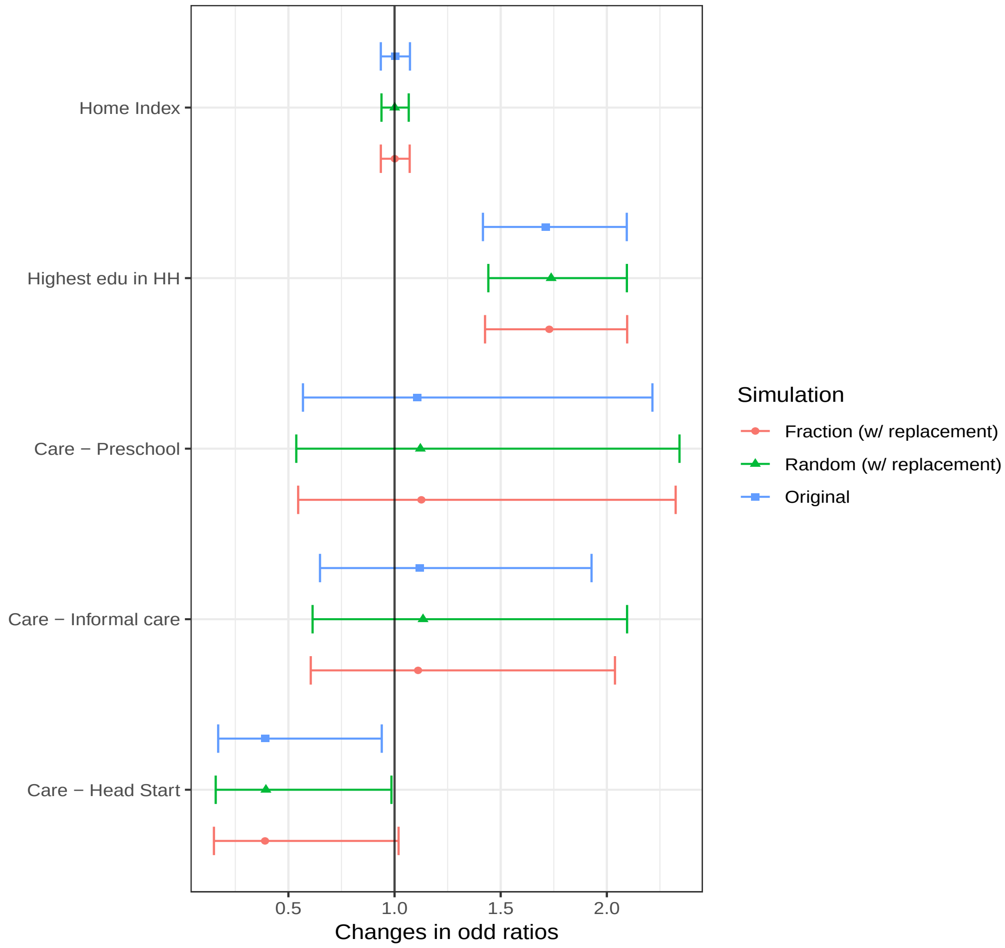


Figure 4.5: Mean coefficients after 1000 simulations for model of (1) high school vs (0) GED

Similar to the previous model, Table 4.4 in the appendix shows the same model but excluding the after school group from the preschool category. The results hold and only vary by 1-2% odds for all covariates.

Figure 4.5 shows the same simulation as before. As the sample size decreases in each model, these graphs become more informative as they can pinpoint specific covariates which would be too uncertain to even interpreted.

The estimates show stability across all models yet the uncertainty grew as the intervals became slightly larger across all simulations. However, it is not clear whether it was because of their substantive value became weaker or simply because the total sample size decreased.

Table 4.5 shows the same models as before but for the dependent variable GED vs no qualification. Relative to the previous two dependent variables, this one attempts to capture the relationship that care arrangement is associated with getting *at least any* qualification (GED) over no qualification. Note that the sample size for this model is substantially lower than for the two previous models, as it excludes those who graduated high school, the vast majority of the sample.

Table 4.5: Logistic regression with dependent variable graduated GED (1) over not having any qualification (0) expressed as odd ratios with robust standard errors

	Dependent variable:					
	1 = GED / 0 = No qualification					
	(1)	(2)	(3)	(4)	(5)	(6)
Care arrangement: Head Start	3.92** (2.42)	6.25** (4.65)	6.12** (4.52)	5.89** (4.46)	6.59** (5.11)	6.71** (5.12)
Care arrangement: Informal care	1.12 (0.39)	1.15 (0.41)	1.11 (0.41)	1.06 (0.41)	0.97 (0.37)	0.90 (0.36)
Care arrangement: Preschool	1.25 (0.50)	1.60 (0.77)	1.69 (0.81)	1.61 (0.79)	1.67 (0.84)	1.65 (0.81)
Care reason: Child needed play		0.26 (0.22)	0.21* (0.17)	0.20** (0.16)	0.16** (0.13)	0.14** (0.11)
Care reason: Other (Ref: Started/Returned to work)		0.81 (0.43)	0.72 (0.40)	0.73 (0.41)	0.73 (0.42)	0.65 (0.36)
Care reason: Started school		0.99 (0.58)	1.06 (0.62)	1.08 (0.65)	1.14 (0.70)	1.04 (0.64)
Home index			1.15*** (0.06)	1.14*** (0.06)	1.10* (0.06)	1.10* (0.06)
Highest education in HH				1.10 (0.20)	1.10 (0.19)	1.12 (0.19)
Race: Other race (Ref: Black)					1.83 (1.08)	1.83 (1.09)
Race: White					2.09** (0.74)	2.00* (0.71)
Number of siblings						0.84 (0.10)
Constant	0.89 (0.21)	0.89 (0.21)	0.08*** (0.07)	0.07*** (0.07)	0.10** (0.10)	0.13** (0.13)
Number of observations	192	192	192	192	192	192
Log Likelihood	-129.98	-128.48	-124.14	-123.96	-121.7	-120.54
AIC	267.95	270.97	264.28	265.91	265.41	265.08
LOO	1.41	1.44	1.41	1.43	1.43	1.44

Note: *p<0.1; **p<0.05; ***p<0.01

The first model shows that relative to being cared for by your parents, preschool alone is associated with an increase of about 1.25% more odds of achieving a GED relative to being a dropout. Similarly, informal care also shows a positive association with an increase in odds of about 1.12%. Finally, and contrary to all previous models, it seems that Head starts do get some sort of boosting for attaining a GED relative to being a dropout, as the associated increase in odds is of 3.92. Moving on to models (2) and (3) we see that the three estimates hold their direction even after controlling for the reason why they enrolled in that care and the home index. Preschool educated children relative to those cared for by their parents have an associated increase in odds of 1.67 of achieving a GED relative to being a dropout. The Head start relationship also increase, but I presume that this category is very volatile and the odds are extremely large and change dramatically between models. We will see their stableness in the simulation plot next. Once we move on to the last model, we see that when controlling for the education in the household, race and the number of siblings. The relationship disappears for the informal care variable with an decrease in odds of 0.90 but stays stable for the preschool educated children with an increase in odds of 1.65. Finally, head start participants boost very high odds of 6.71, something which I find very implausible and it is probably the result of very few observations within that cell [10].

Figure 4.6 contains the simulations that will test the stableness of these coefficients. It should be noted that this simulation is particularly helpful for this model, as it shows how dependent are the point estimates on the specific sample size of the previous model.

The simulation shows that indeed the Headstart simulations show that whenever the sample is bootstrapped, the estimations become extremely volatile with point estimates of over 50 odds ratios. However, we also see that the preschool and informal care categories, although carrying uncertainty in their estimations, are indeed stable over the two simulations. Moreover, table 4.5 in the appendix shows the same results but excluding the before/after school group from the preschool category and the association for preschool is virtually the same. This gives greater support to the notion the preschool education seems to be related the getting any sort of qualification over being a high school dropout.

[10]It should be noted that all of these models do not suffer from multicollinearity as one might expect, given the home index variable and the highest education in the household. The highest Variance Inflation Factor (VIF) is 1.7, a small estimate compared to the standard in the literature

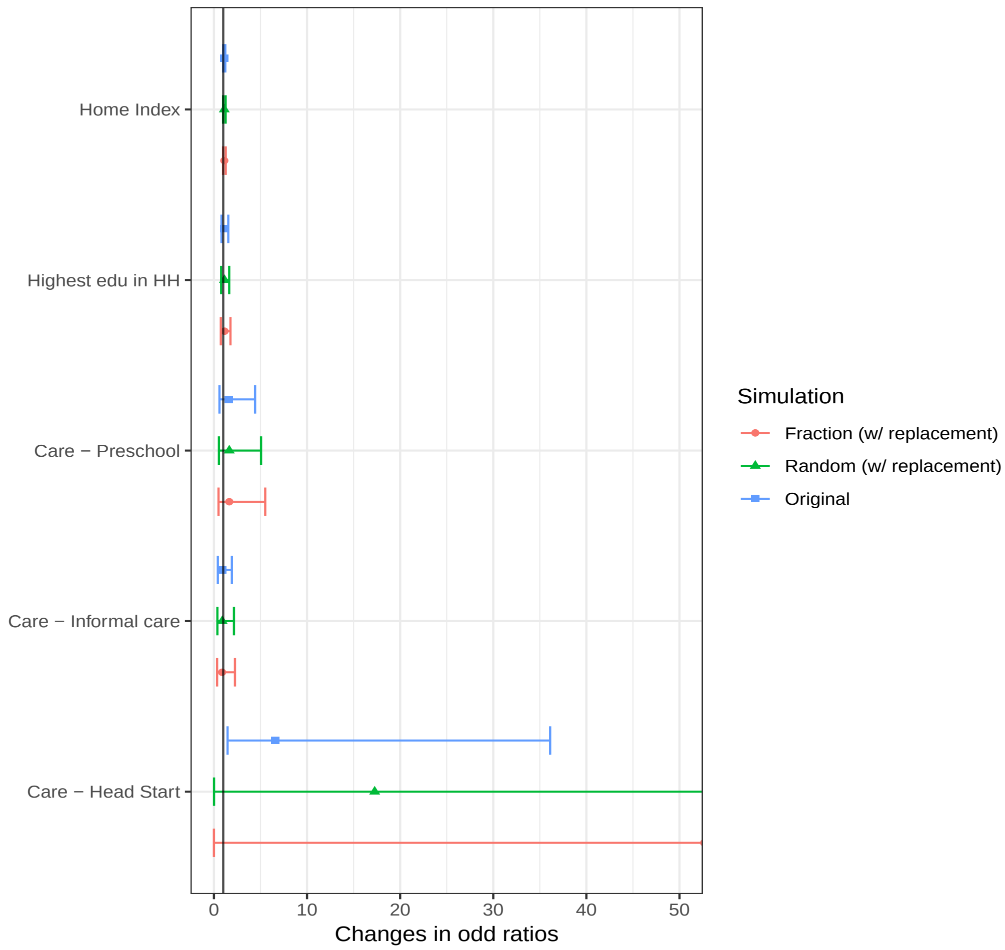


Figure 4.6: Probabilities of attaining a GED vs dropping out by care arrangement and parent's education

The next model is presented to explore whether the quality of the care is important but the coding is different. First, the preschool category is now a dummy against all other types of care (Magnuson et al., 2007). Second, a new dummy is included for whether that care was over 35 hours or less. The models are presented in table 4.6.

Table 4.6: Logistic regression with dependent variable graduated high school (1) over attaining GED + high school dropout (0) with different independent variables and robust standard errors

	Dependent variable:						
	1 = High school / 0 = GED + no qualification						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Care arrange: Preschool	1.73*** (0.34)	1.70*** (0.34)	1.55** (0.34)	1.45* (0.32)	1.15 (0.26)	1.21 (0.27)	1.18 (0.27)
Care intensity: Full time		1.12 (0.20)	1.02 (0.19)	1.06 (0.20)	1.13 (0.22)	1.22 (0.24)	1.19 (0.24)
Care reason: Child needed play			1.89* (0.72)	1.69 (0.64)	1.35 (0.52)	1.48 (0.57)	1.44 (0.56)
Care reason: Started school			0.55** (0.16)	0.58* (0.18)	0.51** (0.16)	0.60 (0.19)	0.55* (0.17)
Care reason: Started/returned to work			1.29 (0.23)	1.24 (0.22)	1.03 (0.19)	1.06 (0.20)	1.06 (0.20)
Home index				1.11*** (0.02)	1.05** (0.03)	1.04 (0.03)	1.05* (0.03)
Highest education in HH					1.74*** (0.12)	1.74*** (0.13)	1.75*** (0.13)
Race: Other races (Ref: Black)						2.82*** (0.91)	2.87*** (0.94)
Race: White						1.47** (0.28)	1.38* (0.27)
Number of siblings							0.83*** (0.05)
Constant	5.44*** (0.48)	5.29*** (0.51)	4.97*** (0.58)	0.68 (0.26)	0.58 (0.25)	0.55 (0.24)	0.67 (0.30)
Number of observations	1384	1384	1384	1384	1384	1384	1384
Log Likelihood	-552.88	-552.66	-547.3	-535.51	-501.49	-494.42	-490.36
AIC	1109.75	1111.33	1106.6	1085.03	1018.98	1008.84	1002.72
LOO	0.8	0.8	0.8	0.78	0.73	0.73	0.72

Note: *p<0.1; **p<0.05; ***p<0.01

The reference category for Preschool is all other types of care and the reference category for full-time is part-time.

These estimates are also positive for preschoolers as they alone have greater odds of completing high school relative to all other types of care. To be more precise, children who participated in preschool have on average odds of about 18% greater to finish high school than all other types of care together. This holds even after controlling for all set of covariates. In a similar vein, those who went to full time have about 19% more odds of achieving high school than part-timers. The home variable and head's education are significant and the head's education is also quite sizable. One take away from this is that when pooling informal care with parental care and Head Start, the preschool effect gains strength and highlights its importance [11]. However, we should not interpret these results as comparable to the previous models because the reference category in the care arrangement variable is different. In other words, the two models show that preschool-educated children are associated with greater chances of graduating high school relative to children cared for by their parents, and to all other types of care pooled together.

As we saw before, the head start group is the smallest among all types of care and their point estimates are very extreme in some situations. To test whether head start is actually driving the preschool estimation, Table 4.6 in the appendix runs the same model but excludes head starters from the model. The odds ratio diminish from 1.18 to 1.12, keeping still an uncertain yet positive relationship.

Table 4.7 and Table 4.8 show the same models with the preschool dummy but for the high school vs GED and GED vs no qualification dependent variables respectively.

[11] I do not include the simulations for this model but the results are the same as in the previous models – it shows robustness in both scenarios

Table 4.7: Logistic regression with dependent variable graduated high school (1) over attaining GED (0) with different independent variables and robust standard errors

	Dependent variable:						
	1 = High school / 0 = GED						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Care arrange: Preschool	1.66* (0.43)	1.72** (0.46)	1.53 (0.46)	1.48 (0.44)	1.16 (0.35)	1.20 (0.36)	1.18 (0.35)
Care intensity: Full time		0.82 (0.19)	0.72 (0.17)	0.73 (0.18)	0.78 (0.19)	0.81 (0.20)	0.80 (0.20)
Care reason: Child needed play			2.53 (1.46)	2.42 (1.39)	1.97 (1.13)	2.09 (1.20)	2.07 (1.20)
Care reason: Started school			0.48* (0.18)	0.49* (0.19)	0.44** (0.17)	0.49* (0.19)	0.47* (0.18)
Care reason: Started/returned to work			1.41 (0.33)	1.38 (0.33)	1.19 (0.29)	1.25 (0.30)	1.24 (0.30)
Home index				1.05** (0.03)	1.00 (0.03)	0.99 (0.03)	1.00 (0.03)
Highest education in HH					1.69*** (0.14)	1.73*** (0.15)	1.73*** (0.15)
Race: Other races (Ref: Black)						2.07* (0.88)	2.10* (0.90)
Race: White						1.05 (0.26)	1.02 (0.25)
Number of siblings							0.91 (0.08)
Constant	10.54*** (1.23)	11.11*** (1.47)	10.28*** (1.63)	3.77*** (1.83)	3.61** (1.96)	3.07** (1.70)	3.36** (1.86)
Number of observations	1292	1292	1292	1292	1292	1292	1292
Log Likelihood	-349.83	-349.46	-344.16	-342.57	-324.66	-322.78	-322.2
AIC	703.65	704.93	700.31	699.14	665.31	665.57	666.4
LOO	0.54	0.54	0.54	0.54	0.51	0.51	0.51

Note: *p<0.1; **p<0.05; ***p<0.01

The reference category for Preschool is all other types of care and the reference category for full-time is part-time.

Table 4.8: Logistic regression with dependent variable graduated GED (1) over not having any qualification (0) expressed as odd ratios with different independent variables and robust standard errors

	Dependent variable:						
	1 = GED / 0 = No qualification						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Care arrange: Preschool	1.10 (0.41)	1.00 (0.38)	1.03 (0.43)	1.10 (0.46)	1.08 (0.46)	1.14 (0.49)	1.11 (0.47)
Care intensity: Full time		2.02^{**} (0.69)	2.49^{**} (1.04)	2.61^{**} (1.12)	2.62^{**} (1.14)	2.85^{**} (1.23)	2.67^{**} (1.17)
Care reason: Child needed play			0.62 (0.44)	0.48 (0.36)	0.46 (0.34)	0.36 (0.30)	0.35 (0.29)
Care reason: Started school			1.41 (0.80)	1.55 (0.88)	1.53 (0.87)	1.50 (0.87)	1.45 (0.84)
Care reason: Started/returned to work			0.70 (0.27)	0.69 (0.27)	0.66 (0.27)	0.58 (0.23)	0.60 (0.24)
Home index				1.16^{***} (0.06)	1.15^{***} (0.06)	1.11^{**} (0.06)	1.11^{**} (0.06)
Highest education in HH					1.12 (0.21)	1.12 (0.21)	1.13 (0.21)
Race: Other races (Ref: Black)						1.52 (0.86)	1.52 (0.86)
Race: White						2.07^{**} (0.75)	1.98^{*} (0.72)
Number of siblings							0.90 (0.10)
Constant	1.07 (0.17)	0.91 (0.16)	0.96 (0.20)	0.07^{***} (0.06)	0.06^{***} (0.06)	0.09^{**} (0.09)	0.11^{**} (0.10)
Number of observations	192	192	192	192	192	192	192
Log Likelihood	-132.88	-130.67	-129.64	-124.75	-124.5	-122.41	-121.9
AIC	269.76	267.34	271.28	263.49	265	264.83	265.79
LOO	1.41	1.4	1.44	1.4	1.42	1.43	1.44

Note: *p<0.1; **p<0.05; ***p<0.01

The reference category for Preschool is all other types of care and the reference category for full-time is part-time.

In both models the preschool coefficient is positive, where in the first the odds are similar to the first model (1.18 odds) and in the second it diminishes to 1.11. In both instances the preschool dummy seems to hold it's relationship. For the first model, this means that preschool educated children are associated with 1.18 greater odds of attaining the 'best' qualification, namely a high school degree over a GED or having no qualifications. The second model shifts the dependent variable for comparing those who have a high school degree vs those who have a GED diploma. The model shows that preschool educated children also have 1.18 greater odds of attaining a high school degree over a GED diploma. Finally, in the last model, we see that even when restricted to getting 'any type' of qualification, preschoolers have 1.11 greater odds of achieving a GED over having no qualification whatsoever. Table 4.7 and Table 4.8 in the appendix presents the previous two models but excluding the head start category and we see both results also hold.

Having said all of that, it should be discussed that all of these preschool estimations show a great deal of uncertainty and do not meet traditional thresholds of significance based on p values. However, as recent evidence throughout much of the scientific world has showed (Bernardi et al., 2017; Collaboration et al., 2015; Gelman, 2013; Gelman and Loken, 2013), p values should not be a binary criteria to accept or reject results in a scientific paper. Instead, the results should embrace uncertainty and show the results as is. Moreover, the acceptance of the results should not be based merely on a significance criteria but on related prior evidence, plausibility of mechanism, study design, real world costs and benefits and the novelty of the findings.

Based on all of these criteria, when we compare these results with past findings such as Deming (2009) and all the relevant research in the literature and hypothesis section, the relationship between preschool education and achieving the best qualification or *any* qualification, becomes more plausible both in relation to prior evidence but in terms of the plausibility of the mechanism. Having said that, I present all results with their corresponding uncertainty but build the reliability of the findings on several specifications and previous theoretical and empirical findings.

Having said that, there is evidence pointing to hypothesis 1, 2 and 3. Both models that have the high school vs GED + no qualification as the dependent variable show that the preschool education category has the strongest association of boosting graduating high school over any other qualification or no qualification. Moreover, the models which compare the dependent variable high school vs GED, all show that preschool educated children had greater odds of getting the best qualification for those who have any qualification. Finally, the models which compare GED vs no qualifications, although the most uncertain, also show that preschool educated children have the greatest odds of achieving any sort of qualification relative to being a high school dropout.

These three hypothesis also show robustness in the alternative specifications in the appendix. However, the three models do have a great deal of uncertainty and further research should attempt to replicate and explore these relationship with the aim of improving even further the reliability of the results.

The fourth hypothesis posited that the relationship between preschool and educational attainment would be moderated by the SES origin of the student. In other words, there would be an interaction between preschool and the education of the parents. In order to test for the interaction, Table 4.9 reruns the first model of all the models presented here but introduces an interaction term between the care arrangement and the highest education in the household. Note that this interaction is included in the model with the biggest sample size given that estimating an interaction requires much more statistical power than other main effects.

Table 4.9: Logistic regression with dependent variable graduated high school (1) over attaining GED + high school dropout (0) with care arrangement and high education interaction with robust standard errors

	Dependent variable:						
	1 = High school / 0 = GED + no qualification						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Care arrangement: Head Start	0.59^{*} (0.18)	0.62 (0.21)	0.62 (0.22)	0.63 (0.23)	0.75 (0.28)	0.71 (0.27)	2.00 (1.54)
Care arrangement: Informal care	1.30 (0.24)	1.39^{*} (0.27)	1.31 (0.26)	1.05 (0.21)	1.14 (0.24)	1.08 (0.23)	1.36 (0.55)
Care arrangement: Preschool	1.89^{***} (0.41)	1.86^{**} (0.46)	1.68^{**} (0.42)	1.16 (0.30)	1.33 (0.35)	1.27 (0.33)	1.04 (0.48)
Care reason: Child needed play		1.83 (0.68)	1.70 (0.65)	1.49 (0.59)	1.51 (0.61)	1.47 (0.59)	1.43 (0.58)
Care reason: Other (Ref: Started/Returned to work)		0.83 (0.21)	0.80 (0.21)	0.81 (0.22)	0.82 (0.22)	0.76 (0.21)	0.73 (0.20)
Care reason: Started school		0.50^{**} (0.16)	0.55^{*} (0.18)	0.53^{*} (0.17)	0.58 (0.19)	0.54^{*} (0.18)	0.54^{*} (0.18)
Home index			1.11^{***} (0.02)	1.06^{**} (0.03)	1.04^{*} (0.03)	1.05^{*} (0.03)	1.05^{*} (0.03)
Highest education in HH				1.74^{***} (0.12)	1.74^{***} (0.13)	1.74^{***} (0.13)	1.85^{***} (0.25)
Race: Other race (Ref: Black)					2.69^{***} (0.87)	2.72^{***} (0.89)	2.74^{***} (0.90)
Race: White					1.40^{*} (0.27)	1.31 (0.25)	1.29 (0.25)
Number of siblings						0.83^{***} (0.05)	0.83^{***} (0.05)
Care arrangement: Head Start * Highest edu in HH							0.60 (0.20)
Care arrangement: Informal care * Highest edu in HH							0.89 (0.15)
Care arrangement: Preschool * Highest edu in HH							1.09 (0.20)
Constant	5.00^{***} (0.65)	5.00^{***} (0.65)	0.72 (0.28)	0.62 (0.27)	0.58 (0.26)	0.71 (0.32)	0.68 (0.34)
Number of observations	1384	1384	1384	1384	1384	1384	1384
Log Likelihood	-549.19	-544.3	-533.05	-500.14	-493.97	-489.48	-487.63
AIC	1106.38	1102.6	1082.11	1018.28	1009.94	1002.97	1005.27
LOO	0.8	0.8	0.78	0.73	0.73	0.72	0.73

Note: *p<0.1; **p<0.05; ***p<0.01

Moving on to the interpretation, note the main effects for the care arrangement and the highest education in the household are not to be interpreted as usual given that they are now relative to the interaction terms. For easier interpretation of the model, we will only look at the interaction terms at the bottom of the table. The three interactions can be interpreted as the difference in the slope of the highest education in the household between the specific category in the interaction and the reference category (parental care). For Head Start * Highest education interaction, we can see that the difference in the slope is negative, suggesting that the difference in the slope between this group and parental care has an odds ratio of about 0.6. For informal care, this difference is also negative with an odds ratio of 0.89, suggesting that their slopes are indeed different from each other. Finally, the preschool slope relative to the parental care slope also seems to be slightly different with a slope of 1.09. Despite these three coefficients having some sort of difference (neither of them are at 1, so the same odds) the uncertainty of the coefficients is big as the standard errors are sometimes twice the size of the coefficients.

For easier understanding, Figure 4.7 shows the predicted probabilities of attaining a high school degree rather than GED for different care arrangements and different educational level of the parents from the first model.

The X axis represents the care arrangement while the Y axis represents the probability of graduating high school rather than obtaining a GED. The first panel (left) is for low educated heads, the second panel is middle educated and the third one is for highly educated (right). The transparent black dots in each panel shows the distribution of predicted probabilities with each education-care arrangement combination. Moreover, the red dot pinpoints the specific average predicted probability for each group. To gauge the uncertainty, each group shows the 95% credible interval based on the standard error of the predicted probabilities.

Moving on to the interpretation, in the highly educated panel there does not seem to be a strong boosting effect from going to preschool relative to other types of care but Head Start seems to have lower chances as expected. Also note that children who were cared for by their parents have the most uncertain estimates aside from Head Start. For the low educated panel, the one in which we are most interested, the preschool effect does not seem to have higher probabilities than other care arrangements (except Head Start). One

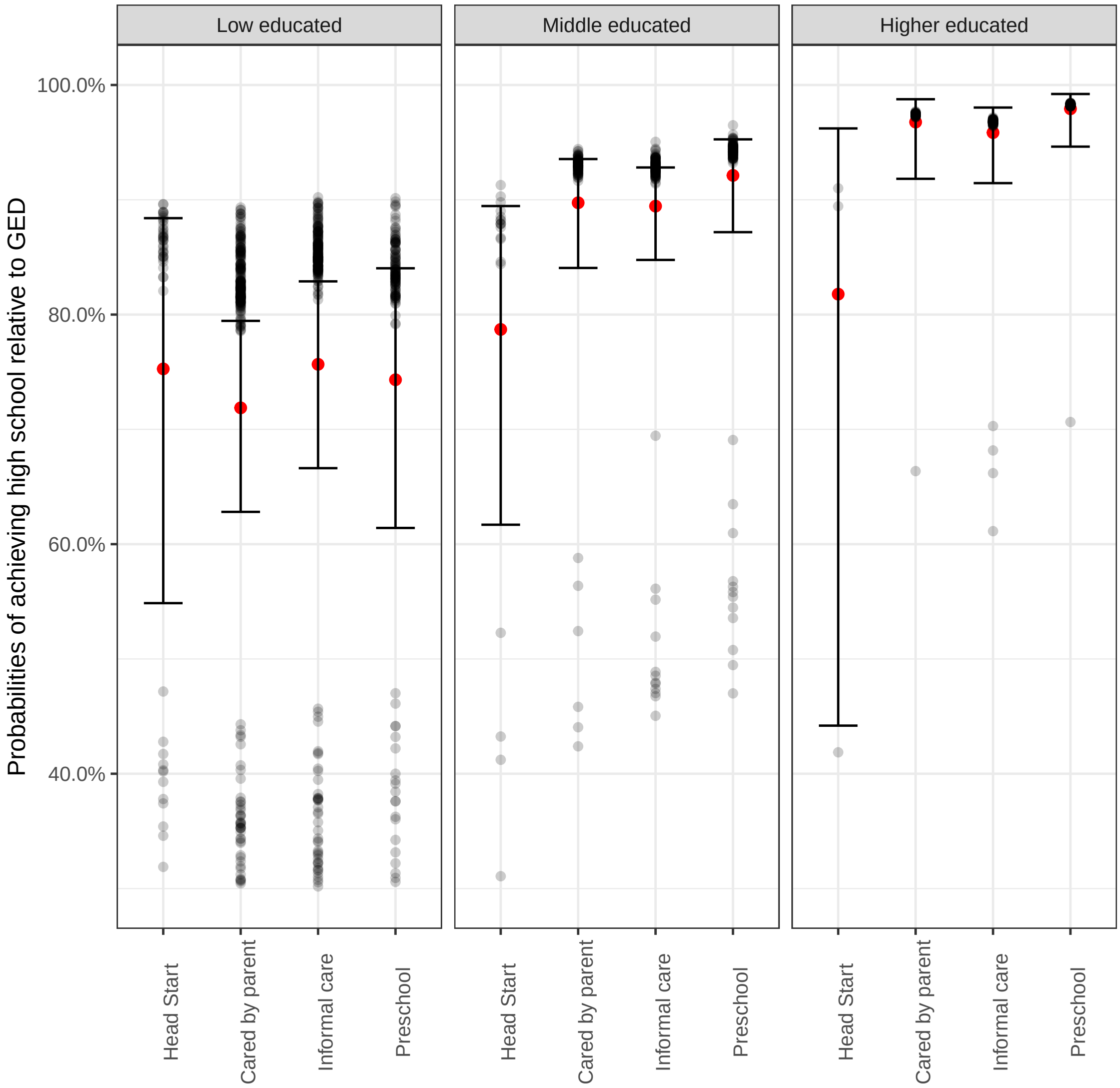


Figure 4.7: Probabilites of attaining a high school degree by different care arrangements and parent's education

pattern that gives credibility to these estimates is how each panel after the low educated has slightly higher probabilities than the previous panel, something to be expected. The fourth hypothesis posited that there would be an interaction between these two. The results point out that there seems to be some evidence albeit very weak; it seems that the fourth hypothesis is unclear as lower educated groups do not benefit more.

All of the models above show that in all instances preschool care arrangements seem to be positively related to higher qualifications relative to lower or null qualifications, although more strongly in some than in others. These results provide evidence in favor of the first three hypothesis even with alternative specifications. However, the fourth hypothesis, whether low educated children benefited more the care arrangement, the evidence is much weaker and there doesn't seem to be a strong relationship. That is, children low educated settings did not benefit more than other SES groups, at least with the present specification and uncertainty. Next we move to the sensitivity section, which tests whether this preschool association has potential to be stable in other settings.

4.6.1 Sensitivity analysis

As I have cautioned several times above, because some of the categories have too few observations, together with measurement error, the models might be lacking strength to capture statistically significant differences. For this reason I have included a series of sensitivity analysis that test the robustness of the results. More concretely, the tests look to measure how strong and generalizable they are under different scenarios.

The first analysis conducts some simulations following the work of Gelman and Carlin (2014) with Type S and Type M error. For type M (M for magnitude) I estimate the probability that the preschool effect will be significant under a hypothetical replication of the same model. For type S (S for sign) I estimate the probability that the preschool estimate is in the wrong direction, that is whether the true effect is negative rather than positive. Because these probabilities are calculated using the model's sample size and the standard error of the preschool coefficient, the hypothetical simulations account for the uncertainty of the current models. For an explanation of the mathematical proof and the specific R libraries that estimate the probabilities, see Gelman and Carlin (2014)

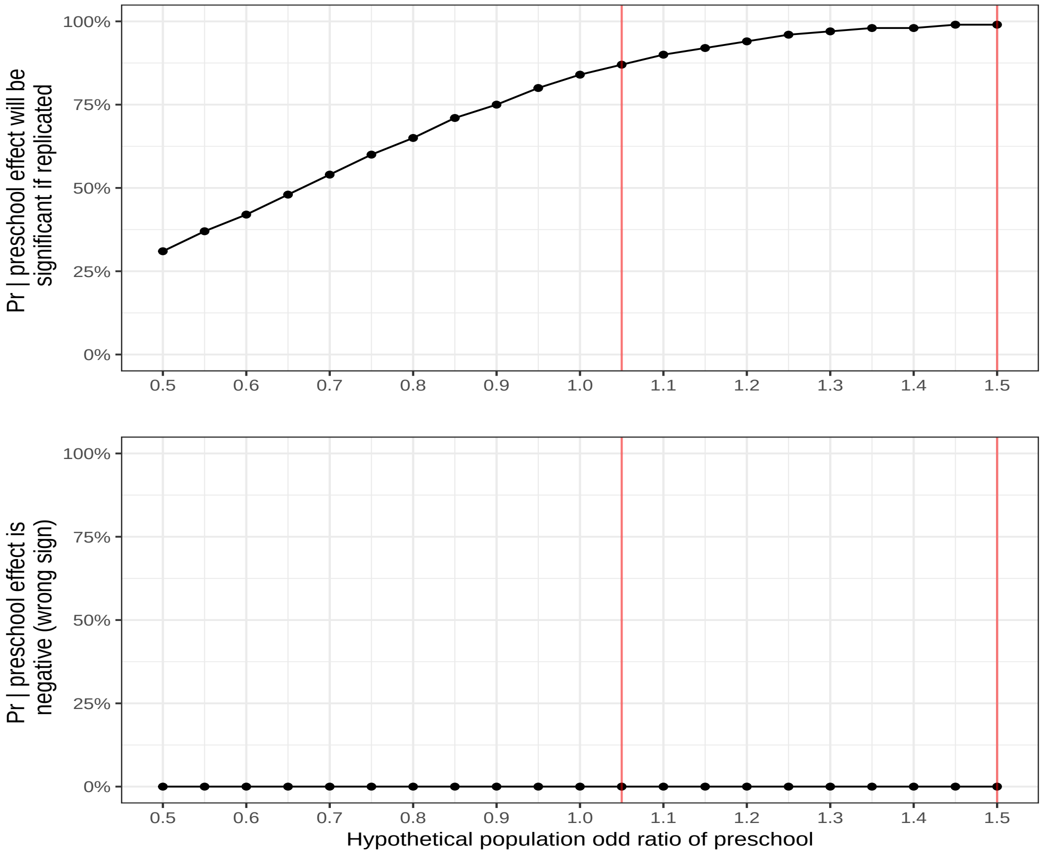


Figure 4.8: Sensitivity checks for Type M and Type S error

The probabilities discussed above are computed for hypothetical population estimates of the 'true' preschool effect (Gelman and Carlin, 2014). Because I do not know this 'true' effect, I simulate the probabilities for 21 hypothetical odd ratios ranging from 0.5 (negative effect, or wrong sign in this case) to 1.5 (a sizable effect). These hypothetical odd ratios include 1, whether there is no effect at all. Figure 4.8 contains both plots.

The reader should note that the preschool estimate varied between 1.08 and 1.50 in all the models, which is why both plots highlight that region with two vertical red lines. Looking at the plot in the top, we see that the probability of achieving significance in a replication is above 90% for all our current values. But even more, the probability that the estimate is in the wrong direction is null (bottom plot). This is true for all hypothetical odd ratios for preschool.

These results provide evidence that the lack of significance in all results might be due to the low statistical power of the model. This low statistical power is inevitable given the few respondents that are in the GED category. It is important to highlight that there is no other data set that allows us to follow respondents in over 20 years that has this rich source of personal information. For this reason it is imperative to exploit this dataset as much as possible without indulging in faulty statistical practice. Taking into consideration the recent evidence from Bernardi et al. (2017) which shows the misuse of statistical significance thresholds in declaring scientific findings, these results are important because of their substantive importance rather than for arbitrary thresholds. Considering that and the fact that the coefficient seems to be robust, it is safe to assume the strength and direction as indicative.

Finally, the last test measures how many observations the models needs to attain significance. Any model in which we add new observations will eventually reach significance, so the interpretation of this test should be very careful. I do not suggest that we artificially add N observations to the model to attain significance, as this is a pure statistical artifact and bad statistical practice. The strength and substantive meaning of the coefficient together with the previous simulations establish some confidence in the relationship. However, to give even more strength to the evidence of the relationship I perform the simulation below.

The objective behind this simulation is to gauge if the model needs a lot, or simply a few more observations to turn the results significant; this test should simply be informative and indicative of the strength of the results if the model needs very few observations to attain a coefficient statistically different from 1 (because they are expressed as odd ratios).

The simulation is as follows: sample 10 random GED holders from the data and re-add them to the data, increasing the sample size by 10. Record the new p-value and unless it is below a significance level (0.05 in this case), continue adding 10 observations. When the model reaches significance, stop. Repeat the previous loop 500 times and calculate the average number of observations needed to attain significance in each of the N iterations.

The X axis of figure 4.9 shows the number of times the significance loop was repeated. Remember that each time the significance loop has ended, it means the preschool effect is significant. Each number in the X axis means that the significant loop was repeated N times and the median number of observations was recorded. By recursively doing this I can assess how reliable the number of observations is.

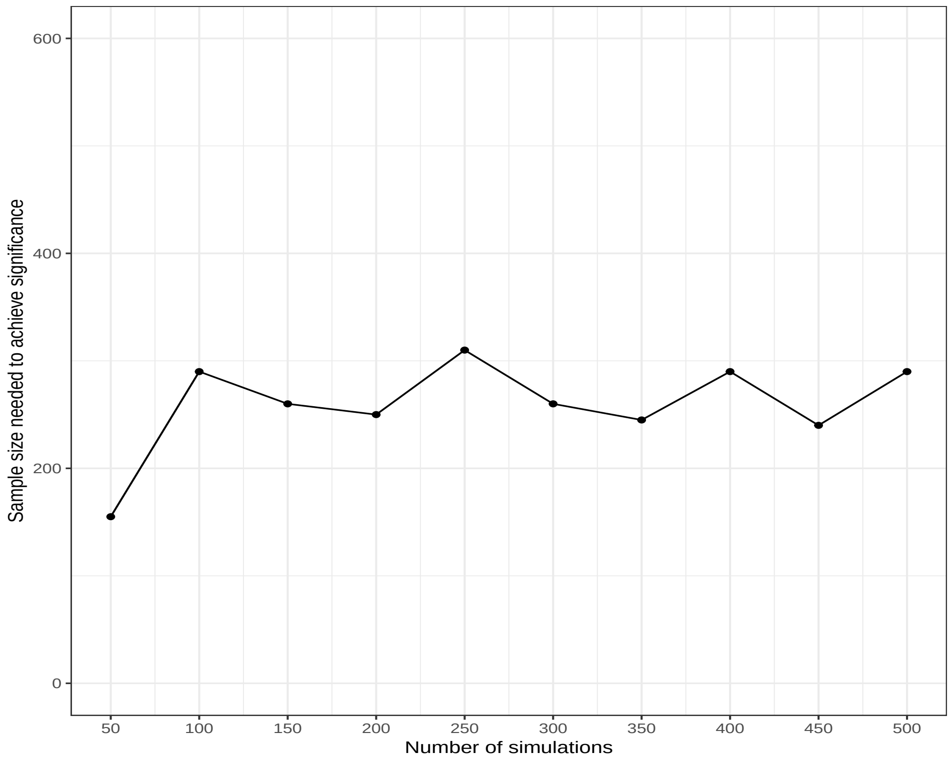


Figure 4.9: Number of observations needed to achieve significance for different simulations

The plot shows that, on average, the line is flat. Meaning that there is not much variability between simulations. On top of this, the average number of observations from all the simulations is 260, quite a small number. This final test gives some robust evidence that the first three models need only a fairly low number of observations to reach the accepted statistical significance in the literature. All of the sensitivity tests presented here reinforce the notion that the results have some meaning and are not just random noise.

4.7 Limitations

As much as these results make sense and pass most of the robustness checks, the study still has limitations. Studies that attempt to link brief activities in the early ages to other activities 20 years later should be in almost all scenarios be interpreted as correlational. No result from all of the above can be interpreted as causal effects as there are major risks of endogeneity in most of the variables in the models, as well as on unobserved variables such as parental effort in their children's upbringing. For that reason, the results should be strictly interpreted as indicative of a relationship rather than in a cause-effect direction.

In a similar line, I have tried to control for strong proxies of self-selection such as the education of the parents and the cultural environment of the household. Moreover, this chapter controls for the reason why the child enrolled in a certain care, a source of self-selection which is hardly controlled for. Additionally, it also introduces the reason why they enrolled in that care, something not done on previous studies such as Elango et al. (2015) and Campbell et al. (2002). However, there are other possible sources of information that could self-select children into better care. For example, a working class family that is overly concerned with their children's education could make an effort to put their child in preschool instead of the typical care associated with their socio-economic status. This is a positive behavior that is uncaptured by the current models. All in all, note that all of these results suffer from possible self-selection although the models attempt to adjust for the most known sources of self-selection such as the learning mileau of the household and the reason why they enrolled in certain types of care.

Another concern is the time span in which children might request a GED examination. The PSID team begun interviewing children in 1997 and followed up several times until 2014-2015 when the last available wave was conducted. This time frame means that parents with children who were 2 year old in 1997 would be 18-19 year olds in 2014-2015. Similarly, children who were 5 year olds in 1997 would be around 22 in 2014-2015. Children who were between 16 and 35 year olds in 2014-2015 have enough time to have registered for the GED tests as test takers can already take the test at age 16 with parental permission. Moreover, this becomes even more plausible for the specific sample in this paper considering that around 72% of GED test takers are between ages 16 and 29. Specifically, 22% of test takers are between ages 16 and 19. This figure has been fairly stable for the last decade as can be seen from the official statistics from the National Center for Education Statistics extracted from `https://nces.ed.gov/programs/digest/d15/tables/dt15_219.60.asp`. As can be seen from figure 4.1 in the appendix, the age distribution of the children in the 1997 sample and shows that the age structure is not concentrated on the early years but rather on ages between 8 and 12. This means that there is a reasonable number of respondents in an age span that will allow to capture respondents who had enough time to take the test. However, there is the limitation that some of these respondents will take the test later on in their lives. Considering the evidence presented here, I find it unlikely.

Finally, all of the results presented above contain a great deal of uncertainty and measurement error due to the low sample size and noisy measurements. For example, asking respondents about the care arrangements of their children retrospectively increases the risks of recall error and imprecision. However, given that this data is perhaps the among the few resources to study these types of questions, it is important to exploit it as much as possible while taking care of the interpretation and statistical tests. I have tried to find solutions to these errors by providing simulations under which the results hold similar values as in the initial models however these are not perfect. Efforts should be directed towards a replication of these results under more rigorous conditions and test whether similar specifications hold.

4.8 Conclusion

This chapter attempts to establish a relationship between early education and following particular educational trajectories later in their career. Previous research on the benefits of attending high quality early education has established that participating in quality care in the first few years previous to starting school is associated with improved chances of attaining a high school degree and a university degree. Moreover, it also links participating in early education to a reduction in potential criminal activities, improved health outcomes and higher income. Most of these benefits are overly effective on children coming from disadvantaged backgrounds relative to their better-off peers.

Despite these findings, most of the research concerned with educational attainment in adulthood is broad in terms of coding. In nearly all studies educational attainment means graduating high school relative to not graduating while ignoring other types of attainment such as a obtaining a GED diploma.

This chapter looks to concentrate on exploring the relationship between attending preschool and completing a high school degree over having a GED diploma or having no qualifications. In a similar line, the results are also extended to include those who obtained a high school degree over a GED diploma and those which obtained a GED diploma over being a high school dropout. These three comparisons form what can be defined as competing educational trajectories: getting the best qualification over getting any or no qualification (first hypothesis), getting the best qualification for those who have any qualification (second hypothesis) and finally, getting any qualification (third hypothesis). Additionally, I test the long standing hypothesis of whether the previous relationship differs by the SES origin of the student. These interactions concentrate on whether low SES children have a 'boosting' effect relative to children from other SES backgrounds as it is standard in the literature.

The first results show that across all models being in preschool relative to being cared for by one's parents is associated with higher odds of finishing a high school degree relative to either GED or being a dropout. For example, in two different variants of the preschool category, children who participated in preschool relative to those cared for by their parents and preschoolers against all other types of care saw an increase in odds of 25% and 18% respectively. These odds were specifically for attaining a high school degree rather than attaining a GED or being a dropout. The results hold even when controlling for the reason why they enrolled in that specific care, a usual unobserved source of self-selection. Moreover, these results are robust when excluding the population of those who completed head start and those which went to after school, as they might be driving the results upwards due to their low sample size and volatility. We also see that other types of informal cares such as being cared for by a relative is associated with an odds increase of 7% of attaining a high school degree over having a GED diploma or dropping out. This highlights the value of the preschool quality as in all cases, it had greater odds when compared to other types of informal care.

The second set of results show that across all models being in preschool relative to being cared for by one's parents is associated with higher odds of finishing a high school degree relative to attaining a GED. More specifically, children who participated in preschool relative to those cared for by their parents saw an increase in odds of 11% relative attaining a high school degree rather than attaining a GED. Moreover, even when comparing preschool against any other type of care, the odds increased to 18% showing robustness. These results highlight that nearly half of association discussed from the first model was attributable to those respondents having no qualifications. It seems that a lot of this association works through the boosting of children having no qualifications attaining a high school diploma.

The third and final set of results show that even with small sample sizes, this preschool association stands when comparing the odds of achieving a GED over being a dropout. That is, children cared for in preschool settings compared to those cared for by their parents saw an odds increase of 60% of achieving a GED over no qualifications. When comparing preschool over any other types of care, the odds diminishes to 11%, a size more comparable to the estimates of other models. The comparison between GED and non high school graduates is very uncertain given the low sample size. However, these results gain credibility, when they hold their association even when excluding head starters and those who were to after school care arrangements.

In all models presented preschool seems to have a stronger association than informal care, both in magnitude and in direction, except for the model that compares the competing odds of achieving a high school degree relative to attaining a GED diploma. This is puzzling as it suggests that for those who have any degree, informal care is as strongly associated to achieving a high school degree as it is to being in preschool. Although the preschool association is still present when comparing preschool against all other types of care, the fact that informal care or preschool are equally associated is something new. Perhaps it would be the case that informal care given by non-relatives has some sort of quality when compared to preschool. However, this is clearly not the case when comparing high school vs GED/No qualification and GED vs No qualification.

In a similar line, and investigating the 4th hypothesis, I test whether this relationship is stronger for children coming from low SES origin and find that the findings are mixed. This relationship slightly increased with each SES category suggesting that preschool did have a slight 'boosting' effect but the relationship is too weak. Further research should attempt to replicate this under more other research designs and with greater statistical power.

This last result is at odds with evidence from previous studies (J. Heckman, 2006; J. Heckman and Kautz, 2013). A possible explanation for this is weak statistical power. There is enough evidence to detect a relationship for the preschool attendants (as it is evident but very weak) but given the low sample size and the few degree of freedoms it is very difficult to capture any real relationship in an interaction. This makes sense as research shows that you need over twice the sample to estimate an interaction effect relative to a main effect (Gelman, 2015).

These results are tested under some hypothetical scenarios to confirm whether with the current uncertainty and low sample size, the coefficients are strong enough to replicate in a new study. The simulations show that the coefficients would indeed replicate with a probability between 75% and 95% for the current values of the preschool coefficient. Moreover, since some of these coefficients are not statistically significant, I simulate the number of observations needed to attain significance as an indication of the strength of the results. For over 500 simulations with a bootstrapped sample, the results show that with a mere 260 more observations, the preschool coefficient would yield statistically significant results. Note that this can be interpreted only as indicative and not as a proper statistical test. However, all of these simulations point towards the same direction which gives credibility to the estimates presented in the chapter.

This chapter attempts to establish a relationship between early education and later educational choices. This is one of the most debated areas of early education research. However, early education interventions have other features which are being criticized and need further attention. For example, we still do not know exactly what it is about early interventions that allows children – specifically children coming from lower educated background – to benefit from these programs. Theoretically, intellectually demanding environments should improve the abilities of children but the evaluations are not very clear on which factors are the ones that help children: is it the teachers? is it the school structure? is the interaction with other children? Further research should help answer these questions as it would put us closer to understanding how these interventions work.

Further research should also attempt to replicate this study under more rigorous designs when new data sources become available in the future [12]. Given that there are very few sources of data available with proper experiments, then better designs using this data might be a possibility. Moreover, using this same survey users should disentangle whether specific types of care histories are more prone to graduating high school over other types of attainment. Following the rich historical data on families from the Panel Study of Income Dynamics (PSID), researchers can track down all types of care that the child received and not only the last one before starting school. This will help to pinpoint the care trajectories that are associated with improved educational attainment rather than

[12] In late 2018 PSID will release the next follow up of the CDS panel which will allow researchers to replicate this analysis using 7-8 extra years.

just the one type of care before the beginning of formal schooling.

Bibliography

Anderson, Laurie M, Carolynne Shinn, Mindy T Fullilove, Susan C Scrimshaw, Jonathan E Fielding, Jacques Normand, Vilma G Carande-Kulis, Task Force on Community Preventive Services, et al. (2003). "The effectiveness of early childhood development programs: A systematic review." In: *American journal of preventive medicine* 24.3, pp. 32–46.

Baker, Michael, Jonathan Gruber, and Kevin Milligan (2015). *Non-cognitive deficits and young adult outcomes: The long-run impacts of a universal child care program.* Tech. rep. National Bureau of Economic Research.

Balfanz, Robert, Liza Herzog, and Douglas J Mac Iver (2007). "Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions." In: *Educational Psychologist* 42.4, pp. 223–235.

Barros, R Paes de and Rosane Mendonca (1998). "The impact of three institutional innovations in Brazilian education." In: *Organization matters: Agency problems in health and education in Latin America.*

Bauchmuller, Robert, Mette Gortz, and Astrid Wurtz Rasmussen (2014). "Long-run benefits from universal high-quality preschooling." In: *Early Childhood Research Quarterly* 29.4, pp. 457–470.

Beeghley, Leonard (2015). *Structure of social stratification in the United States.* Routledge.

Bernardi, Fabrizio, Lela Chakhaia, and Liliya Leopold (2017). "'Sing me a song with social significance': The (mis) use of statistical significance testing in European sociological research." In: *European Sociological Review* 33.1, pp. 1–15.

Bradbury, Bruce, Miles Corak, Jane Waldfogel, and Elizabeth Washbrook (2015). *Too many children left behind: The US achievement gap in comparative perspective.* Russell Sage Foundation.

Bradley, Robert H, Bettye M Caldwell, Stephen L Rock, Holly M Hamrick, and Pandia Harris (1988). "Home observation for measurement of the environment: Development of a home inventory for use with families having children 6 to 10 years old." In: *Contchapterorary Educational Psychology* 13.1, pp. 58–71.

Brunello, Giorgio and Daniele Checchi (2007). "Does school tracking affect equality of opportunity? New international evidence." In: *Economic policy* 22.52, pp. 782–861.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos (Feb. 11, 2011). *Making Schools Work: New Evidence on Accountability Reforms*. World Bank Publishing. 251 pp. ISBN: 0821386794. URL: https://www.ebook.de/de/product/14536372/barbara_bruns_deon_filmer_harry_anthony_patrinos_making_schools_work_new_evidence_on_accountability_reforms.html.

Campbell, Frances A, Craig T Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson (2002). "Early childhood education: Young adult outcomes from the Abecedarian Project." In: *Applied developmental science* 6.1, pp. 42–57.

Cascio, Elizabeth U and Diane Whitmore Schanzenbach (2013). *The impacts of expanding access to high-quality preschool education*. Tech. rep. National Bureau of Economic Research.

Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang (2016). *The Fading American Dream: Trends in Absolute Income Mobility Since 1940*. Tech. rep. National Bureau of Economic Research.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez (2014). *Where is the land of opportunity? The geography of intergenerational mobility in the United States*. Tech. rep. National Bureau of Economic Research.

Children Trends (Dec. 2016). *Children in Poverty: Indicators on Children and Youth*. Tech. rep.

Cunha, Flavio, James Heckman, Lance Lochner, and Dimitriy V Masterov (2006). "Interpreting the evidence on life cycle skill formation." In: *Handbook of the Economics of Education* 1, pp. 697–812.

Currie, Janet and Duncan Thomas (1998). *School quality and the longer-term effects of Head Start*. Tech. rep. National Bureau of Economic Research.

Deming, David (2009). "Early childhood intervention and life-cycle skill development: Evidence from Head Start." In: *American Economic Journal: Applied Economics* 1.3, pp. 111–134.

Elango, Sneha, Jorge Luis Garcia, James Heckman, and Andres Hojman (Nov. 2015). *Early Childhood Education*. Working Paper 21766. National Bureau of Economic Research. DOI: 10.3386/w21766. URL: http://www.nber.org/papers/w21766.

Esping-Andersen, Gøsta (2014). "Welfare regimes and social stratification." In: *Journal of European Social Policy*, p. 0958928714556976.

Esping-Andersen, Gøsta, Irwin Garfinkel, Wen-Jui Han, Katherine Magnuson, Sander Wagner, and Jane Waldfogel (2012). "Child care and school performance in Denmark and the United States." In: *Children and Youth Services Review* 34.3, pp. 576–589.

Garces, Eliana, Duncan Thomas, and Janet Currie (2002). "Longer-term effects of Head Start." In: *The American Economic Review* 92.4, pp. 999–1012.

Gelman, Andrew (2015). *The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective*.

Gelman, Andrew and John Carlin (2014). "Beyond power calculations assessing type s (sign) and type m (magnitude) errors." In: *Perspectives on Psychological Science* 9.6, pp. 641–651.

Gelman, Andrew and Jennifer Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Harris, Douglas N and Tim R Sass (2011). "Teacher training, teacher quality and student achievement." In: *Journal of public economics* 95.7-8, pp. 798–812.

Havnes, Tarjei and Magne Mogstad (2011). "No child left behind: Subsidized child care and children's long-run outcomes." In: *American Economic Journal: Economic Policy* 3.2, pp. 97–129.

Heckman, James (2006). "Skill formation and the economics of investing in disadvantaged children." In: *Science* 312.5782, pp. 1900–1902.

Heckman, James J and Tim Kautz (2012). "Hard evidence on soft skills." In: *Labour economics* 19.4, pp. 451–464.

Heckman, James, John Eric Humphries, and Tim Kautz (2014). *The myth of achievement tests: The GED and the role of character in American life*. University of Chicago Press.

Heckman, James and Tim Kautz (2013). *Fostering and measuring skills: Interventions that improve character and cognition*. Tech. rep. National Bureau of Economic Research.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz (2010). "The rate of return to the HighScope Perry Preschool Program." In: *Journal of public Economics* 94.1, pp. 114–128.

Heckman, James and Lakshmi K Raut (2016). "Intergenerational long-term effects of preschool-structural estimates from a discrete dynamic programming model." In: *Journal of econometrics* 191.1, pp. 164–175.

Heckman, James and Yona Rubinstein (2001). "The importance of noncognitive skills: Lessons from the GED testing program." In: *The American Economic Review* 91.2, pp. 145–149.

Ludwig, Jens and Deborah A Phillips (2007). *The benefits and costs of Head Start.* Tech. rep. National Bureau of Economic Research.

Luster, Tom and Harriette McAdoo (1996). "Family and child influences on educational attainment: A secondary analysis of the high/scope Perry Preschool data." In: *Developmental Psychology* 32.1, p. 26.

Magnuson, Katherine, Christopher Ruhm, and Jane Waldfogel (2007). "Does prekindergarten improve school preparation and performance?" In: *Economics of Education review* 26.1, pp. 33–51.

McKey, Ruth Hubbell et al. (1985). "The Impact of Head Start on Children, Families and Communities. Final Report of the Head Start Evaluation, Synthesis and Utilization Project." In:

Melhuish, Edward C (2011). "Preschool matters." In: *Science* 333.6040, pp. 299–300.

Moore, Kristin A, Zakia Redd, Mary Burkhauser, Kassim Mbwana, and Ashleigh Collins (2002). *Children in poverty: Trends, consequences and policy options.* Child Trends Washington.

Ou, Suh-Ruu (2005). "Pathways of long-term effects of an early intervention program on educational attainment: Findings from the Chicago longitudinal study." In: *Journal of applied developmental psychology* 26.5, pp. 578–611.

Reardon, Sean F (2011). "The widening academic achievement gap between the rich and the poor: New evidence and possible explanations." In: *Whither opportunity*, pp. 91–116.

Reynolds, Arthur J, Judy A tchapterle, Suh-Ruu Ou, Irma A Arteaga, and Barry AB White (2011). "School-based early childhood education and age-28 well-being: Effects by timing, dosage, and subgroups." In: *Science* 333.6040, pp. 360–364.

The White House (2014). *Fact sheet: Invest in US: The White House Summit on Early Childhood Education.* Technical report. Office of the Press Secretary, Washington, DC.

Van de Werfhorst, Herman G and Jonathan JB Mijs (2010a). “Achievement inequality and the institutional structure of educational systems: A comparative perspective.” In: *Annual review of sociology* 36, pp. 407–428.

Waldfogel, Jane (2006). “What do children need?” In: *Juncture* 13.1, pp. 26–34.

4.9 Appendix

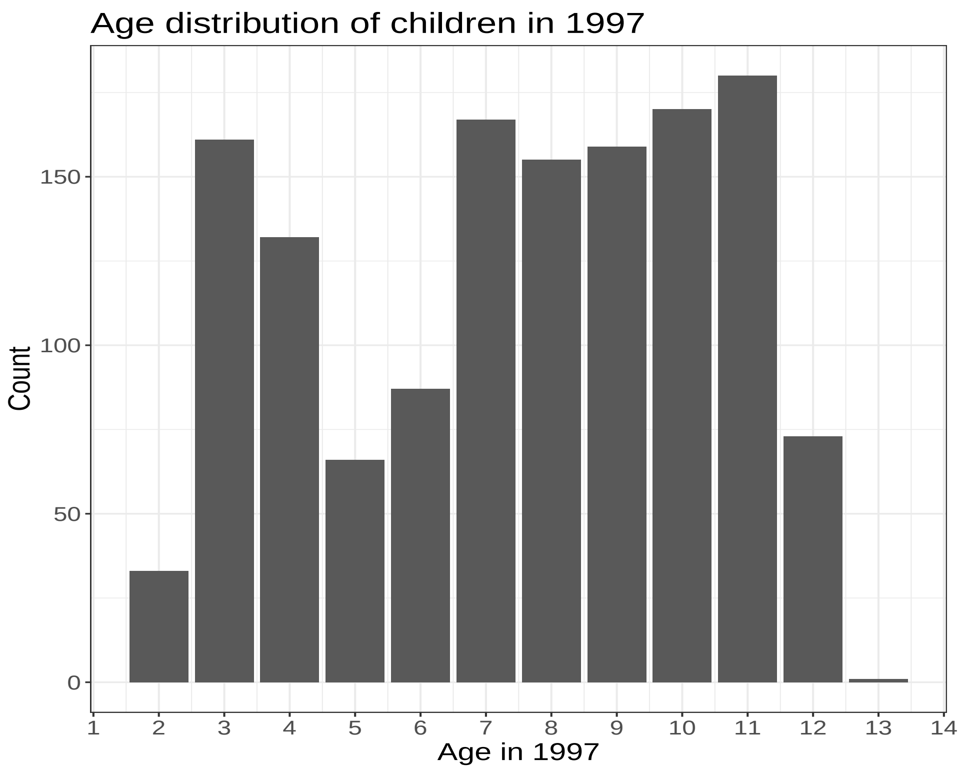


Figure 4.1: Age distribution of children in 1997

Care Arrangement	< High school	GED or above	High school or above
Cared by parent	37	33	350
Head Start	4	14	53
Informal care	34	33	428
Preschool	18	20	360

Table 4.1: Sample size of respondent's within the care arrangement they received in early childhood by the respondent's education in 2014-2015

	Non-missing		Missing	
Variables	Count	Percentage	Count	Percentage
Care arrangements				
- Cared by parent	412	30.0%	385	38.0%
- Head Start	68	5.0%	17	2.0%
- Informal care	487	36.0%	380	37.0%
- Preschool	392	29.0%	238	23.0%
Highest education in HH				
- > Bachelors	326	24.0%	234	23.0%
- Bachelors degree	76	6.0%	56	5.0%
- High school	404	30.0%	307	30.0%
- Less than HS	263	19.0%	213	21.0%
- Some college or 2 year college	290	21.0%	227	22.0%
Care reason				
- Started/Return work	578	61.0%	424	68.0%
- Child needed playmates	139	15.0%	69	11.0%
- Other	153	16.0%	86	14.0%
- Started school	77	8.0%	45	7.0%
Race of head				
- Black	560	41.0%	367	36.0%
- Other	142	10.0%	122	12.0%
- White	657	48.0%	544	53.0%
# of Siblings				
- 0	257	19.0%	297	29.0%
- 1	642	47.0%	419	41.0%
- 2	315	23.0%	215	21.0%
- 3	90	7.0%	47	5.0%
- 4	30	2.0%	20	2.0%
- 5	18	1.0%	15	1.0%
- 6	0	0.0%	1	0.0%
- 7	2	0.0%	0	0.0%
- 8	0	0.0%	0	0.0%
- 9	0	0.0%	1	0.0%
- 10	5	0.0%	3	0.0%

Table 4.2: Composition of independent variables between missing and non-missing values in educational attainment in 2014-2015. This tables attemps to find any differences in self-selection.

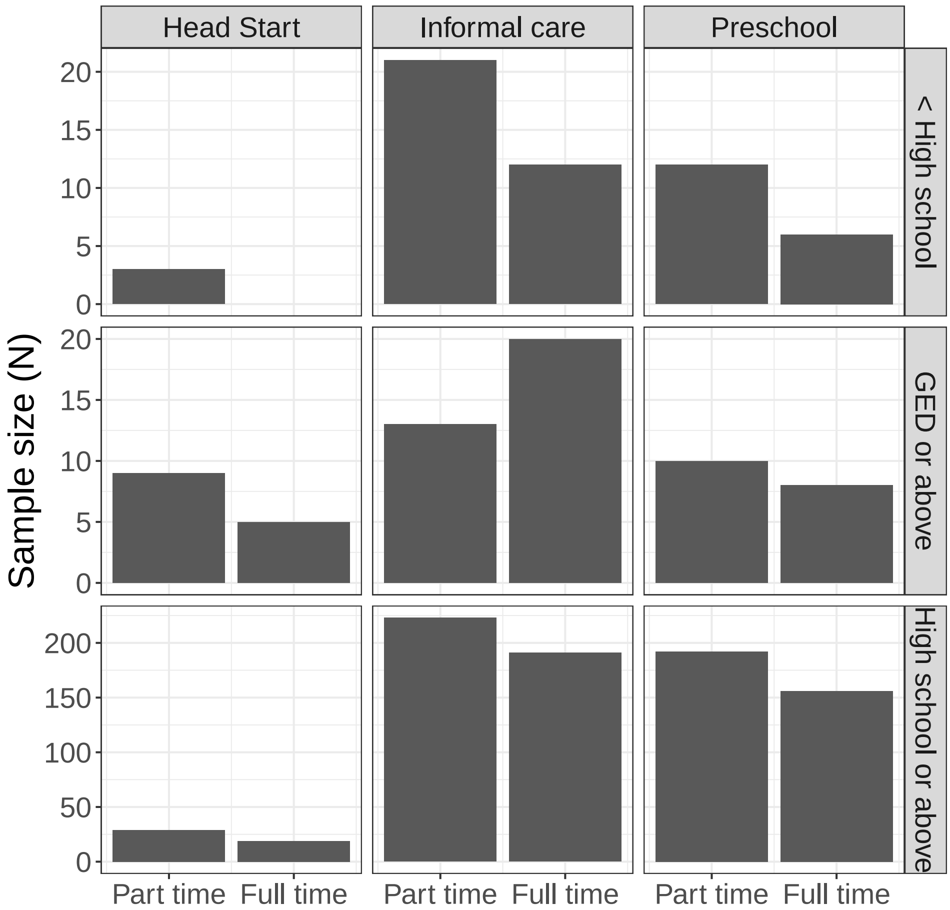


Figure 4.2: Type of care by intensity and respondent's education (raw sample size)

Table 4.3: Logistic regression with dependent variable graduated high school (1) over attaining GED + high school dropout (0) expressed as odd ratios with robust standard errors. The preschool category excludes before/after school

	Dependent variable:					
	1 = High school / 0 = GED + no qualification					
	(1)	(2)	(3)	(4)	(5)	(6)
Care arrangement: Head Start	0.59* (0.18)	0.64 (0.22)	0.64 (0.22)	0.66 (0.24)	0.78 (0.29)	0.74 (0.28)
Care arrangement: Informal care	1.30 (0.24)	1.41* (0.28)	1.33 (0.26)	1.05 (0.22)	1.15 (0.24)	1.09 (0.23)
Care arrangement: Preschool	1.89*** (0.41)	1.88** (0.46)	1.71** (0.43)	1.20 (0.31)	1.37 (0.36)	1.29 (0.34)
Care reason: Child needed play		1.76 (0.66)	1.64 (0.63)	1.38 (0.55)	1.40 (0.57)	1.37 (0.55)
Care reason: Other (Ref: Started/Returned to work)		0.80 (0.20)	0.77 (0.20)	0.78 (0.21)	0.78 (0.22)	0.75 (0.21)
Care reason: Started school		0.49** (0.16)	0.53* (0.17)	0.52** (0.17)	0.57* (0.19)	0.53* (0.18)
Home index			1.11*** (0.02)	1.06** (0.03)	1.04 (0.03)	1.05* (0.03)
Highest education in HH				1.74*** (0.12)	1.74*** (0.13)	1.74*** (0.13)
Race: Other race (Ref: Black)					2.70*** (0.88)	2.74*** (0.90)
Race: White					1.43* (0.27)	1.36 (0.26)
Number of siblings						0.84** (0.06)
Constant	5.00*** (0.65)	5.00*** (0.65)	0.72 (0.28)	0.61 (0.27)	0.58 (0.26)	0.71 (0.32)
Number of observations	1372	1372	1372	1372	1372	1372
Log Likelihood	-545.74	-540.78	-529.54	-496.68	-490.36	-487.49
AIC	1099.47	1095.55	1075.09	1011.35	1002.72	998.97
LOO	0.8	0.8	0.78	0.74	0.73	0.73

Note: *p<0.1; **p<0.05; ***p<0.01

The preschool category excludes before/after school.

Table 4.4: Logistic regression with dependent variable graduated high school (1) over attaining GED (0) expressed as odd ratios with robust standard errors. The preschool category excludes before/after school

	Dependent variable:					
	1 = High school / 0 = GED					
	(1)	(2)	(3)	(4)	(5)	(6)
Care arrangement: Head Start	0.36*** (0.13)	0.36** (0.14)	0.36** (0.14)	0.38** (0.16)	0.41** (0.18)	0.40** (0.17)
Care arrangement: Informal care	1.23 (0.31)	1.32 (0.36)	1.29 (0.35)	1.06 (0.30)	1.16 (0.33)	1.12 (0.33)
Care arrangement: Preschool	1.65* (0.48)	1.53 (0.49)	1.46 (0.48)	1.02 (0.34)	1.13 (0.38)	1.09 (0.37)
Care reason: Child needed play		2.92* (1.62)	2.86* (1.60)	2.46 (1.41)	2.50 (1.44)	2.46 (1.43)
Care reason: Other (Ref: Started/Returned to work)		0.78 (0.24)	0.77 (0.24)	0.79 (0.26)	0.79 (0.26)	0.78 (0.26)
Care reason: Started school		0.51 (0.21)	0.54 (0.22)	0.51 (0.21)	0.54 (0.23)	0.52 (0.22)
Home index			1.05* (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)
Highest education in HH				1.68*** (0.14)	1.71*** (0.16)	1.71*** (0.16)
Race: Other race (Ref: Black)					1.97 (0.86)	2.01 (0.88)
Race: White					0.99 (0.25)	0.96 (0.24)
Number of siblings						0.89 (0.09)
Constant	10.61*** (1.93)	10.61*** (1.93)	4.12*** (2.08)	3.86** (2.15)	3.27** (1.84)	3.66** (2.07)
Number of observations	1281	1281	1281	1281	1281	1281
Log Likelihood	-343.41	-338.44	-337.06	-320.52	-318.84	-318.2
AIC	694.82	690.88	690.12	659.05	659.67	660.41
LOO	0.54	0.54	0.53	0.51	0.51	0.51

Note: *p<0.1; **p<0.05; ***p<0.01
The preschool category excludes before/after school.

Table 4.5: Logistic regression with dependent variable GED (1) over high school dropout (0) expressed as odd ratios with robust standard errors. The preschool category excludes before/after school

	Dependent variable:					
	1 = GED / 0 = No qualification					
	(1)	(2)	(3)	(4)	(5)	(6)
Care arrangement: Head Start	3.92** (2.42)	6.67** (5.00)	6.68** (4.96)	6.42** (4.90)	7.30** (5.75)	7.20** (5.58)
Care arrangement: Informal care	1.12 (0.39)	1.16 (0.42)	1.13 (0.42)	1.07 (0.42)	0.97 (0.38)	0.92 (0.36)
Care arrangement: Preschool	1.32 (0.54)	1.78 (0.87)	1.95 (0.94)	1.85 (0.92)	1.95 (0.98)	1.87 (0.94)
Care reason: Child needed play		0.24* (0.20)	0.18** (0.15)	0.18** (0.14)	0.14** (0.11)	0.13** (0.10)
Care reason: Other (Ref: Started/Returned to work)		0.76 (0.41)	0.66 (0.37)	0.67 (0.38)	0.66 (0.38)	0.63 (0.36)
Care reason: Started school		0.93 (0.55)	0.99 (0.58)	1.01 (0.61)	1.06 (0.66)	1.02 (0.63)
Home index			1.16*** (0.06)	1.15*** (0.06)	1.11** (0.06)	1.11* (0.06)
Highest education in HH				1.12 (0.20)	1.12 (0.19)	1.13 (0.19)
Race: Other race (Ref: Black)					1.86 (1.10)	1.87 (1.11)
Race: White					2.23** (0.80)	2.15** (0.78)
Number of siblings						0.89 (0.12)
Constant	0.89 (0.21)	0.89 (0.21)	0.07*** (0.06)	0.06*** (0.06)	0.09** (0.08)	0.11** (0.11)
Number of observations	191	191	191	191	191	191
Log Likelihood	-129.21	-127.57	-122.8	-122.55	-119.97	-119.65
AIC	266.43	269.13	261.59	263.11	261.94	263.3
LOO	1.41	1.44	1.4	1.42	1.42	1.44

Note: *p<0.1; **p<0.05; ***p<0.01
The preschool category excludes before/after school.

Table 4.6: Logistic regression with dependent variable graduated high school (1) over attaining GED + high school dropout (0) with different independent variables and robust standard errors

	Dependent variable:						
	1 = High school / 0 = GED + no qualification						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Care arrange: Preschool	1.64^{**} (0.32)	1.60^{**} (0.32)	1.39 (0.31)	1.32 (0.30)	1.08 (0.26)	1.15 (0.27)	1.12 (0.27)
Care intensity: Full time		1.12 (0.21)	1.05 (0.21)	1.10 (0.22)	1.18 (0.25)	1.25 (0.27)	1.21 (0.26)
Care reason: Child needed play			2.55^{**} (1.18)	2.12 (0.99)	1.53 (0.74)	1.57 (0.77)	1.56 (0.76)
Care reason: Started school			0.53^{*} (0.18)	0.56^{*} (0.19)	0.48^{**} (0.17)	0.55^{*} (0.20)	0.50^{*} (0.18)
Care reason: Started/returned to work			1.20 (0.22)	1.14 (0.21)	0.93 (0.18)	0.96 (0.19)	0.96 (0.19)
Home index				1.12^{***} (0.02)	1.06^{**} (0.03)	1.04^{*} (0.03)	1.05^{*} (0.03)
Highest education in HH					1.78^{***} (0.13)	1.77^{***} (0.14)	1.78^{***} (0.14)
Race: Other races (Ref: Black)						2.62^{***} (0.85)	2.65^{***} (0.88)
Race: White						1.56^{**} (0.31)	1.46^{*} (0.29)
Number of siblings							0.83^{***} (0.05)
Constant	5.77^{***} (0.53)	5.62^{***} (0.57)	5.39^{***} (0.66)	0.64 (0.25)	0.55 (0.24)	0.53 (0.24)	0.64 (0.30)
Number of observations	1313	1313	1313	1313	1313	1313	1313
Log Likelihood	-510.21	-510.02	-504.73	-491.97	-458.15	-451.78	-447.92
AIC	1024.41	1026.04	1021.46	997.93	932.3	923.55	917.83
LOO	0.78	0.78	0.78	0.76	0.71	0.7	0.7

Note: *p<0.1; **p<0.05; ***p<0.01

The reference category for Preschool is all other types of care except Head start

Table 4.7: Logistic regression with dependent variable graduated high school (1) over attaining GED (0) with different independent variables and robust standard errors

	Dependent variable:						
	1 = High school / 0 = GED						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Care arrange: Preschool	1.65** (0.33)	1.61** (0.33)	1.32 (0.30)	1.27 (0.29)	1.02 (0.25)	1.10 (0.27)	1.08 (0.27)
Care intensity: Full time		1.13 (0.22)	1.04 (0.21)	1.08 (0.22)	1.13 (0.24)	1.20 (0.27)	1.16 (0.26)
Care reason: Child needed play			3.26** (1.63)	2.74** (1.38)	2.04 (1.06)	2.09 (1.09)	2.05 (1.07)
Care reason: Started school			0.57 (0.20)	0.61 (0.22)	0.53* (0.20)	0.61 (0.23)	0.55 (0.21)
Care reason: Started/returned to work			1.28 (0.24)	1.23 (0.23)	1.01 (0.20)	1.04 (0.21)	1.04 (0.21)
Home index				1.12*** (0.02)	1.06** (0.03)	1.04 (0.03)	1.04 (0.03)
Highest education in HH					1.78*** (0.13)	1.76*** (0.14)	1.76*** (0.14)
Race: Other races (Ref: Black)						2.66*** (0.90)	2.70*** (0.93)
Race: White						1.59** (0.32)	1.48* (0.31)
Number of siblings							0.83*** (0.05)
Constant	5.61*** (0.53)	5.46*** (0.57)	5.11*** (0.63)	0.67 (0.27)	0.56 (0.26)	0.56 (0.26)	0.68 (0.32)
Number of observations	1225	1225	1225	1225	1225	1225	1225
Log Likelihood	-482.68	-482.48	-476.36	-465.38	-433.56	-427.31	-423.45
AIC	969.37	970.97	964.72	944.77	883.12	874.63	868.9
LOO	0.79	0.79	0.79	0.77	0.72	0.71	0.71

Note: *p<0.1; **p<0.05; ***p<0.01

The reference category for Preschool is all other types of care except Head start

Table 4.8: Logistic regression with dependent variable obtained GED (1) over high school dropout (0) with different independent variables and robust standard errors

	Dependent variable:						
	1 = GED / 0 = no qualification						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Care arrange: Preschool	1.27 (0.47)	1.16 (0.44)	1.31 (0.57)	1.39 (0.61)	1.38 (0.61)	1.54 (0.68)	1.50 (0.66)
Care intensity: Full time		1.84^{*} (0.65)	2.06 (0.90)	2.17^{*} (0.97)	2.20^{*} (1.00)	2.37^{*} (1.07)	2.23^{*} (1.02)
Care reason: Child needed play			0.40 (0.37)	0.31 (0.29)	0.27 (0.24)	0.18^{*} (0.16)	0.18^{*} (0.16)
Care reason: Started school			1.19 (0.77)	1.27 (0.83)	1.25 (0.82)	1.20 (0.82)	1.15 (0.78)
Care reason: Started/returned to work			0.81 (0.32)	0.78 (0.32)	0.73 (0.31)	0.64 (0.27)	0.66 (0.28)
Home index				1.14^{***} (0.06)	1.14^{**} (0.06)	1.10^{*} (0.06)	1.10^{*} (0.06)
Highest education in HH					1.16 (0.23)	1.15 (0.22)	1.16 (0.23)
Race: Other races (Ref: Black)						1.84 (1.04)	1.83 (1.04)
Race: White						2.26^{**} (0.85)	2.17^{**} (0.82)
Number of siblings							0.91 (0.10)
Constant	0.93 (0.16)	0.81 (0.15)	0.85 (0.19)	0.07^{***} (0.07)	0.07^{***} (0.06)	0.09^{**} (0.09)	0.11^{**} (0.11)
Number of observations	174	174	174	174	174	174	174
Log Likelihood	-120.39	-118.87	-118.16	-114.18	-113.8	-111.31	-110.94
AIC	244.79	243.74	248.33	242.37	243.59	242.62	243.88
LOO	1.42	1.42	1.46	1.43	1.44	1.45	1.46

Note: *p<0.1; **p<0.05; ***p<0.01

The reference category for Preschool is all other types of care except Head start

CHAPTER 5

Conclusion

This thesis has investigated and brought forward the notion that the institutional characteristics of an educational system can have varying effects based on the socio-economic group of individuals. Moreover, it highlights the importance of these characteristics as they can influence long term inequality between different socio-economic groups. In this chapter I summarize the main conclusions of each of the three empirical articles (Chapter 2 trough Chapter 4), discuss the limitations of each study and describe some of the puzzles that future research should follow.

Chapter 2 studied the impact of a decentralization experiment in Mexico to increase parental participation in the school's decision making on school-level and student-level outcomes. Chapter 3 investigated whether the cognitive achievement gaps between developed and developing countries are related to the tracking and vocational setup of each country. Finally, chapter 4 took a longitudinal approach and studied the relationship between quality early education and graduations rates in the United States.

Chapter 2 concentrated specifically on the role of decentralization using a very rigorous randomized controlled trial. This chapter focused on a case study of Mexico, because the country has experimented greatly with such interventions and there was a great opportunity to use a proper experiment to study the question of interest. In contrast, the third chapter took an international perspective. It documents how the variation in achievement gaps in over 30 countries can be explained by institutional features such as tracking and vocational enrollment. The complementation between both articles lies in the fact that one tackles a question by concentrating on the peculiarities of a single country, while the other takes a simple idea and scales it to over 30 countries to confirm world-wide patterns. In a similar way, chapter 4 is related to the previous chapters in three aspects. First, it studies one specific feature, namely early education, which is tightly related to the institutional features studied in the first two chapters. Secondly, it is a case study of the United States, placing particular importance on the peculiarities of a case study similar to the second chapter. Thirdly, it complements the third chapter by concentrating on the long-term importance of an institutional feature. The third chapter studies how tracking and vocational intensity explain differences in the achievement gap in over 32 countries but it does so in a limited way since tracking and vocational intensity do not change over time. Chapter 4 tackles this limitation by studying whether a single institutional feature can be associated to outcomes up to 20 years later.

Chapter 2 finds that increasing parental participation in the parent's association of the school increases test scores in Mathematics and Natural sciences. In other words, the findings suggest that increasing parental participation improved the cognitive performance of the children. These findings extend the work of Gertler et al. (2012) as they investigate test scores at the school level in a very limited fashion. Moreover, this piece of research improves on the understanding of decentralization reforms in light of the work of Bruns et al. (2011). In a cautionary note, Bruns et al. (2011) note that there are mixed findings on the impact of this reform on student test scores but they warn that this could very well be due to the lack of randomized controlled trials to properly study the outcomes. With the evidence from chapter two we can begin to validate the reform as also having impact on cognitive abilities. However, the findings from the chapter also show a negative side effect of decentralization: the impact is not uniform.

Children coming from high socio-economic groups had a greater impact on improving cognitive abilities relative to children from low socio-economic groups. To put it bluntly, both groups of students received exactly the same treatment and some benefited more from this than others. The topic of an SES gradient in decentralization reforms is something rarely discussed in most debates on the topic (Bruns et al., 2011) and these findings suggest that it should be incorporated. The specific mechanisms through which this happened are still unknown and open to future research. I presume that the link behind this inequality came through the fact that parents from high socio-economic groups had richer experiences with their children in terms of helping with homework, integration at school and higher accountability to teachers, among other things. These mechanisms are still hypothetical but they are in line with most of the work of Waldfogel (2006) which documents staggering improvements for children when cared for in rich environments.

These findings are not bullet-proof and have limitations that need to be addressed in future research. The main limitation of the design is that test scores were asked directly to students. The risks associated with this decisions is that children might have exaggerated their marks. Throughout much of the chapter, the descriptive statistics show that the scores do not present over concentration in top marks, something which would be expected if they lied. However, regardless of that, there are also high risks of measurement error. This dependent variable is censored at the bottom meaning that scores can only be recorded for grades between 6 and 10 and anyone who scored at 5 or below is lumped into one category. This reduces variation and limits the design to a narrower research question. Another important limitation which should be considered in future research is the fact that that test scores are measured four times in four years but there is no control on whether the tests becomes harder/easier over time. This could confound the effect of the treatment seriously and some attempts should be made to harmonize the dependent variable as much as possible.

Chapter 3 and 4 are different from the previous chapter because they do not have a strong design in terms of causality. However, the results from chapter 2 help to show that changes in some of the institutional features of an educational system can have positive effects but also unequal effects. The other two chapters concentrate on whether actual inequality can be linked to existing educational features and whether these features can be associated with long term patterns.

Chapter 3 studied the evolution of the achievement gaps internationally and tested whether the variation in this evolution could be explained by the tracking setup of a country. The results indicate the there is high variation in the evolution of the achievement gap across all countries. For example, the United States shows a marked *decrease* in the achievement gap while France shows an *increase* for the time span between 2000-2015. Once disaggregated, these dynamics are explained by completely different phenomena. In the United States, the decrease is mainly driven by the fact that children from the high socio-economic groups are decreasing their overall performance faster than the lower socio-economic group is increasing. That is, the top SES students are reducing their average score at a rapid rate while the bottom SES students are not improving as fast as the rate of the top group. For France, on the other hand, the top SES students are increasing their scores while the bottom SES students are decreasing their scores. In other words, both groups are distancing themselves and complete top-to-bottom inequality is growing.

Recent evidence from Chmielewski (2016) follows a methodology of calculating global averages in achievement gaps and estimating summary gaps for over 50 years of data. The evidence brought forward here suggests that neither of those approaches is faithful to the dynamics of the data. There is too much between-country variation and within-country variation across time to reliably estimate summary indicators. These summary indicators do no capture the steep increases/decrease found in the 15-year achievement gap. Moreover, the dynamics within each of these gaps are very different between countries making it difficult to evaluate the validity of the results. Most research has concentrated on identifying patterns in changing achievement gaps and attempting to explain why they differ between countries but rarely has the research agenda focused on identifying where the gaps are coming from. Future research should follow the origin of each of these gaps to understand better how to explain them.

The chapter complements the previous findings by attempting to explain why some countries show these marked differences. The results indicate that 40% of the between-country variance in achievement gaps is explained by the current tracking and vocational setup of the country. Note that the definition of tracking in the chapter is more broad than the traditional tracking measurements (Hanushek et al., 2006) as it includes three indicators measuring the percentage of the curriculum that is tracked, the age of selection into tracking, and the number of tracks. This improves on previous empirical results and offers more robust evidence as it captures a more fine-grained definition of tracking. The results also offer a surprising finding: although tracking seems to explain much of the achievement gaps, it has a strong interaction with vocational intensity. On average, whenever tracking is not present, the levels of achievement gaps are very low relative to when there is tracking. When the degree of tracking is high, the achievement gap grows by nearly two standard deviations, quite a substantial number. However, when there is high vocational intensity, the inequality associated with tracking is diminished as the achievement gap narrows by about .7 standard deviations.

The results also bring forward other questions which are puzzling for future research. One of the most important ones raised by these results is, through which mechanisms is vocational enrollment being a strong equalizer when tracking is present? Vocational tracking has been often used as a means of increasing employment rates and promoting fast-track education to technical jobs. However, vocational tracking can also be seen as a remedial opportunity to decrease inequality. These questions have important implications because countries which are severely affected by the inequality produced by their tracking system can begin to counteract their role by introducing policy measures related to vocational enrollment. However, before that happens we need to understand the phenomena much better and that is the role of future research.

One of the main limitations of the chapter is the fact that it attempts to explain a phenomena that changes over time with something that does not change, namely tracking. The methodological design of the chapter applies a partial remedy by calculating the differences between the first and last time point over time and using tracking to explain this cumulative change. Further research should attempt to counteract this limitation by developing indicators at the year level that can measure some of the characteristics of tracking that also change over time. The work of Bol and Van de Werfhorst (2013), together with Brunello and Checchi (2007), is one of the first advances towards this goal as it develops a detailed set of indicators of tracking but not at the yearly level. These advances would help confirm and extend the results outlined above.

The previous question would also improve on our understanding of where tracking is increasing inequality. To be as applied as possible, research needs to understand much better which features of the tracking system can be reformed to reduce the highest level of inequality while being the least disruptive on the current educational infrastructure. These avenues of research are exciting and promising and future research should pay particular attention to these puzzles.

Chapter 4 documents a related but different kind of question. It concentrates on the long-term influence that an institutional feature can have with students. The research question is focused on testing whether quality preschool education is associated with graduating high school over attaining a GED diploma in the United States. The chapter uses the detailed care history of children in more than 17 years of data from the Panel Study of Income Dynamics (PSID), estimates show that being in preschool relative to being cared for by one's own parents is associated with 25% higher odds of attaining a high school degree relative to having a GED/No qualifications. These results are also replicated when comparing the odds against graduating high school vs GED and attaining a GED vs No qualifications with an odds increase of 11% and 60% respectively. These results are evident after a battery of simulations that show that the estimations are stable and keep the expected direction. The results highlight the unique role of quality education when compared to other types of care.

In practical terms, these findings suggest something that had not been properly uncovered in the past: even for remedial diplomas, preschool education seems to be an important degree in being associated with improved chances of graduating high school. And when the student did not graduate high school, preschool educated students still had greater chances of following a GED qualification rather than being a dropout. The relevance of this type of institutional characteristic is that even if it happens at the early stages, it *might* offset a complex chain of events that can increase/decrease the chances of graduating high school even years later after the child experience the care. These types of assertions are of course very strong. This chapter does not use a proper experimental design to make those assertions but the vast literature on early education has indeed found that early education is causally linked to increased graduation rates under rigorous experiments (Elango et al., 2015; Campbell et al., 2002; J. Heckman, 2006; J. Heckman et al., 2010). Based on that literature, the findings from this chapter gain particular relevance because even though it cannot ascertain that preschool caused greater graduation rates over GED graduations, the association becomes a plausible explanation for the rise in graduation rates given that it has already be causally linked to increased graduation rates (Reynolds et al., 2011).

Despite this, the chapter has both substantive and methodological limitations. Methodologically, all statistical models attempt to adjust for a possible self-selection of students into certain types of care. In fact, the chapter introduces a variable rarely available on other research, namely the reason why students were enrolled in that type of care. That variable together with the available information on the cultural and socio-economic levels of the family try to control for self-selection. However, there are other possible mechanisms through which some variables might be endogenous. For example, children who enrolled in preschool might have been because their parents had a greater level of motivation and concern for their future despite their low levels of education. The models presented in the chapter cannot capture this type of behavior, which is just one example among many. The correct approach to testing the direct relationship between preschool education and the odds of graduating among different types of education (High school, GED or dropping out) is to use other sources of information. Most of the data on experiments that allow to compare this is rarely available to researchers for free. There should be stronger efforts to publish such data to properly test these types of question.

Another limitation of this and other types of studies related to this one is the fact that there is usually very low sample sizes in the data which in turn makes the estimation procedure difficult and influences external validity in a negative fashion. An old branch of statistics has been gaining ground in recent years (Gelman and Shalizi, 2013) which allows to use different estimations to come around the problem of sample size by using theory and previous empirical results: Bayesian estimation. Future research should follow this approach and embrace uncertainty whenever the data at hand is not very strong in terms of measurement.

The findings from this chapter raise several questions. First and foremost, considering that preschool is associated with graduation rates, it is imperative to figure out what is it about preschool that improves chances of graduating high school over other types of degrees. This is the current 'black-box' problem in education research as it would more clearly pinpoint specific characteristics that should be changed to improve preschool quality. In a similar line, the results find that there is no interaction between the SES group of the students and their chances of graduating high school. The chapter explores the fact that this could be due to low statistical power but further replication is needed. If in fact there is not an SES gradient, why is this the case? Could it be that because graduating GED is mainly taken up by children from middle and low SES groups, there is not an advantage of coming from middle educated families? These results gain relevance considering that most research in early education finds strong SES gradients (Elango et al., 2015) and the fact that GED enrollment seems to have increased (J. Heckman and Rubinstein, 2001). However this could also be explained by the fact that most of the research has not concentrated on such a selective sample of low SES students, such as

GED holder and dropouts.

The findings from this thesis all converge towards a similar direction. Most of the policy solutions to increasing educational effectiveness seem to be applied uniformly across the educational system. However, these efforts neglect the fact the specific problems need specific solutions. The evidence from this thesis highlights the fact that educational reforms can have winners and losers and at the same time, much of the current inequality can be explained by this same argument. Although this thesis merely adds a grain of salt to the educational literature, future research should build upon it to find ever more convincing evidence in favor of decreasing inequality in a cost-effective fashion.

Bibliography

Bol, Thijs and Herman G Van de Werfhorst (2013). "Educational systems and the trade-off between labor market allocation and equality of educational opportunity." In: *Comparative Education Review* 57.2, pp. 285–308.

Brunello, Giorgio and Daniele Checchi (2007). "Does school tracking affect equality of opportunity? New international evidence." In: *Economic policy* 22.52, pp. 782–861.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos (Feb. 11, 2011). *Making Schools Work: New Evidence on Accountability Reforms*. World Bank Publishing. 251 pp. ISBN: 0821386794. URL: https://www.ebook.de/de/product/14536372/barbara_bruns_deon_filmer_harry_anthony_patrinos_making_schools_work_new_evidence_on_accountability_reforms.html.

Elango, Sneha, Jorge Luis Garcia, James Heckman, and Andres Hojman (Nov. 2015). *Early Childhood Education*. Working Paper 21766. National Bureau of Economic Research. DOI: 10.3386/w21766. URL: http://www.nber.org/papers/w21766.

Gelman, Andrew and Cosma Rohilla Shalizi (2013). "Philosophy and the practice of Bayesian statistics." In: *British Journal of Mathematical and Statistical Psychology* 66.1, pp. 8–38.

Gertler, Paul, Harry Anthony Patrinos, and Eduardo Rodrıguez-Oreggia (2012). "Parental Empowerment in Mexico: Randomized Experiment of the "Apoyo a La Gestion Escolar (AGE)" Program in Rural Primary Schools in Mexico." In: *Society for Research on Educational Effectiveness*.

Heckman, James and Yona Rubinstein (2001). "The importance of noncognitive skills: Lessons from the GED testing program." In: *The American Economic Review* 91.2, pp. 145–149.

Reynolds, Arthur J, Judy A tchapterle, Suh-Ruu Ou, Irma A Arteaga, and Barry AB White (2011). "School-based early childhood education and age-28 well-being: Effects by timing, dosage, and subgroups." In: *Science* 333.6040, pp. 360–364.

Waldfogel, Jane (2006). "What do children need?" In: *Juncture* 13.1, pp. 26–34.