Essays on Education and Inequality

Pilar Cuevas Ruiz

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

UCL Social Research Institute
University College London

March 22, 2024

I, Pilar Cuevas Ruiz, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Declaration

- Chapter 1, 4 and 5 are single-authored by Pilar Cuevas Ruiz.
- Chapter 2 is a joint work with Cristina Borra and Almudena Sevilla.
- Chapter 3 is a joint work with Almudena Sevilla, Ismael Sanz and Luz Rello.

Abstract

This thesis aims to study the intricate linkages between education and inequality, addressing the topic from diverse yet interrelated perspectives.

The first essay provides causal evidence of the returns of mothers' high school curricula to children's health at birth. It explores the impacts on infant health outcomes of a comprehensive education reform in Spain that postponed students' curriculum choices and integrated more general education into the high school curriculum. Using a dose-response difference-in-differences (DiD) research design applied to linked population registers, it finds that the reform significantly reduces very low birth weight and preterm births. The positive outcomes are attributed to improved labor market opportunities for mothers and better family planning.

The second essay evaluates the impact of a cost-effective, computer-assisted learning (CAL) language program on student academic performance in the region of Madrid (Spain). By using artificial intelligence and machine learning, the program tailors content to students with learning difficulties, improving their writing and reading skills. The study reveals that students using the CAL program perform better in standardized Spanish language tests, especially those at the lower end of the test-score distribution, and also reports positive spillover effects in other subjects like mathematics. The findings suggest that CAL programs can effectively mitigate literacy problems at young ages and assist teachers in managing diverse learning levels.

In the third essay, the thesis examines gender gaps in response to competitive pressure using data from high school and university entrance examinations in the region of Andalusia (Spain). Women are found to underperform compared to men in high-stakes settings like university entrance examinations, and they are more affected by performance shocks occurring on the same day. These gender differences, more pronounced for students applying to more competitive university programs and science fields, are not explained by gender differences in unobserved ability or effort provision.

Impact Statement

This thesis provides insights into three national and international policy-relevant topics: inter-generational transmission of inequality, literacy gaps, and gender gaps in education.

Intergenerational Transmission of Inequality

Poor health at birth leads to worse long-term outcomes such as lower educational attainment, earnings, and higher risk of disability (Currie, 2011). According to UNICEF-WHO (2019), the prevalence of low birth-weight (under 2,500 grams) has increased from 7.2% to 7.6% between 2000 and 2015 in high income countries. Perin et al. (2022) provide updated estimates of cause-specific mortality for children under 5 years across all 194 WHO Member States from 2000 to 2019, revealing that complications arising from preterm births account for 17.7% of deaths in this age group. While the influence of maternal education - measured in years - is well documented, less attention has been paid to specific dimensions of maternal education that are nevertheless important for infant health, such as the content of school curricula (Almond et al., 2018). The first essay fills this gap by providing the first empirical evidence on the relationship between mothers' educational curricula and their children's health at birth, as well as mechanisms through which more general education among mothers could improve birth outcomes. It shows that children whose mothers were exposed to the broader educational curriculum had improved health outcomes at birth. Thus, it emphasises the potential benefits of integrating a broader and more general educational curriculum into schools and underscores the importance of considering the impact of educational policies on health outcomes. The findings urge policymakers to prioritise the content of education, alongside its quantity, to pave the way for improved health outcomes for future generations.

Literacy Gaps

Despite a rise in expenditure on schooling of over 15% in the past decade alone, a significant number of students in OECD countries complete compulsory education without achieving basic literacy skills (Vignoles, 2016; Gust et al., 2022). Poor

literacy skills lead to significant economic losses and employment disadvantages, prompting a large number of educational interventions to enhance literacy. These interventions include improving teacher skills, new literacy curricula, and personalized student support (Slavin et al., 2011). However, they are costly and challenging to scale, often requiring more instructional time, new teaching methods, or additional personnel, potentially stigmatizing students with tailored content.

The second essay provides experimental evidence on a cost-effective and inclusive approach: a scalable computer-assisted learning (CAL) program, called DytectiveU, designed to improve literacy skills. Using administrative data on standardized tests combined with survey data, the second essay shows that the software was successful not only at increasing students' performance in Spanish language but also in subjects like mathematics. These results are driven by students at the bottom of the language test-score distribution, who are usually left behind in traditional instruction. Thus, the second essay has important policy implications. CAL programs, like DytectiveU, that provide personalized instruction without requiring teacher assistance, can offer a cost-effective solution for scaling up while avoiding the segregation of students. These programs can be viewed as strategic tools to address the long-standing challenges teachers face in catering to diverse learning needs within a classroom.

Gender Gaps in Education

Despite narrowing gender disparities in education and the labor market, women's educational choices often result in lower expected earnings than men's (Bertrand, 2020). Gender differences in response to competitive pressure account for about 20% of this disparity (Buser et al., 2014). Women generally underperform in highly competitive tests but excel in less competitive environments (Jurajda and Münich, 2011; Ors et al., 2013; Cai et al., 2019; Montolio and Taberner, 2021).

The third essay investigates the gender gap in academic performance under varying levels of competitive pressure, with a focus on university entrance examination, like the Cito exam (Netherlands), the SAT (United States), and the Baccalauréat (France). Using population-level data from the region of Andalusia

(Spain), this essay uncovers a significant gender gap in high-stakes university entrance exams compared to lower-stakes high school evaluations. These findings inform policy recommendations for redesigning standardized tests to mitigate gender performance gaps, suggesting adjustments in exam stakes or diversifying university admission criteria to better include qualified women in competitive, high-paying fields.

Acknowledgements

First and foremost, I extend my deepest gratitude to my supervisors, Cristina Borra, Almudena Sevilla, and Nikki Shure, for their unwavering support throughout my PhD journey. Cristina, your passion for research has been incredibly infectious. You introduced me to the world of research when my path was unclear, and that guidance has left an indelible mark on me. Your wisdom, patience, and kindness are qualities I deeply admire and aspire to emulate. Almudena, words fall short to express my gratitude for the doors you opened for me, leading to prestigious universities and esteemed colleagues. Your faith in me and mentorship across all aspects of academia have been invaluable. Your dedication to me, despite being at such an advanced and respected stage in your career, is truly inspirational. Nikki, I am immensely thankful for your kindness and advice, which significantly lightened this journey. Working as a teaching assistant with you at the Social Research Institute, IoE, has been a profoundly enlightening experience.

I am also grateful to my fellow PhD students at the IoE's Institute for Social Research and the LSE's Social Policy Department. Sharing both the highs and lows with you made this often solitary journey feel less lonely. My heartfelt thanks also go to the members of Parentime - Greta, Valentina, and Francesca - for your warm and welcoming spirit during the beginning of my PhD journey amidst a pandemic and lockdown in a foreign country. Your kindness was a beacon of light during those challenging times. My appreciation also goes to the members of the LSE's Social Policy Department. Their kindness, advice, and feedback in the last year of my PhD have immensely enriched this thesis.

I extend my gratitude to the Ramon Areces Foundation, whose generous funding through the Scholarship for Postgraduate Studies, XXXIV Call for Extension Studies Abroad in Social Sciences, made my PhD journey possible.

Special thanks are due to my family, to whom I dedicate this work with all my love. A special mention to my grandparents, whose sacrifices to break the cycle of inequality and their commitment to education paved the way for me to be where I am today. Though they may never read these words, I want to acknowledge their

understanding and support as I pursued my dreams thousands of kilometers away.

My sincerest appreciation goes to my parents, Paqui and Antonio. You have been the epitome of hard work and the most inspiring role models I could have asked for. Your unwavering belief in me and your unconditional support in every decision I have made have been the cornerstones of my journey.

Lastly, my eternal gratitude is for Manolo, my pillar. Your unconditional support and trust in me have been the foundation upon which I built my resilience throughout this journey. Thank you for always being there, especially in the darkest times, when your words would light up the path and make everything clear again.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Understanding the complex relationship between education and inequality has been a subject of interest for a long time. This interest partly stems from the intricate connection between intergenerational mobility and equality of opportunities (Brunori et al., 2013), revealing a persistent transmission of academic and economic outcomes from parents to offspring (Black and Devereux, 2010). The complexity of these dynamics is further amplified when considering gender disparities, where discrepancies in educational choices systematically lead to significant wage gaps in the labor market (Bertrand, 2020). This thesis is rooted in the human capital production framework based on Cunha and Heckman (2007) and Heckman (2007). This model posits that a child's skill formation process is governed by a multistage technology, which is a function of a child's abilities from the preceding stage in combination with environmental factors, and past investment in education or other forms of human capital that are important for a child's later life outcomes. Cunha and Heckman (2007)'s model integrates a broad spectrum of factors—ranging from cognitive and noncognitive abilities to health stocks—and examines how these elements are influenced by a combination of genetics, environment, and investment. Significantly, it emphasizes the concepts of 'self-productivity' and 'dynamic complementarity', illustrating how capabilities foster and enhance one another over time. This framework is pivotal in understanding the complex underpinnings of educational and health outcomes, as it captures both the critical and sensitive periods of development and the multiplier effects of early investments in human capital.

Building on this theoretical foundation, the first essay of thesis delves into the origins and intergenerational patterns of inequality, with a specific focus on the impact of *in utero* investments on health endowments. It engages with a broad body of literature that has looked at a wide range of determinants of health at birth, as reviewed by Currie et al. (2010), Aizer and Currie (2014), Björklund and Salvanes (2011), Almond and Currie (2011), and Almond et al. (2018). These determinants span from prenatal substance abuse, maternal exposure to air pollution during pregnancy, nutrition, poverty, cash and near-cash transfers, health, stress, participation in social programs, and education. This includes papers by Currie and Moretti (2003), Noonan et al. (2007), Fertig and Watson (2009), Lindeboom et al. (2009), Ludwig and Currie (2010), Currie et al. (2010), Aizer (2011), Hoynes et al. (2011), Almond and Mazumder (2011), Lindo (2011), McCrary and Royer (2011), and Carneiro et al. (2013), among others.

Previous studies on the effects of maternal education on infant health at birth have yielded conflicting results.Currie and Moretti (2003) used college openings in the US as an instrument and found positive effects on birth weight and gestational age. Similarly, Grytten et al. (2014) used the Norwegian compulsory education reform of 1960 and observed a positive relationship between years of education and health at birth. In contrast, McCrary and Royer (2011) exploited school entry age policies in the US and found no significant effects on fertility and infant health. Two studies in the UK, Lindeboom et al. (2009) and Carneiro et al. (2013), reported limited effects of maternal education on infant health at birth using the 1947 compulsory schooling reform and variation in schooling cost during a mother's adolescence, respectively. Employing data from the 1970 British Cohort Study, Conti et al. (2010) shed light on these findings by showing that women who are more likely to attend college possess certain characteristics that enable them to obtain higher returns from additional education in terms of earnings and health behavior compared to those who are at risk of dropping out of high school and are forced to stay in school. However, while these studies have focused on the effects of maternal education as measured by years of schooling, there has been limited research on

other aspects of maternal education, such as the specifics of their educational curriculum. This is due to the challenge of disentangling the effects of additional years of schooling from changes in curricula, as these tend to be modified simultaneously by the policy instruments explored in the literature.

By using high-quality population-level data containing detailed information on children's health endowments and maternal characteristics from birth registers, as well as exploiting a unique policy shock, I am able to disentangle the effects of educational curricula from additional schooling in the first essay. In particular, to study the impact of returns to more general education on offspring's health at birth, I take advantage of the staggered introduction of a national comprehensive educational reform during the 1990s across provinces in Spain. The reform exposed students between the ages of 14 and 16 to more general education while leaving the years of education unchanged. The new comprehensive system was implemented in a staggered manner across provinces over a 10-year period, during which the old and new high school systems coexisted. To identify the effects of more general education on children's outcomes, I constructed an index of exposure to the policy at province level using manually-collected data on the share of 14-year-old students under each high school system during the transition period and implemented a dose-response difference-in-differences (DiD) approach (Callaway et al., 2021). Thus, I compare the health-at-birth outcomes of children born to mothers with different levels of exposure to the policy in a sample of mothers who were enrolled in high school during the transition period.

The difference-in-differences estimates show that children born to mothers with greater exposure to the general curriculum through the education reform tended to have better health outcomes. Specifically, I find that the reform led to a 27.14% and 27.5% reduction in the incidence of very low birth weight and preterm births, respectively. The analysis rejects decreases larger than 10% and 11.48% in the likelihood of low birth weight (less than 2,500 grams) and late preterm (less than 37 weeks), respectively. Several identification checks confirm the validity of the estimates, ensuring that exposure to the educational reform was exogenously de-

termined, thereby ruling out endogeneity concerns due to anticipation effects and selection bias, as well as the potential for confounding effects due to peer composition changes simultaneosly triggered by the reform.

By exploring the mechanisms that may explain how more general education among mothers maps onto improved children health at birth, the first essay also contributes to the literature that has leveraged comprehensive policy reforms to learn about the effects of modifying quality aspects of education on adult labor market outcomes (e.g. Oosterbeek and Webbink, 2007; Hall, 2012; Bertrand et al., 2020; Bellés-Obrero and Duchini, 2021), adult health outcomes (e.g.Palme and Simeonova, 2015; Basu et al., 2018; Fischer et al., 2021), and marriage market outcomes (e.g. Anderberg et al., 2019). Findings on the mechanisms suggest that the observed reductions in the incidence of very preterm births and very low birth weight among mothers exposed to the policy reform are driven by increased maternal labor market opportunities and better family planning practices, rather than increased women's earnings via different occupational choices or assortative mating.

In reflecting on policy implications of the first essay's findings, several context-specific factors and research caveats should be considered. First, our analysis is limited by the demographic scope of the sample, which includes mothers up to the age of 33. This limitation narrows our lens, potentially obscuring the full picture of how educational curricula influence maternal outcomes and children health at birth throughout the entire reproductive life of mothers. Second, the focus on children's health at birth, although theoretically founded and providing valuable insights, leaves the long-term effects of maternal educational curricula on offspring largely unexplored. A future line of research could investigate how the benefits of maternal education influence children's outcomes over time, potentially affecting children's educational attainment, employment prospects, and health status in adulthood. Third, examining the impact of curriculum changes on other mechanisms, such as prenatal care visits, maternal mental health issues (e.g., anxiety or depression), and maternal earnings, remain unexplored and would provide deeper insights. Last, the specific components of the comprehensive curriculum content

that drive the observed health benefits also remain unexplored. Identifying which components—whether literacy, numeracy, or specific subject matter—are most effective in improving infant health outcomes would provide targeted guidance for educational policy development.

In exploring the broader applicability of the findings from the first essay, the approximately 27.5% reductions in very preterm and very low weight births closely align with results seen in other interventions targeting mothers to reduce these rates. Maternal nutrition programs, for instance, have demonstrated similar efficacy in reducing the risks of very low weight and preterm birth by about 53% and 54%, respectively. The importance of the first chapter's results are further amplified given the reported complications stemming from very preterm births. These births are responsible for 52% of infant deaths before the age of 5 in the US (Barfield, 2018), particularly in similar settings in developed countries with high quality healthcare systems and similar prevalence rates. Thus, findings from the first essay imply that two additional years of general curricula would reduce infant death before the age of 5 in 15.4% and would be equivalent to about half of the impact of participating in a nutrition program.

The second essay focuses on the role of new technologies in upgrading education as a tool to reduce the inequality of opportunity among schools and students. Given the large number of students completing compulsory education without achieving basic literacy skills despite the rise in expenditure on schools over the past decade (Vignoles, 2016; Gust et al., 2022), this essay evaluates the impact on students' performance outcomes of an innovative computer-assisted learning (CAL) program, designed to address literacy problems in primary school-aged students through advanced cognitive modeling.

Prior studies evaluating interventions that introduce alternative approaches to teaching literacy show promising results in improving writing and reading skills. Many of these interventions focus on changing how teachers teach literacy or modifying the curriculum, such as providing new pedagogical approaches like synthetic phonics or introducing a more structured daily literacy hour (Machin and McNally,

2008; Machin et al., 2018). Others provide teachers with specific training in different instructional strategies to enhance teacher quality (see, e.g., Jacob, 2017; Loyalka et al., 2019; Johnson et al., 2019; Kerwin and Thornton, 2021; Carneiro et al., 2022). Personalized curriculum interventions, including tutoring with content tailored to individual student needs (Lavecchia et al., 2020; Fryer and Howard-Noveck, 2020; Carlana and Ferrara, 2021), selective tracking based on ability (Duflo et al., 2011; Bouguen, 2016; Özek, 2021), extracurricular support (Lavy et al., 2022), and special education needs programs (Keslair et al., 2012), have also proven effective. However, scaling these educational interventions can be challenging due to costs and inclusivity, often requiring more resources for new teaching methods and additional staff for smaller classes or extracurriculars. Moreover, personalized interventions risk stigmatizing students by highlighting individual difficulties.

Educational CAL programs have emerged as a scalable and low-cost solution to overcome the long-standing challenge of managing heterogeneous learning levels within the classroom (as reviewed by Bulman and Fairlie (2016) and Escueta et al. (2020)). The CAL program evaluated in the second essay, called DytectiveU, is a evidence-based computer game that offers over 42,000 linguistically-patterned exercises based on linguistic patterns and natural language processing techniques (Rello et al., 2017a; Rello et al., 2017b). Key features include personalized instruction adapting to students' age and performance, dynamic adaptability through performance metrics, and versatile deployment across various devices for use in schools, after-school centers, or at home.

To assess the effect of the CAL program on academic achievement, I use population-level data on external standardized testing and survey data from teachers, school heads, and families, leveraging the differential timing of the DytectiveU software's deployment across 308 public primary schools in the Region of Madrid, Spain. Initially introduced in 103 schools in the 2018-2019 academic year by the Ministry of Education of Madrid and the *Change Dyslexia* social organization, it expanded to 206 more schools over the next two academic years. The voluntary nature of DytectiveU's adoption means our sample is limited to those schools that

chose to implement it, allowing to compare, given the non-significant differences at baseline, student performance outcomes in schools first exposed to DytectiveU with those exposed later. To measure academic gains, I use administrative records from the 2019 external standardized tests in math and language for 3rd and 6th graders. These mandatory tests for students in the Spanish education system were administered three months after DytectiveU's implementation, offering insights into students' school performance. Additionally, I use detailed CAL data from student interactions with DytectiveU to estimate a dose-response model and assess actual compliance, and to conduct descriptive research on the delivery of personalized instructional content.

I find that the intervention improves language performance, primarily driven by low-achievers, and has positive spillover effects on their mathematics performance. Students in DytectiveU schools scored 0.08 to 0.12 standard deviations higher on the 2019 Spanish standardized test, with the most substantial gains (0.22 to 0.37 standard deviations) observed in the lowest-performing students; these effects extended to mathematics, with gains of 0.14 to 0.21 standard deviations. Several identification checks confirm the validity of these findings, including using demographic shocks as an instrumental variable, employing a generalized difference-in-differences strategy, performing falsification tests, and imputation of missing data. Further analysis suggests that the observed academic gains are more likely attributable to the software's personalized educational content and its broad applicability across different learning levels, rather than to changes in teaching strategies.

The second essay also extends the literature on the effectiveness of CAL programs in two ways. First, it addresses the gap in experimental evaluation of CAL language programs, which are less common than math-focused programs (Banerjee et al., 2007; Mo et al., 2014; Roschelle et al., 2016; Muralidharan et al., 2019). The very few studies on the effectiveness of CAL language programs often face limitations in external validity due to small sample sizes and specific contexts. The rich population-level data on standardized testing and the sophisticated processing module of the DytectiveU software alleviate external validity concerns of CAL

language programs and reveal their potential to enhance academic performance. Second, this essay digs into the conditions under which CAL programs are most effective. Earlier findings, including those by Muralidharan et al. (2019) and Banerjee et al. (2007), highlight the importance of personalization and adaptivity in CAL programs for developing countries. The second essay builds on this by examining the mechanisms of DytectiveU's success in a developed-country setting, where the educational landscape differs significantly from that of developing countries, focusing on features like content personalization, adaptiveness, feedback, intervention approach, and teacher roles. The findings emphasize the importance of differentiated, adaptive content and timely feedback, suggesting that the effectiveness of reading skill enhancement relies more on the program's capacity to cater to individual learning paths than teachers' roles or implementation methods.

The second essay leaves open multiple avenues for future research as several student outcomes and key elements of the CAL program have yet to be fully investigated. The estimates from the second essay are based on short-term academic outcomes. Further research might investigate whether the positive effects observed in the short-term persist over time, and whether they translate into improved long-term academic performance or better labor or behavioral outcomes. More work is also needed to better understand the elements that underpin the effectiveness of CAL programs. While the second essay provides some suggestive evidence on the importance of delivering personalized and adaptive content, additional studies could disentangle these factors and individually assess their impacts. The role of feedback in the learning process and the influence of deployment modalities (e.g., at-home versus in-school) also remain unexplored. These inquiries are essential for a more nuanced understanding of how distinct aspects of CAL software design contribute to educational gains.

The wider applicability of the second essay's findings is necessarily limited. While the Spanish educational system and the government schools comprising the sample for this analysis offer a valuable context to demonstrate the effectiveness of Computer-Assisted Language (CAL) programs in developed countries, where

pupil-teacher ratios are smaller and issues with teacher qualifications or attendance are less pronounced compared to developing countries, it is important to recognize that the external validity of this study is not a one-dimensional construct. The success of the technology-aided intervention evaluated further hinges on a combination of factors. These include the content design characteristics of the DytectiveU CAL program and the intervention's various elements, such as deployment methods and the application approach, whether as a curriculum substitute or a home-school supplement. Moreover, our findings reflect the outcomes from schools that were both willing and motivated to participate, adding another layer to the context of the second essay's findings.

The third essay explores gender differences in performance and attitudes towards competitive pressure in educational settings, which may account for a significant part of the gender gap in educational choices and labor market outcomes (Buser et al., 2014; Bertrand, 2020). While women tend to underperform compared to men in highly rewarded and more competitive tests, the opposite is true in less competitive settings or when the stakes are lower (Jurajda and Münich, 2011; Ors et al., 2013; Cai et al., 2019; Montolio and Taberner, 2021, among others). However, examining gender differences in response to increased competitive pressure remains a challenge, given the difficulty in isolating increased competition and stakes from other factors in real-life settings, such as gender differences in effort provision, and the self-selection of students into different types of examinations.

To address sample selection bias and external validity concerns, I use administrative records for the universe of students in the region of Andalusia (Spain) from 2010 to 2019 and compare gender differences in performance in high- and low-stakes settings. The Spanish university admission system offers an insightful framework for exploring gender performance gaps in competitive, high-stakes environments. Students take similar tests in high school, accounting for 30% of their university access score, and face a university entrance examination, comprising the remaining 70%, which closely mimics the content and structure of high school tests. Admission thresholds for university programs are demand-based, with a centralized

algorithm allocating students to universities based on their access scores. Despite a high pass rate in the entrance exam (about 90%), achieving a sufficient access score is crucial for entry into the most competitive programs. The system's design, along with the population-level administrative records on high school and university entrance exams, enables two methods to discern gender differences under increased competitive pressure. The first method compares gender disparities in student performances in high-stakes versus low-stakes exams, while the second uses the proximity to the highest admission threshold in the university entrance exam to vary competitive pressure. This approach is grounded in the hypothesis that female students might underperform under such pressure, potentially leading to a more pronounced gender gap in performance on the university entrance examination for students with scores close to the highest threshold.

In line with previous studies, I find that women's performance relative to men's drops during high-stakes university entrance exams compared to lower-stakes high school assessments, with an average decline of 0.31 standard deviations. This gender gap is most pronounced in fields like social and legal sciences, engineering, and architecture, and less so in arts and humanities. Closer to the highest threshold, the gender gap increases to 0.7 standard deviations in favor of men, underscoring the impact of competitive pressure, particularly in access to high-demand university programs. To further explore the impact of competitive pressure on gender performance, I examine how men and women react to performance shocks during the university entrance exam. The university entrance examination's structure, testing up to six subjects over three days, enables analysis of gender differences in response to earlier test performances. The findings reveal that women's performance drops more significantly than men's following performance shocks, particularly in subjects crucial for the university access score, but this trend is not observed in core subjects, where the stakes are lower, or among students not close to the highest threshold. Two additional analyses are run to confirm that these findings are not driven by gender differences in effort provision or ability abilities as measured differently in low-stakes high school assessments compared to high-stakes university

entrance examination.

This essay contributes to the extensive literature on gender differences in response to competitive pressure in educational settings. Prior studies have explored gender gaps in response to increased competition (Ors et al., 2013, Morin, 2015, Iriberri and Rey-Biel, 2019), increased stakes (Azmat et al., 2016, Montolio and Taberner, 2021), and both increased competition and stakes (Jurajda and Münich, 2011, Pekkarinen, 2015, Cai et al., 2019). These studies collectively show that although women may outperform men in less competitive or lower-stakes environments, their performance generally declines relative to men's in highly competitive or high-stakes settings. The uniqueness of this essay's setting and its use of extensive population-level student administrative records set it apart from prior research. First, by focusing on low-stakes high-school examinations, which constitute 60% of the university access score, this approach ensures that the analysis captures gender differences in response to competitive pressure, rather than differences in effort provision, offering a more accurate comparison than previous studies that examined performance between actual and mock examinations. Second, with the university entrance examination being the primary requirement for university admission in Spain, there is limited selection bias among students entering the examination based on their performance in the last two years of high school. Third, the study uses rich population-level data for a decade, allowing for a detailed analysis of the impact of competitive pressure on gender gaps across various fields, thereby avoiding the sample selection bias prevalent in many studies that use laboratory experiments or quasi-natural experiments in highly selected samples.

Understanding the limitations and reach of the third essay is crucial for informing policymakers and guiding future lines of research. A significant factor requiring further exploration is the role of stereotype threat in explaining the gender gap arising in low- versus high-stakes settings. A pivotal question not answered in this essay is: to what extent does the observed gender gap in performance under high-stakes conditions stem from a combination of increased competitive pressure and the influence of stereotype threat? Another confounding factor that cannot be

entirely ruled out relates to the differences in evaluation characteristics between high- and low-stakes settings. Whereas low-stakes high school evaluations employ non-blind grading, high-stakes university admissions are assessed through a blind grading process. This distinction raises concerns about potential teacher grading bias, which may deferentially impact gender performance gaps. Similarly, the environment in which tests are administered constitutes another confounding factor. Examinations conducted in standard high schools for low-stakes assessments versus those administered externally at universities for high-stakes situations might additionally trigger higher pressure. The latter environment, being less familiar to students and potentially more intimidating, could exacerbate stress levels, thereby affecting performance, particularly among female students who may be more sensitive to such environmental factors. Moreover, for the design of long-term strategies aimed at mitigating gender gaps in the labor market, it is also necessary to further investigate how these gender gaps in academic settings translate into later students' decision-making and success in their subsequent careers.

The external validity of the third essay is broadened by its potential applicability across various international educational contexts, given the similarity of the Spanish university entrance examination system to other standardized testing systems worldwide, such as the Cito exam in the Netherlands, the SAT in the United States, and the Baccalauréat in France. These evaluations not only serve to assess academic achievement but also to stratify access to university programs based on performance. Therefore, the findings regarding gender differences in performance under competitive pressure in the Spanish context may offer valuable insights into similar dynamics in other countries. However, the external validity of this essay is not free from limitations that may affect the direct applicability of the findings to other contexts. Differences in educational culture, gender norms, and the specific design of each country's examination system can influence how gender gaps in performance emerge and are addressed. Thus, while the third essay provides a valuable framework for understanding the effects of competitive pressure on gender performance differences, additional research tailored to each unique educational context is

necessary to fully leverage these insights for policy-making and educational reform.

Chapter 2

The Causal Impact of Maternal Educational Curricula on Infant Health at Birth

2.1 Introduction

Better health at birth leads to improved long-term outcomes such as higher educational attainment, earnings, and lower risk of disability (Currie, 2011). Previous research identifies several factors that influence health at birth, including mothers' education, behavior, access to resources, and mental and physical health. However, while the influence of maternal education – measured in years – is well documented, less attention has been paid to specific dimensions of the latter that are nevertheless important for infant health, such as the content of school curricula (Almond et al., 2018). This is largely due to the challenges of randomly assigning different educational curricula to mothers while controlling for other factors, such as their innate abilities, and observing the health outcomes of their offspring (Aizer and Currie, 2014). In this paper, we aim to fill this gap by providing the first empirical evidence on the relationship between mothers' educational curricula and their children's health at birth. To this end, we leverage a policy reform in Spain that integrated more general education into the high school system by postponing students' curriculum choices.

According to standard theories of human capital formation, there are two main theoretical predictions as to why increasing mothers' general schooling could optimize infant health (Cunha and Heckman, 2007; Heckman, 2007). First, educational programs that prioritize general knowledge over specialization tend to impart skills that are more transferable across occupations and thus increase students' earning potential, particularly in the context of shifting demand and technology-induced changes in the labor market (Goldin, 2001;Hanushek et al., 2017). Additionally, positive assortative mating may amplify the impact of a woman's education on household income through a multiplier effect, as a woman's education is causally connected to her partner's education (Behrman and Rosenzweig, 2004). Women with greater purchasing power will tend to acquire more and higher-quality material health inputs, such as better medical care, food, and housing, which can improve their children's health outcomes (Currie, 2009). Second, unlike vocational education, which focuses on specific practical skills for particular occupations, the transferable and flexible nature of general education may go further towards improving children's health outcomes by developing mothers' information processing skills. Information processing has been shown to play an essential role in transmitting the benefits of education (Thomas et al., 1991). These processing skills can increase an individual's ability to acquire knowledge related to healthy behaviors and effective family planning (Grossman, 1972). For instance, an increase in mothers' ability to learn about healthy habits can lead to a reduction in smoking rates and an increase in the use of prenatal care (Currie and Moretti, 2003). Such improvements in processing skills can furthermore explain mothers' ability to use contraceptive methods effectively, leading to a reduced likelihood of unplanned pregnancies and greater control over the timing of births (Rosenzweig and Schultz, 1989).

To study the impact of the returns of a curriculum shift towards more general education on children's health at birth, we take advantage of the staggered introduction of a national comprehensive education reform in 1990 across the provinces of Spain. The reform exposed students between the ages of 14 and 16 to more general education by delaying their choice of curricular track to the age of 16. Thus, all

students were now required to complete an additional two-year general curriculum before splitting into vocational or academic programs. Prior to the reform, students were selected into either vocational or academic tracks at the age of 14. The new comprehensive system was introduced in a staggered manner across the provinces over a 10-year period, during which the old and new high school systems coexisted. To identify the effects of the educational reform on children's outcomes, we constructed an index of exposure to the policy using manually collected data on the share of 14-year-old students under each high school system during the transition period and implemented a dose-response difference-in-differences (DiD) approach (Callaway et al., 2021). This allows us to compare health outcomes at birth of children born to mothers with different levels of exposure to the policy for a sample of mothers who were enrolled in high school during the transition period.

Using cross-sectional data from a large-scale survey, we show that the comprehensive education reform had the intended effect of delivering more general knowledge and learning skills, while keeping the number of years of schooling constant and maintaining different educational tracks. We first document that women received more general education as a result of the policy reform. In particular, we observe a 33% increase in the share of women enrolled in the new comprehensive system, and a decrease in enrollment in the old academic and vocational tracks of about 13% and 25%, respectively, due to the policy shock. When we look at women who are old enough to have completed their education, we find that the policy reform had no effects on the share of women who completed high school (regardless of track) or obtained a college degree. We also rule out the possibility that the reform had any impact on years of schooling, as measured by age at highest educational attainment. These findings suggest that the potential effects of the policy reform are driven by changes in the stock of knowledge resulting from the curricular change, rather than by any differences in educational attainment or qualifications obtained.

Our differences-in-differences estimates show that children born to mothers with greater exposure to the general curriculum through the education reform tended to have better health outcomes. Using detailed administrative data from birth cer-

tificates, we show lower rates of very low birth weight (less than 1,500 grams) and very preterm birth (less than 33 weeks gestation) among children whose mothers were exposed to the reform. In particular, we find that the reform led to a 27.14% and 27.5% reduction in the incidence of very low birth weight and preterm births respectively. Our data reject decreases larger than 10% and 11.48% in the likelihood of low birth weight (less than 2,500 grams) and late preterm (less than 37 weeks), respectively. To give a sense of the magnitude of our estimates, research by Bitler and Currie (2005) shows that maternal participation in a supplemental nutrition program can lower the probability of very low weight and preterm birth by 53% and 54%, respectively. Thus, our findings imply that two additional years of general curricula are equivalent to about 50% of the impact of participation in nutrition programs.

We conduct a series of identification checks addressing the validity of our research design, and the potentially confounding effects of mixing peers. First, we assess whether exposure to the reform was exogenously determined, that is, not subject to anticipation effects nor the Ashenfelter (1978)'s dip, and free from potential selection bias arising from treatment heterogeneity (Goodman-Bacon, 2021; Callaway et al., 2021; De Chaisemartin and D'Haultfoeuille, 2022). We provide suggestive evidence indicating that there was neither endogeneity in the implementation of LOGSE with respect to pre-reform outcomes, nor selection on gains by provinces, as provinces did not sort into different treatment levels according to expected gains; thus, confirming the assumption of strong parallel trends holds. Second, we evaluate whether the influx of low-achieving peers into the new comprehensive system could have negatively affected the behavior of high-achieving peers and the learning environment, potentially limiting the true impact of additional years of general education (Duflo et al., 2011; Garlick, 2018; see Sacerdote (2011) for a review on peer effects). Reassuringly, we observe no significant difference in the impact of the reform on children born to higher- and lower-achieving mothers nor resident in urban and rural areas, further supporting the finding that peer group mixing due to

the policy change had little effect on infant health at birth. [1]

Next, we empirically test how more general education among mothers maps onto the documented changes in infant health through women's increased earning potential and improved information processing skills, as theoretically predicted by previous literature. We explore these mechanisms using data from birth certificates, including information on mothers' employment, marriage market outcomes, and fertility choices, along with hospital discharge records documenting instances of female hospitalizations due to behavior-related health risks. When examining the impact of the policy change on women's earning potential, we find a 3.22% increase in labor force participation among mothers in our sample. Our analysis does not reveal any significant effects on mothers' occupations or the quality of their partners, as measured by the latter's qualifications. Turning to the effects of more general education on health behavior, mothers who are exposed to the policy change are 3.04% more likely to be married at the time of their first birth, which may indicate greater control over the timing of their pregnancies through effective contraceptive use. We do not find any impact of the reform on the number of hospital admissions for conditions related to risky health behaviors. Mothers exposed to the new curriculum do not differ from those who studied under the previous system either in terms of fertility patterns or in their age at their first birth, which is particularly important for ruling out sample selection bias, as we only sampled women who had become mothers. Overall, these findings suggest that the observed reductions in the incidence of very low birth weight and very preterm births among mothers exposed to the policy reform are driven by increased maternal labor market opportunities and improved family planning, rather than by an increased ability to avoid risky behaviors or increased earnings due to different occupational choices or positive assortative mating.

Our paper contributes to a broad body of literature that aims to understand

[1] It is not surprising that we find no peer composition change effects in our setting. Plausibly, this is due to Spain's near-universal high school enrollment rate (95%), and the fact that vocational training was provided in both ordinary high schools and vocational schools before and after the comprehensive policy reform, which minimized the disruption of existing peer groups (Servicio de Estudios Estadísticos, 1994.)

the origins and intergenerational transmission of inequality, as reviewed by Currie et al. (2010), Aizer and Currie (2014), Björklund and Salvanes (2011), Almond and Currie, 2011 and Almond et al. (2018). Much of this literature has looked at a wide range of determinants of health at birth, such as prenatal substance abuse, maternal exposure to air pollution during pregnancy, nutrition, poverty, cash and near-cash transfers, health, stress, participation in social programs, and education. This includes papers by Currie and Moretti (2003), Currie and Neidell (2005), Noonan et al. (2007), Fertig and Watson (2009), Lindeboom et al. (2009), Ludwig and Currie (2010),Currie et al. (2010), Aizer (2011), Hoynes et al. (2011), Almond and Mazumder (2011), Lindo (2011), McCrary and Royer (2011) and Carneiro et al. (2013), among others.

Previous studies on the effects of maternal education on infant health at birth have produced conflicting results. Currie and Moretti (2003) use college openings in the US as an instrument and find positive effects on birth weight and gestational age. Similarly, Grytten et al. (2014) use the Norwegian compulsory education reform of 1960 and observe a positive relationship between years of education and health at birth. In contrast, McCrary and Royer (2011) exploit school entry age policies in the US and find no significant effects on fertility and infant health. Two studies in the UK, Lindeboom et al. (2009) and Carneiro et al. (2013), report limited effects of maternal education on infant health at birth using the 1947 compulsory schooling reform and variation in schooling cost during a mother's adolescence, respectively. Employing data from the 1970 British Cohort Study, Conti et al. (2010) shed light on these findings by showing that women who are more likely to attend college have certain characteristics that enable them to obtain higher returns to additional education in terms of earnings and health behavior than those who are at risk of dropping out in high school and are forced to stay in school. However, while these studies have focused on the effects of maternal education as measured by years of schooling, there has been limited research on other aspects of maternal education, such as the specifics of their educational curriculum. This is due to the challenge of disentangling the effects of additional years of schooling from changes in curric-

ula, as these tend to be modified simultaneously by the policy instruments analyzed in the literature. Our setting provides a unique opportunity to separate out the effects of additional years of schooling and curricular changes thanks to high-quality population-level data with detailed information on children's health and maternal characteristics, as well as a particular education policy shock. Our results complement the findings of Conti et al. (2010) by providing causal evidence that infant health at birth is also affected by mothers' educational curricula, thus offering a novel approach to reconcile the ongoing debate on the effects of maternal education on infant health.

Our unique policy shock and the quality of the data also allow us to empirically explore whether mothers receiving a more general education can result in improved birth outcomes, as suggested by previous work. In doing so, we contribute to the literature that has leveraged comprehensive policy reforms to learn about the effects of modifying aspects of educational quality on adult labor market outcomes (e.g. Oosterbeek and Webbink, 2007; Hall, 2012; Bertrand et al., 2020; Bellés-Obrero and Duchini, 2021; Silliman and Virtanen, 2022), adult health outcomes (e.g.Palme and Simeonova, 2015; Basu et al., 2018; Fischer et al., 2021), and marriage market outcomes (e.g. Anderberg et al., 2019). A common identification challenge in earlier studies of policy reforms such as ours is to disentangle the effects of changes in curricula from other changes in aspects of educational quality, such as shifts in the peer group composition. As students move out of vocational tracks and into comprehensive ones as a result of these education reforms that prioritize general education, they form more mixed peer groups, which has often been often shown to negatively affect educational achievement and other social outcomes such as smoking, drinking, and criminal behavior (Sacerdote, 2011; Galama et al., 2018). However, our analysis finds little evidence of such negative effects of peer mixing, and instead shows that the reform led to a positive pattern of impacts on the female labor market, marriage market, and health outcomes. These results are consistent with our analysis of the effects of changes in peer composition and with the characteristics of the Spanish high school system, which was less conducive to group mixing than

more selective high school systems in other countries, such as Germany or the UK.

This paper is organized as follows. Section 2.2 reviews the institutional framework of the Spanish secondary education systems over the 1970-2002 period.Section 2.3 describes the sample, data and variables used, as well as the reform exposure index that underpins our identification strategy. Section 2.4 discusses the methodology employed to test the effect of more general education on health at birth. Section 2.5 presents our findings on the impacts of general education on maternal education and infant health outcomes, along with the results of several checks to validate our identification strategy. We also discuss the underlying mechanisms through which increased general education may lead to better infant health at birth. Finally, the section 2.6 draws some conclusions.

2.2 Institutional Background: LOGSE Reform

Our analysis exploits a major comprehensive education reform in Spain, known as the LOGSE,[2] implemented in the 1990s, as an exogenous variation in educational curricula to test the effects of more general education on infant health at birth.

2.2.1 Pre-Reform System

Before the introduction of the LOGSE, the Spanish education system was governed by the 1970 LGE.[3] Under this framework, compulsory education (ISCED 1 and 2[4]) was based on a single curriculum framework for basic education, known as EGB (*Educación General Básica*), which covered students aged 6 to 13. Upon completion of EGB, students received a general certificate of admission to further education and were given the choice of following either an academic or a vocational track. The academic track, referred to as *Bachillerato Unificado Polivalente* (BUP; ISCED 3), was a three-year program that emphasized subjects such as mathematics, languages, natural and social sciences, physical education, and religious education. The vocational track, referred to as *Formación Profesional I* (FP I; ISCED

[2]Organic Law 1/1990 (*Ley Orgánica de Ordenación General del Sistema Educativo*).

[3]Law 14/1970 (*Ley General de Educación).*

[4]International Standard Classification of Education (ISCED) adopted by the UNESCO General Conference in its 36th session in November 2011.

3), was a two-year curriculum that focused primarily on practical training with limited exposure to general education. Students who completed either the vocational or academic track could access upper vocational studies, known as *Formación Profesional II* (FP II; ISCED 3). However, only students who had graduated from the academic track were eligible to enroll in the pre-college program, or the *Curso de Orientación Universitaria* (COU), a mandatory requirement for college entry. A visual representation of the main pathways of the Spanish education system prior to the LOGSE is provided in Appendix Figure A.1 for further clarification.

2.2.2 Challenges to the System

In 1990, Spain reformed its high school system. The reform was motivated by two main concerns. The first was the two-year gap between the age at which students completed compulsory education (14 years old) and the legal working age (16 years old) following the passage of the 1980 labor reform.[5] The second concern was related to the overly theoretical and outdated nature of the high school academic program, which was seen as disconnected from the needs of both the labor market and higher education.

2.2.3 Post-Reform System

The new LOGSE postponed the need to choose between academic and vocational education by two years, introduced a new comprehensive system with a greater focus on academic subjects from ages 14 to 16, and extended compulsory schooling to age 16. Specifically, the pre-reform two-track (academic vs. vocational) system was replaced by a new curriculum focusing on general academic subjects such as maths, languages, social sciences, and subjects previously limited to the academic track. Figure A.1 in the Appendix shows the educational curricula before and after the reform. As all students aged 14-16 now received additional general education, those who would have previously chosen the vocational track gained two extra years of general education and those on the academic track received a less specialized cur-

[5]Law 8/1980 of the Workers' Statute.

riculum with a wider range of academic subjects.[6] Upon completion of the new comprehensive system, students could either go on to upper secondary education (*Bachillerato*, ISCED 3), i.e. the academic track, or to lower vocational studies (*Formación Profesional de Grado Medio; Grado medio*; ISCED 3), i.e. the vocational track. Figure A.2 in the Appendix shows the main pathways of the Spanish education system after LOGSE. As in the pre-reform system, lower and upper vocational training continued to be provided at both vocational schools and ordinary high schools after the reform, with the aim of promoting greater inclusion and accessibility to vocational education after completing the comprehensive stage. Starting from the 1991-1992 school year, the new LOGSE also obligated students to remain in school until age 16, either under the old LGE system or the new LOGSE system (Felgueroso et al., 2014; Bellés-Obrero and Duchini, 2021).[7]

2.2.4 LOGSE Reform Implementation Process

The new comprehensive system had to be fully implemented by the 1998-1999 school year, as shown in Figure A.3 of the Appendix. During this 1989-1999 transition period, the old and new education systems coexisted. The time series evidence presented in Figure 2.1 shows the changes in enrollment patterns between the pre- and post-reform systems during the 1990s, as indicated by the share of 14-year-old students enrolled in the two systems. The cohort born in 1975 was the last to graduate fully under the pre-reform high school system, while the 1984 cohort was the first to study exclusively under the post-reform high school system. The Spanish central government allowed education centers ten school years to fully implement the new comprehensive system, and provinces differed in the pace at which they introduced the LOGSE at different levels of education. As we will see in section 2.5.3, the staggered implementation of the reform across provinces and over time is as good as random.

[6] Figure A.1 shows that under the pre-reform system, students who had chosen the academic track were no longer required to take religious education or a subject specific to a particular occupation. In addition, two scientific subjects – one called biology and geology and a second called physics and chemistry – took the place of natural sciences. Four additional subjects specific to academic fields were made available: technology, music and arts, a second foreign language, and classical culture.

[7] In the 1991-1992 school year, the enrollment rate for students aged 14-16 was 95.05% (Servicio de Estudios Estadísticos, 1994)

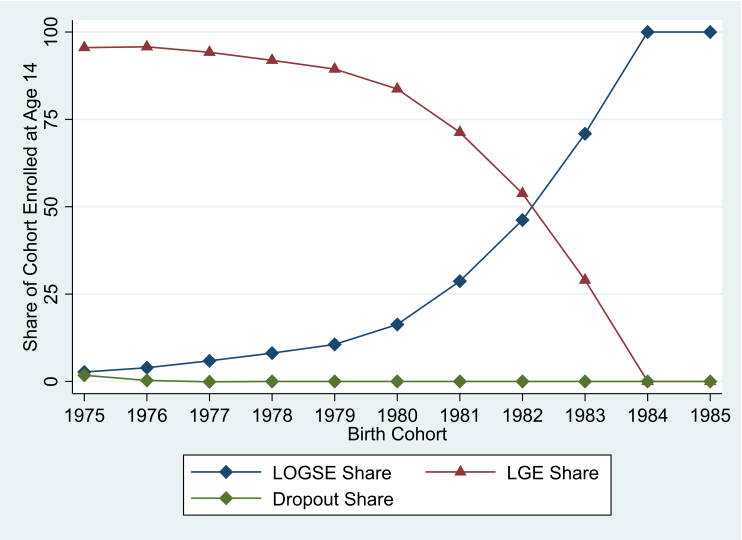


Figure 2.1: The LOGSE and The LGE Enrollment Shares by Birth Cohort
Notes. This figure shows the levels of enrollment of 14-year-olds across the two educational systems by birth cohort, as well as the share of those not in the education system. The 1975 cohort was the first be eligible for the reformed high school system. Source: Annual Education Statistics reports, Statistical Office of the Education Ministry, multiple: 1989-2001.

2.3 Data

Our research employs four primary data sources, linked by the year of birth and the province of residence of women whose offspring's health at birth is being studied.

2.3.1 Childbirth Register Data

Information on health at birth comes from the 2000-2018 childbirth microdata available from the birth statistics of the Spanish Statistical Office. Birth statistics collect data from the birth bulletins, which are filled in when a child's birth is recorded in the Civil Registry. The data recorded in the register include certain characteristics of the birth, such as whether it was single or multiple birth, birth order, gestational age, birth weight, and the parents' demographic and employment backgrounds, as well as the province of registration. In our analysis, we focus on outcomes related to the health at birth of children born to mothers belonging to the 1975-1985 cohorts. We restrict our attention to first births among mothers aged 25 to 33. This age range was chosen because it is expected that most mothers completed their education by age 25, and 33 is the oldest age for the youngest cohort (born in 1985) in our data. Additionally, we exclude immigrants from our sample as mothers born

abroad may not have studied in the Spanish education system and therefore may not be representative of the population we are studying. By focusing on mothers who have completed their education before having children, we are able to examine the full returns of education on infant health at birth. This approach results in an analysis of 1,521,770 first births between 2000 and 2018 to mothers born between 1975 to 1985, ages 25 to 33.

Health at birth is proxied by birth weight and gestational age. Specifically, we look at the incidence of low birth weight (less than 2,500 grams), very low birth weight (less than 1,500 grams), preterm birth (less than 38 weeks), and very preterm birth (less than 33 weeks).[8] We also look at infant mortality outcomes, such as the incidence of fetal death and the rate of survival 24 hours after birth. Summary statistics for these health outcomes at birth can be found in Table A.2 in the Appendix. In our full sample, newborns with low birth weight represent 7% of our sample, while 13% of newborns in our sample are classified as preterm births. Very low birth weight and very preterm births are both around 1%, and the mortality rates at birth are just over 0.1%. This is to be expected, given the high degree of medical contact with pregnant women in Spain, as the public healthcare system guarantees universal coverage for all residents and fully covers prenatal care, delivery, and postpartum care.

We also include variables on mothers' background to further investigate the mechanisms through which the education reform may have affected infant health at birth outcomes. Specifically, we consider mothers' occupation, market participation, partnership choice, and motherhood entry age. Summary statistics for these labor and marriage outcomes can be found in Table A.3 in the Appendix. In our full sample, about 65% of mothers are married, 86% engage in paid work outside the home, and just over 45% practice a recognized profession.

[8]These outcomes have been used in the literature to document the strong association between maternal educational attainment and infant health (Currie and Moretti, 2003;Chou et al. 2010; McCrary and Royer 2011, among others). They have also been used as proxies to measure the impact of health at birth on future child outcomes (Behrman and Rosenzweig, 2004; Almond et al., 2004; Oreopoulos et al., 2006, among others).

2.3.2 Spanish Labor Force Survey

Information on educational outcomes is taken from the Spanish Labor Force Survey (LFS), a quarterly continuous survey administered to approximately 65,000 households and 160,000 individuals. It collects data on the labor force and the population outside the labor market for all individuals aged 16 and over in terms of employment, education, and socio-demographic characteristics. These include educational attainment, age at highest qualification, occupation, nationality, and province of residence. We draw on information from the second quarter of each year, from 1991-2018. Our sample comprises all female respondents aged 17-33 who are Spanish nationals, for a total of 201,701 women born between 1975 and 1985.

Since we are interested individuals' entire educational pathway and not just their highest qualification, we use the LFS data to examine whether individuals have obtained a particular certificate by age 25, prior to the completion of their education. Panel A of Table A.4 in the Appendix presents the summary statistics of the enrollment patterns for our sample. Of the female respondents, 45% followed the academic track, 16% completed the vocational track, and 33% did not achieve a qualification higher than compulsory secondary school education by the age of 25. In Panel B of Table A.4, we focus on four categories in order to observe female degree completion patterns by age 33, once respondents have had the opportunity to fully complete their education. On average, respondents obtained their highest qualification at 20.2 years of age, with 34% finishing high school without continuing their studies, 25% obtaining a vocational qualification, and 36% obtaining a college degree.

2.3.3 Health Data

To evaluate the impact of the reform on adult health, we use hospitalization data from the Spanish MSBD. The MSBD is a administrative and clinical database provided by the Ministry of Health. It collects data directly from public hospitals and contains administrative and detailed medical records on hospitalizations at dis-

charge. We use data from 2004 to 2015[9] and consider several diseases related to health behavior such as diabetes (obesity), cirrhosis (alcohol abuse), lung cancer (smoking), as well as hypertension, which is related to mental health, smoking, and alcohol abuse.[10] Our sample includes female patients between the ages of 25 and 31 born between 1975 and 1985. We use information on the patient's province of residence and birth year to calculate our main adult health outcomes. We document the number of hospitalizations of women due to diabetes, cirrhosis, lung cancer and hypertension for each cohort and province. Table A.5 in the Appendix shows the summary statistics of female hospitalizations for our sample.

2.3.4 LOGSE Exposure Index

Our fourth source consists of data on schooling from the Statistical Office of the Education Ministry. We digitized the province-year data on the number of 14 year-old students enrolled in each academic level during the LOGSE implementation period. Our data spans the period from the 1989-1990 to the 1999-2000 school years, which also includes the first iteration of the LOGSE pilot study, the so-called *Bachillerato Experimental*, which began in 1989.[11] We create an aggregate index to indicate the level of exposure to the LOGSE at age 14, which is broken down by province of residence and year of birth. The 1975 cohort was the first to be exposed to the LOGSE pilot (1989-1990 school year) and the 1985 cohort was the last to be exposed to the previous education system (1999-2000 school year). The index is calculated as follows:

[9] We only work with information from these years due to data availability limitations and changes in the registry.

[10] *See Galama et al., 2018; Basu et al., 2018 and Fischer et al., 2021.*

[11] The *Bachillerato Experimental* was the pilot of the LOGSE, which was approved in the Experimental Reform of Secondary Education in 1983. It was implemented between the 1989-1990 and 1997-1998 school years in a limited number of high schools, mainly in the Basque Country and Navarre regions. *Bachillerato Experimental* was divided into two cycles, beginning with the first cycle of lower secondary education (from age 14 to 16), comprising the first and second years of high school. These two years were compulsory and resulted in a two-year delay in access to vocational studies. The second cycle took place in upper secondary education, comprising the third and fourth years of high school (from age 16 to 18). Students could only continue into the third year if they passed the first cycle.

$$I^{L}_{t,k} = \frac{\text{LOGSE}}{\text{LOGSE} + \text{LGE}}$$

Where $\mathrm{I}^{\mathrm{L}}_{t,k}$ is a continuous treatment function that reflects the implementation of **L**, which is the proportion of students under the LOGSE for cohort t in province k. The numerator represents the absolute number of 14-year-old students enrolled in the new comprehensive system, and the denominator represents the total number of 14-year-old students enrolled in both systems – the new comprehensive system and the old vocational or academic tracks. Note that since the reform induces a change in the curriculum at the age of 14, we exclude students who have to repeat a grade in order to observe the exact proportion of students exposed to the exogenous increase in general education. Hence, the LOGSE exposure index represents the implementation intensity of the reform across provinces and school years during the transition period between the old and new high school systems.

To give a sense of the implementation of the reform, as shown in Figure 2.2, the LOGSE exposure index fluctuated during the transition period between 0 (i.e. no students was under LOGSE) and 1 (i.e. all students under LOGSE in a given province and academic year). Lighter colors correspond to lower levels of LOGSE implementation (captured by $\mathrm{I}^{\mathrm{L}}_{t,k}$) for cohort t in province k. The LOGSE was gradually expanded at varying rates across provinces during the transition period. At the beginning of the period (1992-1993 school year), less than 10% of the 14-year-old student population was under the new system in almost all provinces, but by the end of the transition period (1997-1998 school year), over 60% of 14-year-old students were under the new system.

2.4 Empirical model

In order to identify the effects of more general education on infant health at birth, our research design leverages the staggered implementation of a comprehensive education reform that triggered an exogenous change in the school curriculum. We compare the health outcomes of children of mothers with different exposure to the

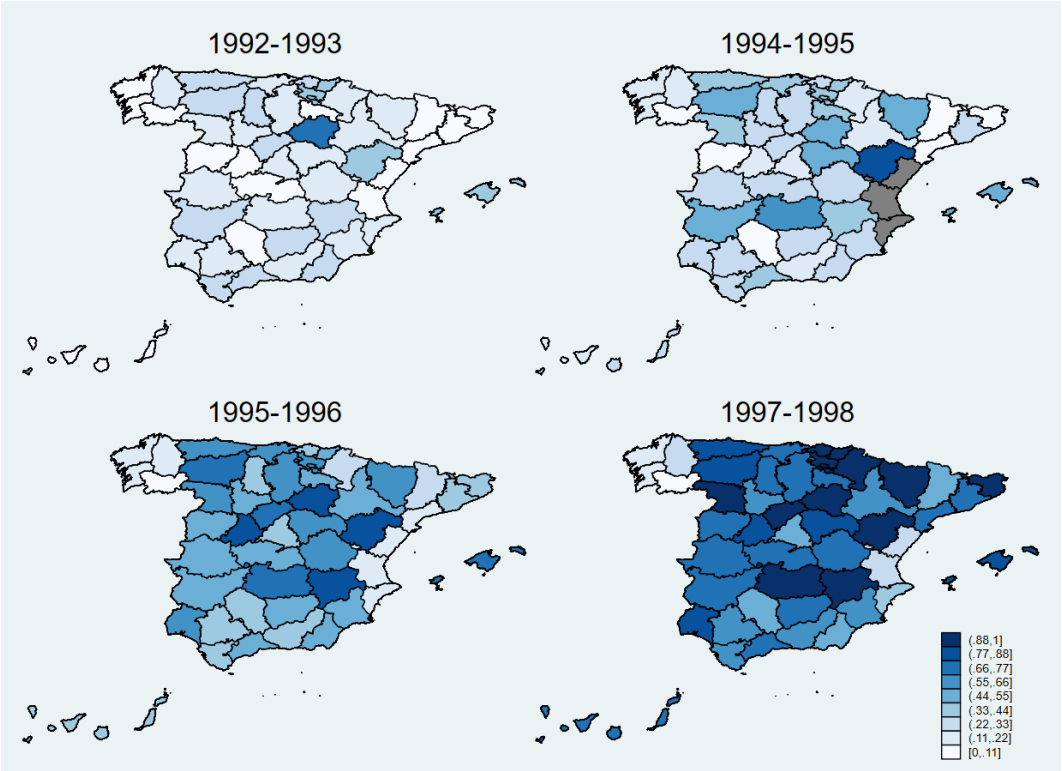


Figure 2.2: Geographic Variation in LOGSE Implementation

Notes: This figure shows the proportion of students under the 1990 LOGSE reform at age 14 (scale of 0 to 1) by province in the 1992-1993, 1994-1995, 1995-1996, and 1997-1998 school years. Increasing levels of LOGSE exposure are indicated by darker blue shading; provinces in grey represent missing data. Source: Education Statistics 1990-1991 and 1999-2000 (Education Ministry).

policy during a 10-year transition period in which the old and new high school systems coexisted. We estimate the following dose-response DiD model separately for each outcome variable:

$$\mathrm{Y}_{itk} = \alpha + \beta \mathrm{I}_{\mathrm{tk}} + y_{\mathrm{t}} + \theta_{\mathrm{k}} + \varepsilon_{itk} \tag{2.1}$$

where Y_{itk} is a child's outcome of interest for their mother i from cohort t and province k. Y_{itk} represents a dependent variable of interest: weight at birth, number of gestational weeks, and indicators for low birth weight (under 2,500 grams), very low birth weight (under 1,500 grams), premature birth (32 to 37 weeks), very premature birth (28 to 32 weeks) and infant mortality at birth.

I_{tk} is our key regressor, which captures the intensity of LOGSE exposure for mothers born in year t in province k. Our coefficient of interest is β, which shows

the relationship between our LOGSE exposure index and health outcomes at birth. A positive coefficient indicates that an increase in mothers' general education is associated with better health outcomes at birth.

Cohort year fixed effects y_t are included to monitor the socio-economic situation of each cohort, while province fixed effects θ_k account for any province-level factors correlated with the education systems. Standard errors are clustered at the province level. ε_{itk} denotes the model error term.[12]

In section 2.5.3, we discuss the additional assumptions needed for two-way fixed effects (TWFE) estimators to be valid in our continuous treatment setting in light of the recent literature regarding DiD applications beyond the canonical binary setting (Callaway et al., 2021).

2.5 Results and Discussion

We present our findings in four sections. In the first section, we evaluate the effect of the LOGSE reform on women's educational attainment, enrollment patterns, and degree completion. Our analysis reveals that while the reform did not significantly impact degree completion rates or years of schooling, it did increase the provision of general education by shifting students from the academic and vocational high school programs into the new comprehensive high school system. In the second section, after addressing sample selection concerns driven by fertility choices, we examine the impact of the reform on infant health at birth. Our findings indicate that the reform reduced incidence of very low birth weight (less than 1,500 grams) and very preterm births (less than 33 weeks). In the third section, we conduct a series of identification and robustness checks. Finally, after discussing the validity of our findings, we empirically test the two theoretical predictions suggested by prior literature, which could explain how the expansion of general education through the reform maps onto the reported changes in infant health at birth. Our analysis shows that the reform led to an increase in labor force participation among mothers and a rise in the likelihood of being married at the time of first birth. We do not

[12] We adjust the standard error for multiple hypothesis testing.

identify any changes in maternal occupation, mate quality, or hospitalization due to behavior-related illnesses.

2.5.1 Effects on Education

Effects on Enrollment

In Table 2.1, we study the impact of the reform on female school enrollment by looking at the educational choices of women aged 17 to 24, i.e. while they may still be enrolled in the education system.[13] We find that the LOGSE exposure leads to an increase of 10.38 percentage points in the share of female students enrolled in the post-LOGSE comprehensive track, a rise of 33.16% (column 2). Meanwhile, the reform reduces the share of women enrolled in the pre-LOGSE academic and vocational tracks by 13.90% and 25.51%, respectively (columns 3 and 4). We find no evidence of a significant impact of the reform on the share of women without any high school credentials (column 1). Overall, the data presented in Table 2.1 suggest that the reform led to a shift of women away from the academic and vocational tracks to the comprehensive system.

Table 2.1: Reform Effects on High School Enrollment

		(1) No Degree	(2) Comprehensive Education	(3) Academic Education	(4) Vocational Education
Index		0.0014	0.1038***	-0.0609**	-0.0467***
		(0.004)	(0.020)	(0.026)	(0.015)
1975's Mean	Cohort	0.011	0.313	0.438	0.183
Std. Dev.		0.10	0.46	0.50	0.39
Observations		109,339	109,339	109,339	109,339
R-squared		0.003	0.019	0.022	0.012

Notes. Standard errors are in parentheses. The estimates are obtained from estimating eq. (2.1) on a sample of female Spanish nationals aged 17-24, born between 1975 and 1985. All specifications include a constant and main controls for birth year and province of residence. Standard errors are clustered at the province level for each specification. Data are from the 1991-2018 Spanish LFS. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

[13]The Spanish LFS does not provide information on current enrollment. We infer enrollment choices by looking at the highest educational attainment before education is completed. See panel A of Table A.4 for the exact definition of educational enrollment choices.

Effects on Degree Completion

Table 2.2 shows the impact of the reform on female educational attainment by looking at LFS respondents' age at their highest qualification and whether they have completed a particular qualification by age 33.[14] Column 1 of Table 2.2 shows that the reform did not change the age at which the highest qualification was obtained and thus had no effect on total years of schooling. Columns 2 to 5 of Table 2.2 show that the LOGSE did not lead to a greater percentage of women finishing their schooling with a particular qualification. Hence, the LOGSE did not succeed in increasing the share of women with a high school diploma or vocational credentials, nor did it affect the share of women with college degrees. In sum, our degree completion estimates suggest that the reform did not increase educational attainment or years of schooling.[15]

All in all, the results from Tables 2.1 and 2.2 suggest that the reform did not affect women's educational attainment, but rather induced a change in the curriculum towards more general education.

2.5.2 Effects on Health At Birth

Table 2.3 reports the reduced-form effects of the reform on our main health outcomes at birth: birth weight, gestational age, and infant mortality at birth. In order

[14]Here, our sample comprises women aged 25 to 33 in order to limit our analysis to women who have completed their education. See Panel B of Table A.4 in the Appendix for the exact definition of degree completion outcomes.

[15]The LOGSE's lack of impact on years of schooling among women also suggests a limited causal role of the reform on fertility patterns. This alleviates a concern about sample selection bias since we only sample women who became mothers. Prior studies on the effects of education on infant health at birth have looked at the effect of years of schooling on fertility patterns in order to rule out potential sample selection bias due to education policy reforms affecting fertility patterns (Currie and Moretti (2003); McCrary and Royer (2011)) Additional years of schooling may reduce teen pregnancies through the "incarceration effect," defined as a delay in fertility equal to the amount of additional time spent in school, since this may reduce the time available for engaging in risky behavior, and hence improve health at birth by preventing pregnancies at a young age (see Black et al. (2004); Cygan-Rehm and Maeder (2013); Geruso and Royer (2018)). Moreover, as Becker (1965)'s quality/quantity trade-off suggests, more years of education may induce women to have fewer children of higher quality. As little is known about the effects curricular changes on fertility, we use the Spanish Administrative Birth Registry to create a panel of annual birth rates by mother's age at the province level to test the impact of the LOGSE on fertility patterns. Column 1 of Table A.9 in the Appendix shows no effects of the LOGSE on birth rates. Column 2 of Table A.9 in the Appendix shows no effects of the reform on the age at first birth. Thus, we can defend that the reform has not led to a selected sample of observed mothers and women more and less exposed to the reform form an equivalent sample.

Table 2.2: Reform Effects on Degree Completion

	(1) Age at Highest Qualification	(2) No Degree	(3) High School Degree	(4) Vocational Degree	(5) College Degree
Index	0.2198	-0.0007	0.0045	0.0334	-0.0330
	(0.184)	(0.005)	(0.018)	(0.028)	(0.028)
1975's Cohort Mean	19.641	0.013	0.360	0.246	0.343
Std. Dev.	4.32	0.11	0.48	0.43	0.47
Observations	85,004	85,348	85,348	85,348	85,348
R-squared	0.033	0.004	0.026	0.011	0.022

Notes. Standard errors are in parentheses. The estimates are obtained from estimating eq. (2.1) on a sample of female Spanish nationals aged 25-33, born between 1975 and 1985. All specifications include a constant and main controls for birth year and province of residence. Standard errors are clustered at the province level for each specification. Data are from the 1991-2018 Spanish LFS. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

to capture the effects of completed education, we only consider mothers aged 25 to 33.[16]

Panel A of Table 2.3 focuses on the effects of the reform on birth weight and the incidence of low and very low birth weight (less than 2,500 grams and less than 1,500 grams, respectively). Column 1 of Panel A shows no effects of the reform on birth weight. The confidence intervals allow us to reject a positive impact of LOGSE on birth weight greater than 0.74%.[17] Column 2 of Panel A shows negative, non-significant effects on low birth weight, and we can reject reductions in the incidence of low birth weight greater than 10%. There is, however, a significant drop of 0.19 percentage points in the incidence of very low birth weight (column 3, Panel A). Given the overall incidence of very low birth weight, this represents a decrease of 27.14%. This result survives a multiple hypothesis testing correction (Romano-Wolf P-value=0.068). Therefore, the estimates from Panel A suggest that the reform had limited positive effects on birth weight.

Next, we evaluate the impact of the reform on gestational length, defined as the number of gestational weeks, as well as the incidence of late preterm and very

[16] See McCrary and Royer (2011) on the different implications of completed versus ongoing education at the time of entering motherhood.

[17] The upper bound of the coefficient interval was calculated as the sum of the point estimate of the coefficient (10.9079) and 1.96 times the standard error (7.011). This result was then divided by the population average (3199.580). The calculation can be summarized as follows: 10.9079 + 1.96 * 7.011/3199.580.

preterm births (less than 36 weeks and less than 33 weeks, respectively). Column 1 of Panel B shows no significant impacts of the reform on the number of gestational weeks. We can reject a positive increase in the number of weeks larger than 0.21%. Column 2 of Panel B shows that there are no significant effects of the reform on the incidence of late preterm births, and we can reject a reduction in the incidence of more than 11.48% In contrast, column 3 of Panel B indicates a significant decrease of 0.33 percentage points in the incidence of very preterm births. Given the overall incidence of very preterm births, this represents a decrease of 27.5%. This result also survives a multiple hypothesis testing correction (Romano-Wolf P-value=0.01). Thus, the reform led to a lower share of very preterm births, confirming the positive effects of the LOGSE on infant health at birth.

Panel C displays the effects of the reform on our infant mortality measures: the likelihood of fetal death and the likelihood of survival 24 hours after birth. Given that Spain has a universal public health care system with a 99.9% chance of survival 24 hours after birth and a 0.01% incidence of fetal death, our finding that the reform had no impact on mortality outcomes is consistent with expectations.

Table 2.3: Reform Effects on Health at Birth

	(1)	(2)	(3)
Panel A: Weight at Birth Measures			
	Weight	Low Weight	Very Low Weight
Index	10.9079	-0.0014	-0.0019*
	(7.011)	(0.003)	(0.001)
Romano-Wolf p-value			0.0689
1975's Cohort Mean	3199.580	0.070	0.007
Std.Dev	506.73	0.25	0.09
Obs	1,446,005	1,446,005	1,446,005
R-squared	0.002	0.000	0.000
Panel B: Gestational Age			
	Weeks	Late Preterm	Very Preterm
Index	0.0229	0.0034	-0.0033***
	(0.032)	(0.006)	(0.001)
Romano-Wolf p-value			0.01
1975's Cohort Mean	39.145	0.132	0.012
Std.Dev	1.90	0.34	0.11
Obs	1,296,160	1,296,160	1,296,160
R-squared	0.001	0.001	0.000
Panel C: Mortality at Birth			
	Fetal Death	Survive 24h after Birth	
Index	-0.0001	-0.0000	
	(0.000)	(0.000)	
Romano-Wolf p-value			
1975's Cohort Mean	0.002	0.997	
Std.Dev	0.05	0.05	
Obs	1,513,676	1,513,676	
R-squared	0.001	0.000	

Notes. Standard errors are in parentheses. The estimates are obtained from estimating 2.1 on a sample of female Spanish nationals aged 25-33 and born between 1975 and 1985. All specifications include a constant and main controls for birth year and province of residence. Standard errors are clustered at the province level for each specification. Data are from the 2000-2018 Childbirth microdata of Vital Statistics (INE). ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

2.5.3 Threats to Identification

We address two threats to identification: (i) the validity of our identifying assumptions, and (ii) the possible confounding impact of peer group changes.

2.5.3.1 Identifying Assumptions

Our identification strategy is based on the staggered introduction of the new comprehensive system across Spanish provinces. First, to alleviate potential anticipation effects, the roll-out of the reform needs to be exogenous with respect to pre-reform education and infant health patterns (Goodman-Bacon, 2021). Second, according to Callaway et al. (2021) and De Chaisemartin and D'Haultfoeuille (2022), the use of continuous dose-response DiD models for investigating treatment effects requires the additional assumption of no selection bias among groups treated with different intensity levels.In what follows, we look at the assumptions of treatment exogeneity and selection bias.

Treatment exogeneity requires that the implementation of the policy is unrelated to prior outcomes. It requires that provinces' past enrollment and health at birth patterns do not influence the implementation of the policy, either because provinces that anticipate larger benefits are more eager to implement the policy (the so-called anticipation effects) or because provinces subject to negative shocks are more likely to be treated earlier (the so-called Ashenfelter (1978)'s dip). We provide two exercises that support the assumption of exogeneity in the implementation process.

First, to test the exogeneity of the implementation of the LOGSE with respect to other macroeconomic outcomes that may affect outcomes related to education and health at birth, we regress the LOGSE exposure index on provincial GDP per capita, female employment, and labor participation rates. The results are displayed in Table A.6 in the Appendix, and show no correlation between the LOGSE exposure and any of the macroeconomic variables. This indicates that the roll-out of the LOGSE was unrelated to any other economic determinants.

Second, to ensure that pre-treatment education and health at birth are not correlated to the implementation of the LOGSE, we perform two placebo checks. We

restrict our sample to five cohorts prior to the first cohort exposed to the reform (1970-1975), and create placebo lead variables that capture differences in the uptake of the reform five cohorts later. If differences in school enrollment patterns and health at birth outcomes across cohorts and provinces are spuriously related to the subsequent implementation of the reform across provinces and cohorts, the coefficients on the placebo lead variables should be statistically different from zero. Tables A.9 and A.10 in the Appendix show that this is not the case, and thus the education and health impacts do not appear to be spuriously correlated to differences in the adoption of the LOGSE across cohorts and provinces.

Additionally, using a continuous treatment requires being able to rule out selection into treatment intensity levels, which is usually referred to as the strong parallel trends assumption (Callaway et al., 2021; De Chaisemartin and D'Haultfoeuille, 2022). In essence, this assumption arises because of potential heterogeneity in gains. If provinces that expect higher gains, self-select into higher intensity treatment levels, then the continuous treatment DiD would be contaminated by this selection bias. So far we have seen that the roll out of the policy was unrelated to provinces' macroeconomic variables and was not likely subject to any Ashenfelter's dip or anticipation effects. However, the assumption of strong parallel trends involves comparing *potential* outcomes across treatment-dose groups, and therefore cannot be tested in most circumstances. Next, we show that the strong parallel assumption holds and the timing of the LOGSE implementation is unrelated to prior education and health at birth patterns across provinces, further supporting the hypothesis of no self-selection into treatment intensity levels.

To examine parallel trends and analyze the dynamics of treatment effects, we conduct an event-study analysis using the TWFE model, incorporating indicators that measure the proximity to the implementation of the LOGSE Reform. Our approach involves estimating the following specification:

$$Y_{itk} = \alpha + \beta_g \times \sum_{g=-6}^{g=+2} D_{g(kt)} + \gamma_t + \theta_k + \varepsilon_{itk} \tag{2.2}$$

where Y_{itk} is binary indicator equals to 1 for very premature birth (28 to 32

weeks) and $D_{g(kt)}$ is a set of indicator variables that take value 1 if, for a child born to a mother from cohort t in province k, the implementation of LOGSE was g cohorts away. [18] [19] To define $D_{g(kt)}$, we transform our LOGSE exposure index into a binary indicator, which takes value 1 when in k province and for t cohort at least 75% of the student population in under LOGSE. Therefore, we compare children born to mothers enrolled in the high school system when more than 75% of their cohort was under LOGSE in their province with those born to mothers from the last set of provinces and cohorts that adopted the reform, considering the latter as a 'never treated' group.

Following Braghieri et al. (2022), as the fully dynamic version of TWFE model in equation (2.2) provides only valid estimates under strong assumptions regarding to treatment effect homogeneity (Sun and Abraham, 2021), to allow for heterogeneity in treatment effects across time and treated units we present the event study figures generated by a set of recently proposed estimators that are robust to treatment effect heterogeneity (Callaway and Sant'Anna, 2021; Borusyak et al., 2021). Figure 2.3 displays the event-study figures and demonstrates that the estimates align with the parallel trends assumption: regardless of the estimator used, the coefficients for cohorts prior to the implementation of LOGSE are consistently near zero, showing no significant pretrends. Figure 2.3 also clarifies the dynamics of treatment effects: all the recently developed robust estimators show treatment effects that increase over time in the postperiods. The increase treatment effects over time could be attributed to the prolonged exposure to the new curricula.

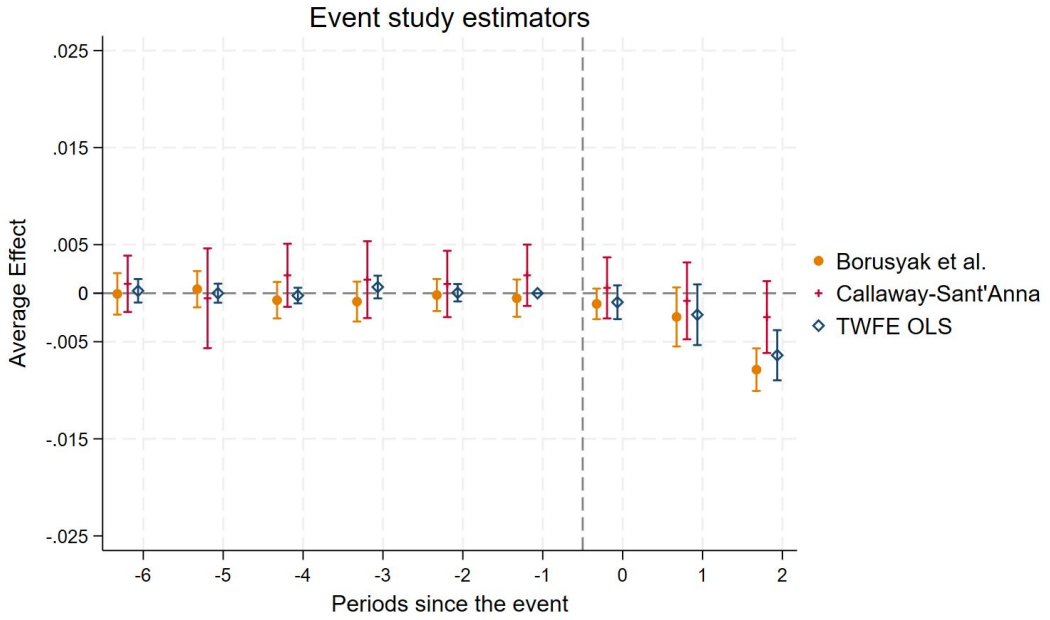
To further test that there is no self-selection into treatment intensity levels and that prior outcomes do not predict LOGSE entry, we follow Ferrara et al. (2012) and Amuedo-Dorantes et al. (2018) and aggregate the data prior to the introduction of the reform at the cohort and province level to estimate the following model:

$$\text{ICohort}_k = \alpha + \beta \text{X}_{1975\text{k}} + \delta Z_{1975\text{k}} + \varepsilon_k \tag{2.3}$$

[18] All other variables are defined as in equation (2.1).

[19] We conduct this exercise for the likelihood of very premature birth (28 to 32 weeks) as this outcome variable has the most precisely estimated coefficient for the reform's effects, as indicated in Panel B of Table 2.3.

Figure 2.3: Effects of the Reform on the Likelihood of Very Preterm Birth Based on Distance to/from LOGSE Implementation



Notes. This figure combines the event-study plots constructed using three different estimators: a dynamic version of the TWFE model (in blue with diamond markers); Callaway and Sant'Anna (2021) (in red with cross markers); and Borusyak et al. (2021) (in orange with circle markers). The outcome is a binary variable that takes value 1 for very preterm birth (28 to 32 weeks). The time variable is the birth cohort of the mother and the treatment group is defined by the province where at least 75% of the student population was under LOGSE during the corresponding cohort period of the mother. We restrict our sample to mothers born between 1975 and 1983, rather than including those born up to 1985 as in estimates from Eq. (2.1) This decision is based on the fact that from 1984 onwards, in all provinces, more than 75% of the student population was under LOGSE. Model 2.2 requires the use of control units as a 'never treated' group, which limits the maximum number of post-periods that can be estimated to two. We estimate six pre-periods because the youngest cohort studying in a province with more than 75% of its population under LOGSE was the 1978 cohort. The bars represent 95 percent confidence intervals. Standard errors are clustered at the province level.

Our dependent variable ICohort$_k$ *is* is the birth year of the first cohort for which the LOGSE exposure index is higher than 0.5 in province k. Our regressors of interest X_{1975k} and Z_{1975k} are aggregated at the cohort and province levels and measured for the 1975 cohort, which was not yet exposed to the new system. X_{1975k} represents either female educational outcomes or children's health at birth outcomes and δZ_{1975k} includes the macroeconomic controls. The results in Tables A.7 and A.8 in the Appendix indicate that neither education outcomes not health at birth outcomes prior to the implementation of the LOGSE predict the year in which the LOGSE came to replace the previous system. Therefore, the incidence of the reform does

not appear to be explained by prior education and health patterns.

2.5.3.2 Potential Confounding Factors

Thus far, our analysis has assumed that the reform's positive effects on infant health at birth are directly driven by the integration of more general education into the high school system. However, a remaining concern is that the policy reform may have affected adult maternal outcomes, and thus infant health at birth, through changes in peer composition. Specifically, the changes brought about by the reform in the structure of high school educational tracks resulted in students moving out of vocational tracks and into the comprehensive system. If the new influx of low-achieving peers into the new comprehensive system had a negative effect on the behavior of high-achieving peers and the learning environment, our health at birth estimates may provide a lower bound of the effects of expanding general education. Educational tracks and peer effects are closely linked, and prior literature suggests important interactions between the school environment, peers, and health. For instance, Robalino and Macy (2018) and Gaviria and Raphael (2001) show that the prevalence of smoking, drug use, and alcohol consumption among high school students can be influenced by their peers. Likewise, Basu et al. (2018) observe that the shift from an early-tracking system to a comprehensive one led to increased rates of depression and smoking, with a greater impact on students with lower cognitive abilities who might have been more susceptible to the decline in the quality of the learning environment.

To address this concern, we conduct two additional exercises to examine whether the effects on health at birth in response to LOGSE exposure vary for students who experience different peer composition changes. First, we use mother's place of residence as a proxy for the peer composition changes induced by the reform and explore how these effects vary between mothers from rural (less than 50,000 inhabitants) and urban (more than 50,000 inhabitants) areas. We hypothesize that students in less populated areas will face fewer changes in the peer composition as a result of the reform than those in more populated areas. In less populated areas, such as small towns and rural areas, exposure to changes in the school envi-

ronment is expected to be limited due to tight-knit social structures. That is, students with extended social connections outside of school, perhaps through family or community ties, may be less influenced by their peers in the school setting due to pre-existing social networks that are less dependent on the school environment (Carrell et al., 2009). If no differences are observed between these two groups, this would suggest that the role of peer composition changes is limited in our analysis. Second, as the reform has no effects on educational attainment, degree completion rates, or type of job, we use mothers' occupation as a proxy for their educational achievement and analyze whether children born to higher-achieving mothers benefit less from the reform than those born to mothers with lower career achievements. We hypothesize that high-achieving students may benefit less from the influence of low-achieving peers, while low-achieving students may benefit more from exposure to higher-achieving peers. We define high-achieving mothers as those with jobs in managerial or professional occupations. If we observe no differential effects of the policy change on children born to mothers with different professional achievements, this would suggest that the impacts on birth weight and gestational age were not affected by changes in peer composition. To perform these two exercises, we estimate the following model:

$$y_{itk} = \alpha + \beta_1 D_i \times I_{tk} + \beta_2 D_i + \beta_3 I_{tk} + \gamma_t + \theta_k + \varepsilon_{itk} \tag{2.4}$$

Where D_i is a dummy variable that takes on a value of 1 if the mother i either (a) has a job in a managerial or professional occupation or (b) lives in an area with less than 50,000 inhabitants. All other variables are defined as in equation (2.1). The interaction $D_i * I_{tk}$ identifies the impact of the reform on the gap in health at birth outcomes between infants born to mothers (a) with high- and low-skilled jobs or (b) living in rural and urban areas. The coefficient of interest, β_1, measures how differences in the health at birth of infants born to mothers exposed to different changes in peer composition are affected by the reform. Table 2.4 presents the results from this exercise. The estimated coefficients corresponding to the interaction between the LOGSE exposure index and mother's occupation are statistically equal to zero

(first row of panel A), which suggests that the reform had no differential effect on the health of infants born to mothers of different abilities. Panel B provides further support for our findings by showing no significant differential effects due to the size of the mother's place of residence. While there appear to be some significant differences in the incidence of low birth weight and the number of gestational weeks at the 10% level (Columns 2 and 4, Panel B), these differences are arguably not relevant due to their varying implications for health at birth and low statistical power. The findings suggest that changes in peer composition caused by modifications to curriculum tracks do not play an important role in our health at birth estimates.

Table 2.4: Peer Composition Changes

	(1) Weight	(2) Low Weight	(3) Very Low Weight	(4) Weeks	(5) Late Preterm	(6) Very Preterm
Panel A: High-Achieving Vs Low-Achieving						
Index*HighAchieving	-2.9574	-0.0014	-0.0008	0.0002	0.0033	-0.0001
	(5.065)	(0.002)	(0.001)	(0.018)	(0.003)	(0.001)
High Skills	15.1862***	-0.0096***	-0.0014***	0.0484***	-0.0143***	-0.0028***
	(3.552)	(0.001)	(0.000)	(0.012)	(0.002)	(0.000)
Index	11.1644	-0.0016	-0.0018	0.0383	0.0012	-0.0033**
	(7.095)	(0.003)	(0.001)	(0.036)	(0.007)	(0.001)
1975's Cohort Mean	3199.580	0.070	0.007	39.145	0.132	0.012
Std.Dev	506.73	0.25	0.09	1.90	0.34	0.11
Obs	1,343,986	1,343,986	1,343,986	1,199,747	1,199,747	1,199,747
R-squared	0.002	0.001	0.000	0.002	0.001	0.000
Panel B: Rural Vs Urban						
Index*Rural	-1.3970	0.0028*	0.0004	0.0293*	-0.0032	0.0006
	(2.368)	(0.001)	(0.000)	(0.015)	(0.002)	(0.001)
Rural	-4.4167	-0.0009	-0.0002	0.0029	-0.0014	-0.0003
	(3.108)	(0.001)	(0.000)	(0.010)	(0.002)	(0.000)
Index	11.7451	-0.0030	-0.0021**	0.0053	0.0053	-0.0037***
	(7.305)	(0.003)	(0.001)	(0.033)	(0.006)	(0.001)
1975's Cohort Mean	3199.580	0.070	0.007	39.145	0.132	0.012
Std.Dev	506.73	0.25	0.09	1.90	0.34	0.11
Obs	1,446,005	1,446,005	1,446,005	1,296,160	1,296,160	1,296,160
R-squared	0.002	0.000	0.000	0.001	0.001	0.000

Notes. Standard errors are in parentheses. The estimates are obtained from estimating eq. (2.4) on a sample of female Spanish nationals aged 25-33, born between 1975 and 1985. All specifications include a constant and main controls for birth year and province of residence. Data are from the 2000-2018 Vital Statistics (INE) childbirth microdata. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

2.5.4 Mechanisms

We have shown that, among women, the reform's main effect was to induce a switch from academic and vocational high school programs to a comprehensive high school system, which extended general education for two additional years. Further, this had no significant effects on women's ultimate rates of high school and college degree completion, or on their years of schooling (Tables 2.1 and 2.2). However, the boost to general education among women did significantly reduce the incidence of very low birth weight and very preterm births (Table 2.3). In this section, we empirically test how extending mothers' general education maps onto the documented changes in infant health by increasing women's earning potential and improving their information processing skills, as theoretically predicted by previous studies.

Effects on Women's Earning Potential

Panel A of Table 2.5 presents the effects of the LOGSE reform on labor market outcomes. Increased general education can improve women's permanent income and maternal prenatal investment through the labor market by increasing skill portability across occupations (Goldin, 2001; Hanushek et al., 2017). Our data shows a sharp increase in the share of mothers joining the labor force at the time of their first childbirth (Column 2, Panel A). The reform significantly increases the likelihood of a mother being engaged in paid work outside the home by 2.75 percentage points, which represents a 3.22% rise. We observe no economically or statistically significant effects on type of female occupation (Columns 3 and 4, Panel A) or on the likelihood of still being enrolled in the education system by age 33 (Column 1, Panel A), which is consistent with our results on degree completion and educational attainment (Table 2.2). These findings suggest that the reform had positive effects on maternal labor force participation by the time of their first birth due to a greater share of mothers being engaged in paid work outside the home, potentially leading to an increase in mothers' earning potential and prenatal investment.

Panel B of Table 2.5 reports the effects of the reform on assortative mating. Increased earning potential resulting from more general education may also contribute to positive assortative mating, leading to higher household permanent income and

prenatal investment through a multiplier effect (Behrman and Rosenzweig, 2004. Our data show no significant effects on mothers' mate quality (Columns 1 and 2, Panel B), as measured by their mate's occupation (McCrary and Royer, 2011). Thus, mothers with greater exposure to the reform do not tend to have children with more- or less-qualified partners compared to mothers less exposed to the reform.

Table 2.5: Reform Effects on Women's Earning Potential

	(1)	(2)	(3)	(4)
Panel A. Labor Market Outcomes	Student	Working Mother	Qualified Job	Non-Qualified Job
Index	-0.0012	0.0275**	-0.0035	0.0029
	(0.001)	(0.014)	(0.012)	(0.010)
Romano-Wolf P-value		0.043		
1975's Cohort Mean	0.005	0.853	0.409	0.264
Std.Dev.	0.07	0.35	0.49	0.44
Obs	1,521,770	1,467,386	1,521,770	1,416,631
R-squared	0.001	0.028	0.014	0.010
Panel B. Assortative Mating	Mate Qualified Job	Mate Non-Qualified Job		
Index	0.0050	-0.0180		
	(0.012)	(0.013)		
Romano-Wolf P-value				
1975's Cohort Mean	0.340	470		
Std.Dev.	0.47	0.50		
Obs	1,521,770	1,521,770		
R-squared	0.015	0.020		

Notes. Standard errors are in parentheses. The estimates are obtained from estimating eq. (2.1) on a sample of first deliveries of mothers who are Spanish nationals aged 25-33 and born between 1975 and 1985. All specifications include a constant and main controls for birth year and province of residence. Romano-Wolf p-values based on 1,000 studentized bootstrap replications. Data are from the 2000-2018 childbirth microdata from the Spanish Statistical Office.

Effects on Health Behaviors and Family Planning

Panel A of Table 2.6 presents the effects of LOGSE on the number of female hospitalizations due to behavior-related illnesses. This analysis is to examine the possible effects of the reform on adult health, which may be driving the positive effects on infant health at birth. By providing more transferable skills through the expansion of general education, young women may be better equipped to learn about the negative consequences of risky health behaviors. We consider several diseases related to health behavior such as diabetes (obesity), cirrhosis (alcohol abuse), lung cancer (smoking), and hypertension, which is related to mental health, smoking, and alcohol abuse (Fischer et al., 2021). We observe no significant effects on any particular type of behavior-related disease. Thus, the documented null effects of the reform on adult health outcomes suggest that its positive effects on infant health at birth are not driven by a lower engagement in risky health behavior.

Panel B of Table 2.6 reports the reform's effects on mothers' marital status and age at first birth. More transferable skills may also improve women's ability to process information about fertility options, leading to greater control over the timing of their pregnancies through effective contraceptive use. Our data show no effects on age at marriage (Column 2, Panel B) but there is a significant increase of 2.35 percentage points in the likelihood of being married at the time of the first birth, representing a 3.05% increase (Column 1, Panel B). This higher share of married mothers due to the policy change could plausibly indicate a greater ability to plan ahead for motherhood, which may also explain the improved health outcomes at birth.

2.6 Conclusions

This paper contributes to the ongoing discussion on how education can best reduce the intergenerational transmission of inequality. We exploit a unique policy shock that integrated more general education into the high school curriculum while keeping the quantity and other aspects of education quality, such as the composition of peer groups, constant. We provide causal evidence that the curriculum under

Table 2.6: Reform Effects on Health Behavior

	(1)	(2)	(3)	(4)
Panel A. Adult Health				
	Lung Cancer	Diabetes	Cirrhosis	Hyperthension
Index	-0.0424	-0.4339	0.1964	-0.2492
	(0.064)	(0.718)	(0.297)	(0.321)
1975's Cohort Mean	0.029	2.351	0.578	0.111
Std.Dev.	0.22	3.37	1.19	0.50
Obs	2,847	2,847	2,847	2,847
R-squared	0.021	0.067	0.031	0.032
Panel B. Family Planning				
	Married	Marriage Age		
Index	0.0235**	-0.0218		
	(0.011)	(0.072)		
Romano-Wolf p-value	0.043			
1975's Cohort Mean	0.768	26.466		
Std.Dev.	0.42	2.77		
Obs	1,521,770	980,853		
R-squared	0.053	0.024		

Notes. Outcomes in Panel A represent the number of hospitalizations due to lung cancer (column 1), diabetes (column 2), cirrhosis (column 3), and hypertension (column 4) for each cohort and province of residence. Outcomes in Panel B show the share of mothers who are married at the time of their first birth (column 1) and their age at motherhood in years (column 2). The estimates are obtained from estimating eq. (2.1) on a sample of female Spanish nationals aged 25-30 and born between 1975 and 1985. All specifications include a constant and main controls for birth year and province of residence. Standard errors are in parentheses. Romano-Wolf p-values based on 1,000 studentized bootstrap replications. Data from Panel A are from the 2004-2015 MSBD. Data from Panel B are from the 2000-2018 childbirth microdata from the Spanish Statistical Office. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

which mothers study also affects infant health at birth. In particular, by leveraging the staggered introduction of a comprehensive educational policy reform across Spanish provinces, we implement a differences-in-differences research design and compare the health outcomes of children born to mothers with different levels of exposure to the reform. Using detailed administrative data from birth certificates, we find that the reform led to a 27.14% reduction in the incidence of very low birth weight (less than 1,500 grams) and a 27.5% reduction in the incidence of very preterm birth (less than 33 gestational weeks).

Our findings are in line with mechanisms proposed in economic theory and previous literature on the effects of general knowledge acquisition on improved earning potential and information processing skills (Grossman, 1972; Thomas et al., 1991; Goldin, 2001). Information on mothers' occupations, marriage market allocation, and fertility choices from the birth register, together with hospital discharge records, allow us to identify possible underlying channels through which the reform may have affected children's health. Our results suggest that the observed positive effects on children's health at birth may be driven by increased maternal labor supply and better family planning, rather than an increased ability to avoid risky behaviors or increased women's earnings via different occupational choices or assortative mating.

Given the recent nature of the policy reform, the estimates of this paper are based on a relatively young sample of mothers (aged 25-33). Further research might investigate the impact of educational curricula on maternal labor, health, and social outcomes, as well as the health of their offspring, throughout their entire reproductive lives. Additionally, while we focus only on health at birth, further work could examine the effects of changes in mothers' educational curricula on children's long-term health outcomes. It would also be worth examining other variables that can serve as proxies for understanding the mechanisms through which infant health at birth is affected by changes in maternal education, such as mothers' prenatal care visits, mental disorders (e.g. anxiety or depression), or earnings.

Overall, our study shows the potential benefits of integrating more general edu-

cation into the high school system. We highlight the importance of considering the impact of educational policies, particularly those concerning school curricula, on health outcomes, as opposed to focusing solely on the quantity of education. This paper furthermore contributes to broader debate on the role of educational curricula in promoting social mobility and reducing inequality.

Chapter 3

Closing Literacy Gaps: A Personalized Technology-Aided Intervention

3.1 Introduction

A significant number of students in OECD countries complete compulsory education without achieving basic literacy skills (Vignoles, 2016; Gust et al., 2022). In 2018, 23% of 15-year-old students (over 10 million pupils) in 79 high- and middle-income countries struggled in this regard.[1] Most worrying, despite a rise in expenditure on schooling of over 15% in the past decade alone, the proportion of low performers increased between 2009 and 2018. As poor literacy skills can lead to significant economic losses and employment disadvantages in the labor market, a number of educational interventions have been implemented to enhance writing and reading skills. These include efforts to improve teachers' skills, the introduction of new literacy curricula, and the provision of personalized attention to struggling students (Slavin et al., 2011). While such actions have been successful, much less is known about how to improve literacy skills in a cost-effective and inclusive way. In this paper, we present evidence on the academic gains of a low-cost and scal-

[1]The results of PISA 2018 (Schleicher, 2019) show that only 77% of students obtained higher than level 1 proficiency in reading. See also Figure 2: https://learningportal.iiep.unesco.org/en/library/pisa-2018-insights-and-interpretations.

able computer-assisted learning (CAL) program designed to enhance the writing and reading skills of students with learning difficulties.

Educational CAL programs have emerged as a solution to overcome the long-standing challenge of managing heterogeneous learning levels within the classroom (for a review, see Escueta et al., 2020). These programs are specialized software packages that aim to improve specific skills, such as math computation or reading comprehension, and have the potential to enhance understanding through emerging artificial intelligence and machine learning techniques. CAL programs offer several advantages over traditional teaching methods. First, they can provide high-quality, engaging, and interactive content - often in the form of games - that can surpass teachers' own knowledge. Second, these programs have the potential to decrease the time gap between a student's solving of a problem and the reception of feedback. Third, by analyzing error patterns, they can precisely target content to clarify specific problem areas, making the learning process dynamically adaptive. Last, CAL programs allow for full personalization, catering to individual learning needs and preferences.

In this paper, we evaluate the CAL program called DytectiveU, developed by an independent Spanish social charity, which aims to address literacy problems in primary school-aged students through advanced cognitive modeling. Specifically, DytectiveU is an evidence-based computer game that offers over 42,000 linguistically-patterned exercises, derived from a list of 1,000+ linguistic errors and supplemented by language resources using natural language processing techniques, such as frequently used words and phonetically and orthographically similar word pairs (Rello et al., 2017a; Rello et al., 2017b). Globally adopted by over 350,000 students, DytectiveU features interactive games where students, acting as detectives-in-training, resolve linguistic challenges in sessions of around 20 minutes. A key feature of the software lies in its processing module, which personalizes instruction based on student inputs like age and past performance, aligning exercises with their cognitive abilities and learning needs. A second key feature is its dynamic adaptability through performance metrics (clicks, hits, speed, accuracy,

efficiency), which tailor future challenges, thus reinforcing weaker skills or advancing stronger ones. DytectiveU is versatile in its deployment, available in schools, after-school centers, and for home study and is platform-agnostic, supporting computers, tablets, or smartphones. It can also be used as a light-touch homework supplement to classroom curriculum.

To assess the effect of the CAL program on student achievement, we leverage the differential timing of the deployment of DytectiveU software across 308 public primary schools in the Region of Madrid (Spain). DytectiveU was initially introduced in 103 primary schools in the 2018-2019 academic year by the Ministry of Education of Madrid in collaboration with the social charity *Change Dyslexia*. It was later extended to 206 primary schools in two subsequent calls in the 2020-2021 and 2021-2022 academic years. Ideally, DytectiveU would have been assigned to schools randomly, rather than relying on schools volunteering to participate. This lack of randomization means that our sample is limited to those schools that decided to implement DytectiveU, where we compare, in the context of non-significant differences at baseline, students' performance outcomes in schools that were exposed to DyectiveU in the first call to those in schools that were exposed later. In addition, to capture actual compliance, we estimate a dose-response model and compare the performance outcomes of students in schools with a higher proportion of students logged into DytectiveU to those of students in schools with a lower proportion, and to those in schools where DytectiveU was introduced later.

To evaluate the DytectiveU software, we use population-level data on standardized testing and survey data from teachers, school heads, and families in the Region of Madrid in the 2018-2019 academic year, along with rich CAL data on the usage of DytectiveU and delivery of the instructional content. We measure student achievement using administrative records on the 2019 external standardized test results of 3rd and 6th grade students in math and language (Spanish). These tests, which are mandatory for all 3rd and 6th grade students enrolled in the Spanish education system, were administered three months after the introduction of DytectiveU across school centers. Designed at the national level, they aim to provide teachers

and parents with additional information on students' relative school performance. The standardized tests are graded blindly and not intended to be competitive in nature. We linked these administrative records on students' standardized test scores to survey data on family, teacher, and school characteristics, collected weeks prior to the tests. We also use detailed CAL data from students' interaction with DytectiveU to assess actual compliance and intensity of use, as well as to conduct descriptive research on the delivery of personalized instructional content.

We find that students in DytectiveU schools improve performance in language, with the academic gains being primarily driven by low-achievers, and positive spillover effects on mathematics performance. Specifically, we document that students in schools using DytectiveU scored between 0.08 and 0.12 standard deviations higher on the 2019 Spanish standardized test relative to students in schools that had not yet used DytectiveU. The presented estimates are intent-to-treat (ITT) given the 50% adoption rate of DytectiveU among students, reflecting a lower bound since noncompliance is ignored.[2] Using the proportion of active students logged into DytectiveU at school and grade level as a key regressor, our dose-response estimates indicate gains in Spanish ranging from 0.16 to 0.21 standard deviations. The analysis indicates that these gains are driven by low-achieving students, who are the primary target of the CAL program. Students in the bottom extremes of the distribution (5th and 10th percentiles) gained between 0.22 and 0.37 standard deviations, while no significant gains are found among students in the top extremes. We furthermore find that students in schools using Dytective also improve their mathematics performance, with estimated gains ranging from 0.14 to 0.21 standard deviations, comparable to the gains observed for Spanish. This suggests that CAL language programs could be an effective tool for subjects other than language, particularly for pupils who need significant help with reading and writing. These findings are robust to the exclusion of outliers, the imputation of mean values on family data, and missing values on test scores where family data is lacking. To provide a sense of the magnitude of our estimates, the reported gains in Spanish are equivalent to

[2]The 50% adoption rate is considered a high compliance rate given that the software was primarily intended to be used by students struggling with reading and writing.

about two months to half an academic year when compared with the conventional learning gains typically observed in national and international assessments over the course of one school year (Woessmann, 2016).

We conduct several checks to address potential identification threats that may arise from the possible self-selection of schools into the deployment and use of the CAL language program. An example of such endogeneity concerns includes omitted variable bias, such as student intrinsic motivation or faculty quality. We use demographic shocks in the previous school year as an instrumental variable (IV) to estimate DytectiveU deployment and the number of students using DytectiveU. The IV estimates confirm the effectiveness of DytectiveU in improving writing and reading skills. Second, we employ data from the prior academic year (2017-2018) to estimate the causal impact of DytectiveU using a generalized difference-in-differences (DiD) strategy. This method controls for unobserved, time-invariant differences between the schools, further confirming the validity of our estimates. Third, we test whether schools' previous performance predicts DytectiveU status or use. We find that schools' performance during the prior academic year does not predict DytectiveU deployment or the number of students logged into DytectiveU. Finally, we conduct two falsification tests, which again support the validity of our findings.

The academic gains reported above may reflect a combination of the DytectiveU CAL software and an increased focus on students struggling with reading and writing, potentially in the form of enhanced teaching strategies or greater attention from faculty. We conduct several exercises to identify these two potential channels. Specifically, we evaluate the impact of the Dytective Test detection tool (a complement of the DytectiveU CAL program designed to detect risks of reading and writing difficulties) alongside a teacher and counselor program introduced in 39 primary schools in the 2017-2018 academic year. We document that neither of these elements had a significant effect on student performance, implying that the intervention did not directly influence student outcomes through changes in teaching strategies. Meanwhile, three pieces of evidence highlight the impact of the DytectiveU software intervention: (i) the CAL data show how instructional con-

tent is individually tailored and dynamically updated with student progress; (ii) we observe uniform effectiveness across grade levels, gender, and maternal education, highlighting DytectiveU's ability to systematically educate all students; (iii) gains come from the extensive margin (number of logged-in students) rather the extensive margin (number of challenges per student), suggesting that the total sessions completed by each student may be of lesser relevance – indeed, as the content fully adapts to each student's individual needs, pupils who complete fewer sessions could achieve similar gains as those who complete a higher number of sessions.

Our findings have important implications for policy discussions on effective strategies to mitigate poor literacy skills in cost-effective and inclusive ways. Several educational interventions that use alternative approaches to teaching literacy have demonstrated promising results in improving writing and reading abilities (see Table B.1 of the Appendix). Many of these interventions focus on changing how teachers teach literacy or modifying the curriculum, such as providing a new pedagogical approach like synthetic phonics or introducing a more structured daily literacy hour (Machin and McNally, 2008; Machin et al., 2018). Others provide teachers with different instructional strategies through specific training to increase teacher quality (see, e.g., Jacob, 2017; Loyalka et al., 2019; Johnson et al., 2019; Kerwin and Thornton, 2021; Carneiro et al., 2022). Personalized curriculum interventions have also been shown to be effective, including providing content tailored to students' individual needs through tutoring (Lavecchia et al., 2020; Fryer and Howard-Noveck, 2020; Carlana and Ferrara, 2021), selective tracking based on ability (Duflo et al., 2011; Bouguen, 2016; Özek, 2021), extracurricular support (Lavy et al., 2022), or special education needs programs (Keslair et al., 2012). Yet, implementing these interventions at scale can be challenging in terms of both cost and inclusion. They often require introducing new pedagogical approaches, which may force teachers to conduct more instruction or place additional demands on their time. Alternatively, such interventions may require more teachers or volunteers to address the demand for extracurricular classes or smaller groups, which can be expensive to implement. In addition, interventions designed to address the specific

difficulties of each student by providing tailored content may be counterproductive as labelling students in order to target their specific needs can generate stigma. The uniqueness of our large-scale intervention and our high-quality administrative data on external standardized tests together with detailed information on school and student characteristics allow us to provide evidence on the short-term impacts of a novel low-cost technology-based intervention on students' academic gains. CAL programs, like DytectiveU, that provide personalized instruction without requiring teacher assistance, can offer a cost-effective solution for scaling up while avoiding the segregation of students, given that they can be implemented as a homework supplement in class or at home.

This paper adds to the growing body of literature on the effectiveness of CAL programs in enhancing academic performance, as reviewed by Bulman and Fairlie (2016) and Escueta et al. (2020). Though prior studies indicate promising results from CAL programs, only a small fraction have been subject to experimental evaluation. Most of these studies focus on CAL math programs (e.g., Banerjee et al., 2007; Mo et al., 2014; Roschelle et al., 2016; Muralidharan et al., 2019). The scarcity of experimental research on CAL language programs might stem from the market's predominant emphasis on math-focused CAL programs, which are considered better suited to personalized learning due to this subject's more objective nature. Furthermore, the limited number of studies evaluating CAL language programs, often with small sample sizes and focusing on specific schools, may raise concerns about external validity. For instance, Wijekumar et al. (2014) thoroughly evaluate a web-based reading comprehension program in Pennsylvania schools. However, the program was implemented after the state had introduced a high-speed network and a one-to-one student-to-computer ratio in 2008, factors that may affect the applicability of the study's findings. We present more information on these studies in Table B.2 of the Appendix.Our rich population-level data and the sophisticated processing module of the DytectiveU software, able to capture individual language error patterns, alleviate external validity concerns of CAL language programs and reveal their potential to enhance academic performance.

Our study also contributes to better understanding the circumstances under which these programs are most effective in improving learning outcomes. Prior studies on the effectiveness of CAL language programs offer mixed conclusions, suggesting that the characteristics of the intervention are important (for a review, see Table B.2 in the Appendix). Recently, Muralidharan et al. (2019) evaluated the deployment of Ei Minsdpark learning software across middle school grades in urban India, and observe substantial improvements in both language and mathematics. These gains appear to stem from the software's personalization and adaptivity features, aligning with earlier findings by Banerjee et al. (2007), who similarly report positive outcomes from a CAL program in a comparable urban Indian context. While CAL programs featuring advanced personalization and adaptivity might be particularly effective in developing countries, where challenges like limited availability of qualified teachers and high pupil-to-teacher ratios are common, these issues may not be as pronounced in developed countries. Our research adds to these studies by exploring the mechanisms behind the effectiveness of CAL programs. First, using DytectiveU user interaction data, we show that the software's ability to deliver personalized and adaptive instructional material are likewise a key factor in the program's success, here in a developed-country where the educational landscape differs significantly from that of developing countries. Second, we conduct a thorough review that focuses specifically on CAL language programs, with the aim of identifying the features that contribute to their effectiveness, including content personalization and adaptiveness, feedback, intervention approach, and teacher roles. Our review highlights the importance of differentiated, adaptive content, and timely feedback. Teacher involvement and intervention approach - whether the CAL program is supplementary or substitute - seem less influential when the aforementioned key features are achieved. Hence, enhancing reading skills seems to depend more on how well the program caters to each student's unique learning path, rather than on teachers' roles or implementation methods.

The rest of the paper is structured as follows: Section 3.2 describes in detail the design and deployment of the CAL language program. Section 3.3 presents the data

sources and the measurement of various outcome variables of interest. Section 3.4 discusses the two main econometric speculations. Section 3.5 sets forth our main set of results and robustness checks. Section 3.6 presents findings of several checks to validate our identification strategy and section 3.7 assesses potential mechanisms that may be driving the observed academic gains, along with a cost-effectiveness analysis. Section 3.8 concludes.

3.2 Program and Context

3.2.1 The DytectiveU CAL software

The key element of the intervention we evaluate consists of the DytectiveU CAL language software, developed by the social charity *Change Dyslexia*, an independent organization dedicated to mitigating dropout rates due to reading and writing difficulties. Specifically, DytectiveU is a scientifically validated computer-based game initially designed to improve reading and writing skills among first to sixth-grade students with dyslexia. It has since been expanded to address a broad range of literacy challenges.[3] The software includes a processing module that, based on student inputs, enables the provision of personalized instructional material. This material is adaptive to each student's progress and is complemented by immediate feedback. In what follows, we provide an overview of the content design of DytectiveU and highlight key components that make it a unique learning technology.

DytectiveU has two parts: *(i)* a web-based game for students, and *(ii)* a back-end interface for supervisors, such as school therapists, counselors, or teachers. Each student creates an avatar so that they can start playing the games. When a child begins playing, their avatar enters a "detective academy" (students are detectives-in-training), where they need to resolve linguistic challenges (sessions) of around 20 minutes consisting of a set of personalized exercises. Through the back-end interface, supervisors can visualize each student's individual performance as well as compare their overall performance with that of all users in their age group.

[3]DytectiveU is based on research led by Dr. Luz Rello at Carnegie Mellon University in collaboration with several universities. The patent derived from this research was filed on April 20, 2017 (application 15/493,060.). It is available for free here: https://www.changedyslexia.org/.

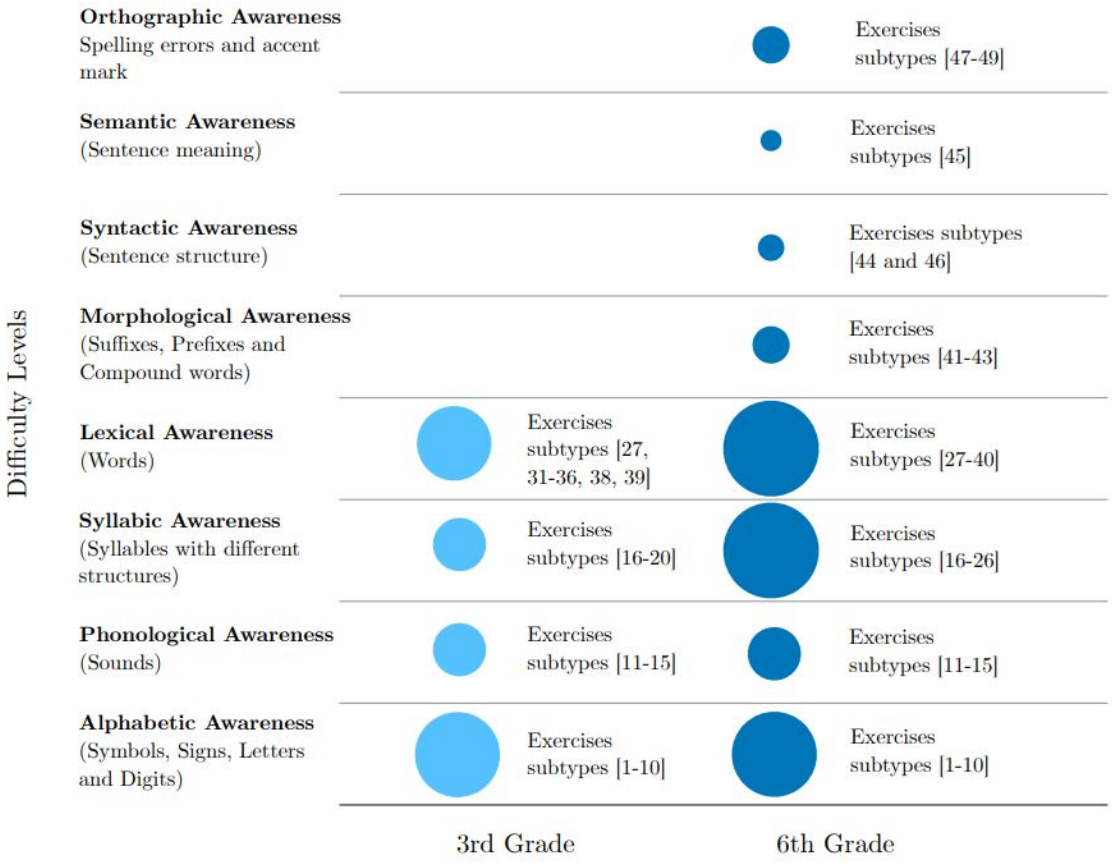
DytectiveU includes a extensive corpus of 42,000 exercises that were manually created by linguists and psychologists using two language resources: (a) patterns extracted via linguistic data mining from an existing list of 1,171 errors made by people with dyslexia (Rello et al., 2017b; Rello et al., 2017a); and (b) language resources created ad hoc using natural language processing techniques, such as lists of the most frequent words for different contexts; lists of words including word pairs that are phonological and orthographically similar; and lists of confusion sets (groups of words that tend to be mistaken for each other), among others.[4] Figure B.3 in the Appendix presents examples of the exercises in DytectiveU. The 42,000 exercises are divided into 49 subtypes, targeting different linguistic elements. The subtypes initially focus on the most basic elements (symbols, letters, sounds, and syllables), and then proceed to address more complicated linguistic units (words and non-words), eventually turning to complex parts of words (morphemes) and sentences. Table B.3 in the Appendix provides examples of exercises by linguistic level. Thus, as the user advances in a subtype, the level of difficulty increases. In addition, each of the 49 exercise subtypes has five levels of difficulty, so as the player proceeds in the game they also make headway within the subtype. While the difficulty levels increase, the linguistic elements are less frequent in number, more complex, or have a greater number of distractors. Distractors are incorrect options in a multiple-choice answer, which resemble the correct option and are meant to "distract" the player.

The software enables the provision of *differentiated instructional material* to each student, effectively addressing limitations in individual teacher knowledge and heterogeneity in learning levels within a classroom. To personalize the instructional material, DytectiveU receives as inputs: a) the age of the student; b) the number of sessions already completed using DytectiveU (if the student has used it before) and; c) the performance of the student in each of the completed sessions. The age of each user is necessary, as the exercises presented reflect the cognitive capabilities of users of their same age. The number of sessions allows to understand if

[4]This approach has been found to improve the writing skills of children with dyslexia after playing for 20 minutes, four times a week, for four weeks (Rello et al., 2014).

the student is having difficulty relative to a specific linguistic capability. Knowing the performance of the student in past sessions makes it possible for DytectiveU to personalize future exercises. To ensure that the exercises target not only the areas needing reinforcement but also those where students show proficiency, each exercise addresses at least three of 17 cognitive abilities and 7 performance measures related to literacy (as shown in Table B.4 in the Appendix). Figure 3.1 depicts how the difficulty levels of the exercises escalate in accordance with the increasing complexity of the linguistic elements for higher school grades.

Figure 3.1: Progression of Exercise Difficulty by Linguistic Awareness and Academic Year



Note: This figure shows the exercise subtypes classified by the linguistic elements they target, as well as by difficulty level and academic year. For further details on the exercises, see Table B.3 in the Appendix.

The content that DytectiveU provides is also *dynamically adaptive*. For each exercise, DytectiveU gathers a set of performance metrics: *number of clicks, hits,*

speed, accuracy;[5] and *efficiency*.[6] These metrics are subsequently mapped onto each of the cognitive abilities and the literacy performance measures (see Table B.4 in the Appendix), which are used to personalize the subsequent exercises. Hence, depending on user performance in comparison with that of the other users of their same age, the tool selects the exercises for the next exercises in order to either strengthen the weakest cognitive skills or challenge the strongest cognitive skills with more difficult exercises.

Finally, the appealing user interface, combined with *immediate feedback*, facilitates children's continuous engagement with the material. When an exercise is solved correctly, the user earns points, and the correct answer is highlighted in green. Conversely, if the answer is incorrect, it is immediately highlighted in red, with the correct answer displayed alongside to prevent discouragement and facilitate timely learning (an example of feedback provision for a specific exercise is provided in Figure B.2 in the Appendix). The more exercises a player completes within a game, the more points s/he is awarded. These points can be used to customize the avatar (Figure B.1 in the Appendix offers an example of profile customization).

3.2.2 The DytectiveU Centers Intervention

DytectiveU CAL language software was first deployed in public schools in January 2019 as the main component of a project called *Help Dyslexia* ("Ayuda a la Dislexia"), designed by the Ministry of Education of Madrid in collaboration with the social charity *Change Dyslexia*. The software was used as a supplement to traditional homework assignments rather than as a substitute for the established curriculum. It was made available to students both in school and at home, providing them with flexibility in usage. Teachers and school counselors had the option of administering DytectiveU during regular lecture hours or students could use it outside the classroom with school counselors, who are qualified psychologists, sometimes specializing in education. Additionally, students could access DytectiveU at home via computer devices or by downloading the mobile application. To maximize the full

[5] Number of *Clicks* divided by the number of *Hits*.
[6] *Accuracy* multiplied by the number of *Hits*.

potential of the software, it was recommended that students complete 64 challenges (sessions of around 20 minutes) spread over 4 times a week for 8 weeks.

The integration of DytectiveU in primary schools was also complemented with a learning difficulty detection tool, the so-called Dytective Test,[7] and a teacher and school counselor training program. The Dytective Test detection tool is a web-based game designed to spot dyslexia or the risk of other learning difficulties in an affordable and scalable way. Players complete linguistically motivated activities for around 15 minutes crafted to reveal differences between individuals with and without dyslexia.[8] DytectiveU is considered a complementary tool to the DytectiveU CAL software, and not all schools administered it to all of their students.[9] The training for teachers and school counselors, conducted by a team of speech therapists and linguists with expertise in reading disorders, consisted of seminars, videos, and the provision of a user manual and usage protocol for the computer platform. The seminars were divided into four 2.5-hour sessions. The first two covered the research behind the program, including the use of artificial intelligence to identify reading difficulties and the design of training exercises, as well as the implementation of the tool in the classroom. The third session addressed any doubts or technical questions about the program, and the final session featured testimonials from participating schools.

In the 2018-2019 academic year, DytectiveU CAL software was deployed across 107 public primary schools in the Region of Madrid. Specifically, students were granted unlimited access from mid-January 2019 until the end of the academic year. The introduction of DytectiveU followed on the heals of a previous initiative

[7]Dytective Test dectection tool is also based on research led by Dr. Luz Rello at Carnegie Mellon University in collaboration with several universities.

[8]Dytective Test is composed of a Machine Learning (ML) model (Random Forests) trained with human-computer interaction data extracted from a gamified test. It was evaluated with 5,059 participants divided into two training sets, 3,644 computer users and 1,395 tablet users. The model is able to classify people as having dyslexia or not with high sensitivity or recall, around 80% depending on the age group (Rello et al., 2020).

[9]On average, 60% of the students logged into DytectiveU used the Dytective Test detection tool. The reasons for not administering the Dytective Test included pre-existing identification of certain students who were at risk for dyslexia or already had known difficulties with reading and writing, as well as time and resource limitations. Students who had already been identified as at risk prior to the offering of the Dytective Test, those did not take the test for any other reason, or those who had a low risk of learning difficulty could still use DytectiveU if they wished to improve their skills.

in the 2017-2018 academic year, during which 39 primary schools in the Region of Madrid implemented the Dytective Test detection tool and benefited from four different seminars on the importance of tackling literacy problems at at early age. In light of high participation and involvement from educational centers and participating families, the program was extended for two additional calls, along with the deployment of DytectiveU. In 2020-2021 call, 187 primary schools participated in the program, and in 2021-2022 call, this number increased to 193 primary schools. Participation was voluntary and interested primary schools had to be public or semi-public. Enrollment consisted of a simple online procedure, requiring the contact information of the school, the responsible coordinator, and the total number of enrolled students. In addition, the school had to demonstrate that the majority of the faculty supported the initiative and that the school board had been informed.

3.3 Data

Our aim is to assess the effectiveness of DytectiveU in improving academic performance. To this end, we use population-level administrative data on standardized testing and survey data from teachers, school heads, and families in the Region of Madrid in the 2018-2019 academic year, along with school and grade level data on the use of DytectiveU. Table B.5 in the Appendix presents the datasets used for the main analysis.

3.3.1 Student Administrative Register

In the Region of Madrid, as in the rest of Spain, all students are required to take standardized tests in Spanish and Mathematics in 3rd and 6th grade.[10][11] The Spanish Government is responsible for establishing the overall framework, design, and specifications of the testing process and ensuring consistency across the country. Meanwhile, the regional governments have the authority to make any necessary adjustments to the test's design, administration, and grading. The purpose of these

[10] Regulated by The Organic Law 8/2013 (Ley Orgánica para la mejora de la calidad educativa, LOMCE).

[11] Approximately 95% of students in the Region of Madrid took the tests. Reasons for non-participation include illness, late arrival, and special education needs.

tests is to provide additional insights into student performance for teachers and parents as well as the pupils themselves. The grading of the tests is blind and the results do not have any impact on academic standing or serve as a basis for publicly ranking schools within the Region of Madrid. Standardized testing in 2019 for 3rd-graders took place on April 9th and 10th, while 6th-graders took the tests on April 29th and 30th.[12] The mathematics test consists of multiple-choice questions while the Spanish test has two parts: a linguistic competence (multiple choice) section and a written communication (written) section.[13] The children are given one hour and fifteen minutes for the Spanish language test and one and a half hours for the mathematics test.

For the main analysis, we use the 2018-2019 student register, which includes information on the student's overall scores in mathematics and Spanish in both 3rd and 6th grades, his/her gender, class group, grade level, and school. The main dependent variable is the 2019 Spanish test score standardized to have a mean of 0 and a standard deviation of 1. From the student register, we also extract the number of students per class, per grade level, and per school. Panel A of Table B.6 in the Appendix reports the definitions of the variables built from the student register.

3.3.2 Family, Teacher, and School Head Questionnaires

Prior to administering the standardized tests, school heads provided personal keys and relevant information to families and teachers so that they could complete a questionnaire using a secure system that ensured confidentiality, including the use of closed envelopes, agendas, or secure electronic communication. They also offered computer resources and assistance to those families in need. The teacher questionnaire was completed by 3rd and 6th-grade Spanish and mathematics teachers. Though school heads did not have access to the content of family and teacher questionnaires, they did receive information on completion status through the IT platform, allowing them to monitor progress in completing this task. School heads

[12] The school calendar for primary schools in Spain typically begins the first week of September and ends in June.

[13] Tests are publicly available here :https://www.comunidad.madrid/servicios/educacion/evaluacion-3o-primaria

were also required to complete a questionnaire through the same platform, in order to gather information about the availability and management of resources within the establishment. The questionnaires were developed by the Ministry of Education and Vocational Training and the Education and Research Counseling of the Region of Madrid.[14] Broadly, they aim to collect data on the socio-economic and cultural conditions of the school centers in order to contextualize the obtained results.

For the purposes of the main analysis, we use the 2018-2019 family, teacher, and school head questionnaires. The family questionnaire contains information on parental socio-economic characteristics such as parents' highest educational level and country of origin. The questionnaire also asks questions related to parental investments, such as the number of digital devices and books at home as well as pre-primary enrollment choices. The teacher questionnaire gathers information on teachers employment characteristics, such as years of experience. Notably, the school head questionnaire includes information on internet connection satisfaction, of particular interest for our analysis. The linkage between the student register and family questionnaire is done at the student ID level. The teacher questionnaire is linked to the student register and family questionnaire at the school ID and class ID levels. The school head questionnaire linkage is done at the school ID level. The variables built from the family, teacher, and school head questionnaires are reported in Panel B of Table B.6 in the Appendix.

3.3.3 DytectiveU CAL Data

The third source of data is derived from students' use of DytectiveU. The *Change Dyslexia* team gathered anonymized individual-level data on DytectiveU usage, including the number of challenges completed and their timing, difficulty level achieved by the time of the standardized test, and student age. The linkage between the Language DytectiveU CAL data and the assessment data, as well as the family, teacher, and school head questionnaires, is done at the school and grade levels due to confidentiality concerns, as it is not possible to identify students at the individual level. For the main analysis, we calculate the ratio of students who actively used

[14]The questionnaires are publicly available here:https://www.boe.es/eli/es/res/2016/03/30/(5)/dof/spa/pdf

DytectiveU before taking the standardized tests during the 2018-2019 academic year. This ratio represents the proportion of students who used DytectiveU relative to the total number of students who took standardized tests at their school and grade level. The number of students taking the external standardized tests effectively encompasses the population in these grade levels, as 95% of students participated in the compulsory standardized testing. This measure serves as a key predictor, representing the extent of DytectiveU implementation at both the school and grade levels. To observe the intensity of use, we calculate the average number of challenges completed by each school at each grade level. We use individual-level data on the number of completed challenges (sessions) and achieved linguistic difficulty level by the time of the standardized tests to descriptively observe the delivery of personalized and adaptive instructional content. Variables obtained from the DytectiveU dataset are reported in Panel C of Table B.6 in the Appendix.

For further insight on the implementation of the Language DytectiveU CAL and school and student compliance, Figure B.4 in the Appendix presents the frequency distribution of schools by grade according to two engagement metrics: the percentage of logged-in students and the number of completed sessions. While around 20% of schools reported less than 10% of their students logging in, over 37% of schools achieved login rates exceeding 70%. This high level of engagement is notable, especially considering the program's focus on students with learning challenges. On average, 46% of the students logged into DytectiveU and completed approximately 13 challenges (sessions of about 20 minutes).[15] Figure B.5 meanwhile delineates student retention across sessions (challenges), contrasting the survival or retention rates between 3rd and 6th graders. In the 3rd grade, around 20% of logged-in students completed more than 25 sessions, whereas in the 6th grade, half of that percentage did so. On average, 4.2% of students completed more than 64 challenges prior to the standardized testing, the recommended threshold for students at high risk of learning difficulties, such as dyslexia.

[15]Rello et al. (2014) documents significant improvements in writing and reading skills after completion of 16 challenges in the Language DytectiveU CAL.

3.3.4 Sample

Schools using the DytectiveU CAL program were not selected using a random or systematic process based on specific criteria. Instead, participation was voluntary. Given this lack of clear criteria or randomization in the selection of schools, we use establishments that participated later as the control group. As both the treatment and control groups participated voluntarily, they have more similarities (for the purpose of evaluating DytectiveU) than other schools that did not express interest in the program.

In order to evaluate the effects of DytectiveU on academic performance, we will examine the effects of the first phase of its deployment, i.e., schools that used DytectiveU for the first time during the 2018-2019 academic year, on performance on the 2019 standardized testing. Schools in the control group received the treatment in subsequent calls between 2020-2021 and 2021-2022. Additional information about the first phase and control groups can be found in Table B.7 of the Appendix.

We aim to compare the academic performance outcomes of schools that did and did not use DytectiveU, while controlling for various student, family, and school characteristics. The validity of the research design depends on the presence of significant differences in academic performance outcomes and observable characteristics between the treatment and control groups at the outset of the study.

Panel A of Table 3.1 shows the main performance outcomes in the pre-intervention year (2017-2018 academic year). Column (1) displays means for all primary schools in the Region of Madrid, Column (2) for schools participating between 2018-2019 and 2021-2022, Column (3) for schools participating in the first call (2018–2019), and Column (4) for schools in subsequent phases that adopted the program in either 2020-2021 or 2021-2022. We see that the treatment group (Column (3)) performs worse in both Spanish language and mathematics compared to the regional average (Column (1)). Panel B of Table 3.1 presents means of observable student and school characteristics for all primary schools, as well as for the treatment and control groups, for the 2018-2019 academic year. The treat-

ment group tends to have a lower proportion of students with college-educated parents,lower average number of books and digital devices at home (Columns (1) and (3) in Panel B) than the regional average (Column (1) in Panel B). Schools in the treatment group also tend to have a higher number of students and report worse internet connection that the regional average (Columns (1) and (3) in Panel B). However, the control group (Column (4)) is more similar to the treatment group (Column (3)) than to the regional average (Column (1)). Column (5) shows no significant differences between the performance outcomes and observable characteristics (with the exception of pre-school enrollment share) of the treatment and control groups for both samples.

Hence, in the context of non-significant differences at baseline between the treatment and control groups and for the purposes of evaluating the effects of DytectiveU on academic performance, we sample 27,571 students enrolled in 3rd and 6th grades in 270 primary schools using the program between the academic years of 2018-2019 and 2021-2022.[16] To give an idea of the relative size of the sample, it represents 23.19% of the entire school population in 3rd or 6th grades in the Region of Madrid in the 2018-2019 academic year.

3.4 Empirical Strategy

To evaluate the impact of DytectiveU on students' academic performance, we use a quasi-experimental design that leverages the timing of DytectiveU's implementation across schools. The treatment group consistes of schools that implemented DytectiveU in the 2018-2019 academic year, while the control group comprises schools that implemented DytectiveU between the 2020-2021 and 2021-2022 academic years. We exploit the staggered introduction of DytectiveU to produce both intent-to-treat (ITT) estimates of the DytectiveU program and dose-response estimates using the proportion of students actively employing DytectiveU at the school

[16] We exclude from our sample schools that had previously implemented the Dytective test detection tool and received teacher training in the prior academic year (2017-2018). This exclusion is due to the possibility that these schools might have already altered their instructional methods for students struggling with reading and writing, influenced by their earlier exposure to the complements of DytectiveU.

Table 3.1: Differences between the Treatment and Control Groups

	(1) All primary schools	(2) Full Sample	(3) Treatment Group	(4) Control Group	(5) P-Value (3) - (4)
Panel A. Students' Performance					
Standardized Score in Spanish	0.006	-0.175	-0.140	-0.191	0.475
Standardized Score in Maths	0.006	-0.224	-0.188	-0.242	0.619
Panel B. Students' Characteristics					
Frac. Students Started After 3yo	0.252	0.282	0.252	0.297	0.010
Frac. Female Students	0.485	0.479	0.480	0.479	0.847
Frac. Inmigrant Students	0.043	0.047	0.040	0.050	0.139
Frac. College Mother	0.541	0.438	0.441	0.437	0.892
Frac. College Father	0.466	0.350	0.341	0.355	0.672
Books at Home					
Frac. Less than 50	0.273	0.323	0.311	0.329	0.397
Frac. More than 50 and less than 50	0.268	0.281	0.288	0.277	0.423
Frac. More than 100	0.257	0.214	0.214	0.213	0.989
Frac. More than 5 Digital Devices at Home	0.558	0.518	0.511	0.522	0.556
Panel C. Schools' Characteristics					
School Size (Number of Students)	134.840	125.699	127.038	125.057	0.835
Class Size (Number of Students)	24.635	24.483	24.408	24.518	0.646
Frac. Students in 3rd Grade	0.502	0.504	0.507	0.503	0.783
Internet Connection					
Severe internet inconvenience	0.117	0.210	0.199	0.215	0.790
Moderate internet inconvenience	0.199	0.295	0.264	0.310	0.466
Mild internet inconvenience	0.239	0.269	0.302	0.253	0.445
No internet inconvenience	0.360	0.198	0.235	0.180	0.368
School Location					
Capital	0.417	0.352	0.386	0.336	0.471
East	0.138	0.221	0.210	0.226	0.785
North	0.067	0.079	0.080	0.079	0.983
West	0.120	0.073	0.064	0.078	0.684
South	0.257	0.275	0.260	0.282	0.739
Frac. Teacher More than 10 years of Experience	0.628	0.666	0.691	0.654	0.443
Number of Students	134,501	27,571	8,939	18,632	
Number of Schools	1,279	270	89	181	

Notes. Columns (1), (2), (3), and (4) show means in outcomes variables (Panel A), student characteristics (Panel B), and school characteristics (Panel C). Column (1) reports data for a sample of students in 3rd and 6th grades from all primary schools in the Region of Madrid, Column (2) for a sample of students in 3rd and 6th grades from primary schools that implemented DytectiveU between the 2018-2019 and 2021-2022 academic years, Column (3) for a sample of students in 3rd and 6th grades from primary school included in the treatment group (schools implementing DytectiveU in 2018-2019 academic year), and Column (4) for a sample of students in 3rd and 6th grades from primary schools included in the control group (schools implementing DytectiveU between the 2020-2021 and 2021-2022 academic years). Scores in Spanish language and mathematics are standardized to have a mean of 0 and a standard deviation of 1 (Panel A). Column (4) reports p-values of the differences between treatment (Column (3)) and control (Column (4)) groups. P-values in Panel A are calculated controlling for student and school characteristics. Source: The data on students' scores in Spanish language and mathematics come from the 2017-2018 assessments and the data on student and school characteristics come from the 2018-2019 family and teacher questionnaires.

and grade levels. The ITT estimates provide insight into the overall impact of the program, while the dose-response estimates identify the effects of the actual treatment exposure.

Intent-to-Treat Estimates

We estimate the ITT effects of implementing DytectiveU (β_1) with an ordinary least squares (OLS) model, as shown in the following equation:

$$y_{\text{isc}} = \alpha + \beta_1 Treat_s + X_i'\beta_2 + Z_s'\beta_3 + C_c'\beta_4 + \varepsilon_{isc} \tag{3.1}$$

Where y_{isgc} is the standardized score on the 2019 Spanish test of student i, in school s, and class group c. $Treat_{sg}$ is the key regressor and is equal to 1 for treatment schools (i.e., schools implementing DytectiveU in the 2018-2019 call) and 0 for controls schools (i.e., schools implementing DytectiveU in the 2020-2021 and 2021-2022 calls). The coefficient of interest is β_1, which represents ITT effects for all students in our sample, regardless of whether or not they actually used DytectiveU . X_i is a vector of personal/family characteristics that consists of the following variables: gender, age, early enrollment in the education system (prior to age 3), foreign birth, parental college degree attainment, number of books and digital devices at home. In addition, the specification controls for a vector of school characteristics (Z_s), including school size, location, and reported internet connection, and a vector of class group characteristics (C_c), which includes class size and teacher's years of experience. Standard errors are robust and clustered at the school level.

Dose-Response Estimates

The ITT estimates are based on an average participation rate of approximately 50% among students in treated schools. However, to accurately assess the impact of implementing DytectiveU and estimate the expected treatment effects at varying levels of the CAL program engagement (with further assumptions), we establish a dose-response relationship between the number of students who actively use DytectiveU and academic gains using:

$$y_{\text{isgc}} = \alpha + \beta_1 Coverage_{sg} + X_i'\beta_2 + Z_s'\beta_3 + C_c'\beta_4 + \varepsilon_{isgc} \tag{3.2}$$

Where $Coverage_{sg}$ is the proportion of students actively using DytectiveU at the school and grade level, which is zero for control schools, while all other variables are defined as in equation (3.1). The coefficient of interest, β_1, reflects the average academic gain resulting from a unit increase in DytectiveU adoption. Using these dose-response estimates to predict the effects of varying the proportion of students using the CAL program requires further assumptions about (i) common treatment effects between control and treatment groups as we identify responses over a subset of compliers and not the full sample (we give 0 dose to control schools) and, (ii) the linearity of the functional form of the relationship between the number of students using DytectiveU and the average academic gains (since we expect the average academic gains to increase or decrease proportionally with the level of exposure).

Following Muralidharan et al., 2019, we provide additional suggestive evidence that common treatment effects between the control and treatment groups may be a reasonable assumption in this setting. Figure B.6 of the Appendix presents the kernel-weighted local mean smoothed plots that relate the 2019 test scores in Spanish to percentiles in 2018 test scores at the school and grade levels separately for the treatment and control groups, with 95% confidence intervals. We see that the relationship lines between the 2018 and 2019 test scores are within the confidence intervals, indicating that there is no significant difference between the treatment and control group. The treatment group outperforms the control group only in the lower half of the baseline distribution, with no strong evidence of differential absolute magnitude gains across the second half of the distribution. This is consistent with the program's intent to target low-achieving students, with the lower-performing treated schools at baseline being the most likely to benefit from the program. This graphical evidence thus implies that the common treatment assumption holds in our setting.

To investigate the relationship between the number of students using Dytec-

tiveU and their academic gains, we conduct both graphical and analytical analyses to explore the linearity of the functional form. The graphical analysis presented in Figure B.7 suggests a linear relationship, which is further supported by the lack of improvement when adding a quadratic term found to be statistically insignificant (see quadratic estimates in Table B.9 of the Appendix and estimates for the linear relationship in Panel B of Table 3.2). Moreover, the adaptive nature of the intervention implies that DytectiveU can be equally effective regardless of students' initial learning levels or rates of academic progress. This further supports the plausibility of a linear dose-response relationship. Note also that DytectiveU was implemented three months prior to the standardized test, making it plausible that the program did not experience diminishing returns over the relatively short duration of the treatment in this study.

3.5 Results

3.5.1 Effects on Spanish Language Performance

Table 3.2 shows whether DytectiveU had an effect on students' performance on the 2019 standardized Spanish test for 3rd and 6th grades. Panel A shows ITT estimates of eq. (3.1), where the key regressor is a dummy variable equal to 1 for students enrolled in schools that started using DytectiveU in 2018-2019 (first call) and 0 for students enrolled in schools that started using DytectiveU in subsequent calls (2019-2020 and 2021-2022). Panel B shows dose-response estimates of eq. (3.2), where the key regressor is the fraction of students logged into Language DytectiveU by grade and school. We show specifications without controls (Column (1)) and including controls (Column (2)) for student characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of digital devices and books at home), school and class size, school's location and reported internet connection, and teacher's years of experience.

In Panel A, we show that students enrolled in schools that used DytectiveU scored between 0.09 and 0.12 standard deviations higher on the 2019 standard-

ized Spanish test than those in schools that adopted DytectiveU in later academic years. In Panel B, we consider the effects of a higher adoption of DytectiveU: performance on the standardized Spanish test improves on average between 0.18 and 0.23 standard deviations in response to a 1 percentage point increase in the fraction of students using DytectiveU. In each case, the point estimates of the effects are slightly lower when adding controls. The magnitude and significance of the coefficients of interest are also somewhat reduced when including controls for observable student, parent and school characteristics (Column (2)). Note that the sample is downsized when including controls, as the family questionnaire response rate is about 50%. Table B.8 in the Appendix replicates the results shown in Column (1) for the sub-sample of students without missing questionnaire data and reports similar coefficients in terms of significance, size, and sign as the main results shown in Table 3.2. Hence, the low response rate to the family questionnaire does not seem to limit the generalizability of the main findings.[17] In sum, for both treatment measures, we find positive effects of DytectiveU on Spanish language performance. The dose-response estimates are higher than the ITT estimates (18.44% vs 9.41% of a standard deviation). As expected, the estimates in Panel A are more conservative than those in Panel B as the ITT estimates ignore noncompliance. In other words, the effect of DytectiveU on students' performance is higher as the coverage (i.e., larger as the number of students using DytectiveU) increases. These results provide further evidence that DytectiveU is effective at bolstering reading and writing skills.

To provide a sense of the magnitude of our estimates, we compare them with the conventional learning gains observed in most national and international assessments over one academic year. These gains are usually between a quarter and one third of a standard deviation (Woessmann, 2016). Hence, the reported improvement in Spanish performance ranging from 9.41% to 23.23% of a standard deviation is equivalent to two to seven months of schooling. Our findings are in line with previous literature on successful CAL reading and spelling programs (Escueta et al., 2020). Although Rouse and Krueger (2004) and Borman et al. (2009) find weak and

[17] These results are robust to imputation analysis for missing values, see Section 3.5.4.

insignificant effects of the Fast ForWord program in the US due to a lack of adaptive content and fast feedback, two other US studies (Wijekumar et al. (2012) and Wijekumar et al. (2014)) evaluate the Intelligent Tutoring for the Structure Strategy (ITSS) reading comprehension program and find significant positive effects ranging from 0.2 to 0.53 standard deviations on a series of comprehension texts. In Canada, Deault et al. (2009) assess the ABRACADABRA web-based literacy program and report an improvement of 0.35 standard deviations in reading. In the developing world, Muralidharan et al. (2019) look at a CAL reading program called Mindspark in Delhi and report an increase of 0.22 standard deviations in Hindi language reading ability.

Table 3.2: Effects on Spanish Language Standardized Test

	(1)	(2)
Panel A. Intent-To-Treat Estimates		
Treat	0.1207**	0.0941*
	(0.050)	(0.051)
R-squared	0.003	0.119
Panel B. Dose-Response Estimates		
Coverage	0.2323***	0.1844**
	(0.076)	(0.076)
R-squared	0.005	0.120
Mean outcome in control schools	-0.206	-0.001
Sample Size	22,430	8,810
Number of Schools	269	214
Controls	NO	YES

Notes. Outcome variable: standardized score on Spanish test taken at the end of the 2018-2019 academic year. The unit of observation is student i in school s, grade c, and class group c. Intent-to-treat estimates from eq. (3.1) are shown in Panel A. Treat is a dummy variable equal to 1 for students in the treatment schools. Dose-response estimates from eq. (3.2) are shown in Panel B. Proportion of students using DyectiveU is the fraction of students logged into DytectiveU by grade and school. Controls include student characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection, and teacher's years of experience. Table B.8 of the Appendix replicates the results reported in column (1) excluding from the sample students with missing family and teacher questionnaire data. Standard errors (in parentheses) are robust and clustered at the school level.***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

3.5.2 Distributional Effects on Spanish Language Performance

In this section, we explore whether the gains from using DytectiveU differ across the distribution of Spanish test scores. One might expect that the students who most benefit from using DytectiveU are those with lower scores on the Spanish test, indicating that may have difficulty with reading and writing. Table 3.3 displays results from a quantile regression using the standardized score obtained on the Spanish test for the full sample. The results indicate that DytectiveU has a non-uniform effect throughout the distribution, its impact being higher in the first half of the distribution. The point estimates for the ITT estimates (Panel A) are higher in magnitude and significance at the bottom extremes (5th and 10th percentiles) compared to the median and above. The dose-response estimates (Panel B) show a similar pattern. Two potential non-exclusive explanations can be given for this pattern. One is that low-achievers use DytectiveU more intensively. Another is that there is a higher proportion of low-achievers logged into DytectiveU. Although it is unclear whether the cause is the higher number of low-achievers using DytectiveU or the higher usage of DytectiveU by these students, these findings lend further support to the causal effects shown in Table 3.2. Broadly, this exercise suggests that the intervention is particularly effective at increasing performance in Spanish language among low-achievers, who were the target of DytectiveU.

3.5.3 Effects on Mathematics Performance

Although the program focuses on providing personalized linguistic exercises to ameliorate writing and reading difficulties in Spanish, it might also impact on other subjects, such as mathematics. Machin and McNally (2008) find that the reading demands of a math test for 11-year-olds were nearly 70% of those of a dedicated reading assessment, based on text difficulty. We accordingly test outcomes for mathematics, as standardized math tests were administered on the same day and in the same setting as the Spanish tests.

Table 3.4 presents the results for mathematics performance, where the outcome variable is the standardized score on the 2019 mathematics test for 3rd and 6th graders, which took place on the same day and in the same setting as the Span-

Table 3.3: Distributional Effects on Spanish Language Standardized Test

	(1) .05Q	(2) .1Q	(3) .25Q	(4) .5Q	(5) .75Q	(6) .9Q	(7) .95Q
Panel A. Intent-To-Treat Estimates							
Treat	0.2364***	0.1704***	0.0928	0.0884	0.0584	0.0318	0.0109
	(0.066)	(0.058)	(0.058)	(0.055)	(0.050)	(0.065)	(0.070)
R-square	0.105	0.114	0.117	0.118	0.116	0.113	0.104
P-value Parente-Santos Silva test	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel B. Dose-Response Estimates							
Coverage	0.4072***	0.2994***	0.2296**	0.1803**	0.0964	0.0619	-0.0193
	(0.095)	(0.087)	(0.095)	(0.075)	(0.078)	(0.096)	(0.115)
R-square	0.109	0.113	0.118	0.119	0.116	0.113	0.103
P-value Parente-Santos Silva test	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Number of Students	8,810	8,810	8,810	8,810	8,810	8,810	8,810
Control Variables	YES	YES	YES	YES	YES	YES	YES

Notes. Outcome variable: standardized score on the 2019 Spanish test. The unit of observation is student *i* in school*s*, grade level *g*, and class group *c*. Intent-to-treat estimates from eq. (3.1) are shown in Panel A. Treat is an intention to treat dummy equal to 1 for students in the treatment schools. Dose-response estimates from eq. (3.2) are shown in Panel B. Proportion of students using DyectiveU is the fraction of students logged into DytectiveU by grade and school. All regressions include controls for student characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection, and teacher's years of experience. Standard errors (in parentheses) are robust and clustered at the school level. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

ish test. Although both the ITT and dose-response estimates for mathematics show a positive pattern, the effects are only statistically significant for the dose-response specification (Panel B). That latter indicates that, on average, a 1 percentage point increase in the proportion of students using DytectiveU increases the test scores in mathematics by 0.19 standard deviations. Since the reading and writing demands of a mathematics test are lower than those of a Spanish test, the results for mathematics are not necessarily driven by low-achievers in mathematics. Indeed, students who struggle in Spanish may be mid or top-performers in mathematics. Table B.10 in the Appendix reports the results from the quantile regression using the standardized score obtained on the mathematics test. DytectiveU does not have a uniform

effect across the entire math distribution and, in contrast to our estimates for Spanish (Table 3.2), its effects seem to be higher in the second half of the distribution. This pattern of results is consistent with our main results for Spanish, and supports the idea that DytectiveU is highly effective at not only increasing performance in Spanish, but also improving performance in other subjects such as mathematics.

Table 3.4: Effects on Mathematics Standardized Test

	(1)	(2)
Panel A. Intent-To-Treat Estimates		
Treat	0.0707	0.0788
	(0.057)	(0.059)
R-squared	0.001	0.102
Panel B. Dose-Response Estimates		
Coverage	0.1473*	0.1962*
	(0.088)	(0.108)
R-squared	0.002	0.104
Mean outcome in control schools	-0.2382	-0.2392
Sample Size	22,685	9,680
Number of Schools	269	214
Controls	NO	YES

Notes. Outcome variable: 2019 standardized score in mathematics. The unit of observation is student i in school$s$, grade level g, and class group c. ITT estimates from eq. (3.1) are shown in Panel A. Treat is a dummy equal to 1 for students in the treatment schools and 0 for students in control schools. Dose-response estimates from eq. (3.2) are shown in Panel B. Proportion of students using DyectiveU is the fraction of students logged into DytectiveU at the grade and school levels. All regressions include controls for student characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection, and teacher's years of experience. Standard errors (in parentheses) are robust and clustered at the school level. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

3.5.4 Robustness Checks

Finally, we examine whether our results for students' academic performance are robust to several reasonable changes in the estimation strategy. First, we test whether the findings are driven solely by the differences between highly and lowly involved schools. To this end, we exclude outliers from our analysis and replicate our main results (reported in Table 3.2). The results of this exercise are reported in Table

B.11 in the Appendix. We identify outliers as the top and bottom 10% of treated schools based on the proportion of students who logged into DytectiveU. The results are robust to dropping schools with exceptionally high or low usage of the program, suggesting that our main findings are not driven by a small number of extreme cases. Second, we check whether the estimated results for students' Spanish performance are sensitive to the imputation of missing values on family data. Given that the family questionnaire response rate was approximately 50% and not significantly different between the treatment and control groups, missing data are a potential concern that could affect the validity and reliability of our estimates. To address this issue, we consider two different imputation approaches: (i) we assign the mean values of the corresponding control variable at the school level to the students with missing family data and; (ii) we assign missing values to outcomes from students with missing family data. The results of these imputation analyses are reported in Appendix Tables B.12 and B.13, respectively. We find that our main results reported in Table 3.2 are very similar when missing values are imputed using either approach. Specifically, the estimated treatment effects and standard errors are almost identical across the different specifications, indicating that the imputation has little or no impact on the average results. Our main findings are thus robust to potential biases that may arise from missing data, lending further support to the validity and reliability of our estimates.

3.6 Threats to Identification

Thus far, we have shown that DytectiveU improved students' performance on the Spanish test and had positive spillover effects on their mathematics performance. This was particularly true for low achievers in Spanish, the intended target of DytectiveU. However, the fact that neither allocation of the DytectiveU software nor the proportion of student engagement with the software was randomized, raises endogenity concerns. Omitted variable bias, such as student intrinsic motivation or faculty quality, may have simultaneously influenced a school's decision to adopt DytectiveU and the academic performance of its students. This could bias our es-

timates upwards, particularly if schools in the treatment group, which adopted DytectiveU earlier, are inherently more proactive in addressing literacy challenges and used the program more intensively compared to those in the control group, which adopted the program later. In what follows, we conduct a number of identification checks in order to evaluate the validity of our estimates.

First, we assess the effectiveness of DytectiveU by considering a factor that influenced the proportion of students using the program during the 2018-2019 academic year, but that had no direct impact on their performance on standardized tests during that same year. Specifically, following Keslair et al. (2012), we examine demographic shocks in the previous academic year (2017-2018) that indirectly effected the proportion of students using DytectiveU in the 2018-2019 academic year, but did not directly affect the performance of that year's group. To this end, we use the proportion of students born between November and December within the previous year group as our instrumental variable (IV). In the Spanish system, students born at the end of the calendar year are typically the youngest in their grade level and may be more prone to experiencing early learning difficulties(Crawford et al., 2007; Dhuey and Lipscomb, 2010). It can consequently be inferred that a cohort with a higher proportion of students born between November and December may require a greater allocation of resources due to having more pupils vulnerable to this challenge. Figure B.8 in the Appendix displays the distribution of the standardized scores from the 2018 Spanish test separately by schools with a low, mid, and high proportion of students born at the end of the year. As observed in Figure B.8, the test score distribution of schools with a high proportion of students born at the end of the year is to the left of the distribution of schools with a low proportion of students born at the end of the year. In other words, schools with a higher proportion students born between November and December perform worse in Spanish compared to schools with relatively older students. Under the assumption that the time resources are fixed across academic years, adverse shocks in the previous year reduce access to DytectiveU, as schools that had a higher proportion of younger students the year prior may have less available time for students needing language

support due to the necessity of addressing the previous year's challenges. This is a plausible assumption as the adoption of DytectiveU is not associated with increases in school budgets or the allocation of more guidance counselors. The overall level of resources directed towards remedial education in a given school is determined at the regional level and is stable over time.[18] This factor therefore provides a interesting tool for identifying the overall effect of the proportion of students using DytectiveU on their average performance.

The results of our IV estimates for the effectiveness of DytectiveU are presented in Table 3.5. The F-stats from the first stage of Panel B confirm that our demographic shock instrument–the proportion of students born between November and December in the previous year's group–is a valid instrument for estimating the share of pupils using DytectiveU. This is supported by the negative and highly statistically significant IV coefficients, which indicate that schools with a higher proportion of younger students tend to have a lower proportion of pupils logged into DytectiveU. However, as shown in Panel A, the instrument is not as effective for estimating the decision to implement the software earlier. This discrepancy could be due to the fact that the proportion of younger students in the previous year's group, while related to the share of students using DytectiveU, is not a perfect proxy for the decision to implement DytectiveU. The latter is mainly made by the school head and can be influenced by a variety of factors not related to the number of struggling students, such as administrative priorities.

Our results from the second stage regressions, which measure the impact of DytectiveU on performance on the standardized Spanish test, remain statistically significant and have the same sign as our main results presented in Table 3.2. Specifically, in response to a 1 percentage point increase in DytectiveU participation, performance on the standardized Spanish test improves by 1.36 units of standard deviation (Column (2), Panel B). Overall, these results confirm that DytectiveU is effective at improving writing and reading skills, and support the use of our de-

[18]Typically, one Educational Guidance and Psycho-Pedagogical Team (*Equipo de Orientación Educativa y Psicopedagógica*) is shared among several pre-primary and primary schools, and the time allocated for remedial education for each school does not vary substantially from one year to the other (Commission, 2022).

mographic shock instrument as a valid IV for estimating the proportion of students using DytectiveU.

These IV estimates may provide an upper limit to the effect of DytectiveU for two non-exclusive reasons. First, as indicated by Keslair et al., 2012, adverse shocks in previous years may also limit access to other resources not directly related to students with special language needs. Second, it is possible that the IV estimates exceed the OLS estimates because the IV is able to capture the Local Average Treatment Effect (LATE)[19] of DytectiveU on student performance. In our OLS estimates, the treatment effect is assumed to be the same for all students, while our IV estimates capture the effects of DytectiveU on a specific subgroup of students who may benefit more from using it, i.e., students from schools that necessarily had to use the software less intensively due to a lower availability of resources caused by having a higher proportion of struggling students in previous years.[20]

A second way to investigate potential selection bias within the sampled primary schools is to use data from the prior academic year (2017-2018) to estimate the causal impact of the intervention using a generalized difference-in-differences (DiD) strategy. The strategy compares the before-after difference in outcomes between students in schools where DytectiveU was introduced and students in schools that did not change their DytectiveU status between the two academic years (2017-2018 and 2018-2019). We estimate the following specification:

[19]This concept was first introduced by Angrist and Keueger (1991) in their study of changes in compulsory schooling laws, where they found that the effect of extending obligatory schooling on student outcomes varied depending on the characteristics of the students and their families.

[20]Previous research finds that IV estimates tend to be larger than OLS estimates. Jacob and Lefgren (2004) observe somewhat larger IV estimates than OLS estimates for the effect of summer school and grade retention on student achievement. Onda and Seyler (2020) examine the process of reclassifying English learner students as English proficient in Minnesota and find that low compliance rates in school districts as they gradually adopted the state's reclassification policy led to larger instrumental variable (IV) estimates. Eren et al. (2017) analyze the effects of summer school and grade retention on high school completion and juvenile crime in Louisiana using a fuzzy regression discontinuity design (RDD). They find that being retained in fourth or eighth grade increases the probability of dropping out by 11% and 10% respectively, and that the LATE is larger for retention. The LATE estimates of actual grade retention on the number of additional grades attained are also higher.

Table 3.5: Identification Check #1. Addressing the non-random assignment and adoption of DytectiveU. IV of the impact of the proportion of students born at the end on the calendar year in the previous academic year on Spanish performance.

	(1)	(2)
Panel A. Intent-To-Treat Estimates		
Treat	1.7429	1.6123
	(1.243)	(1.501)
Cragg–Donald F stat	2.654	1.841
First-Stage Regression Results for Treat		
Prop. Students born at the end of academic year in 2017-2018	-1.0901**	-0.7536
	(0.457)	(0.695)
Panel B. Dose-Response Estimates		
Coverage	1.5198**	1.3672**
	(0.719)	(0.685)
Cragg–Donald F stat	9.352	10.362
First-Stage Regression Results for Coverage		
Prop. Students born at the end of academic year in 2017-2018	-0.8595***	-0.9813***
	(0.235)	(0.344)
Sample Size	13,013	5,388
Controls	NO	YES

Notes. Outcome variable: 2018 standardized score in Spanish test. The unit of observation is student i in school s, grade c and class group c. Controls include students' characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection and teacher's years of experience. Standard errors (in parentheses) are robust and clustered at school level. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

$$Y_{isct} = \alpha + \beta_1 \text{Treat}_s + \beta_2 \text{Post}_t + \beta_3 \text{Treat}_s \times \text{Post}_t + X_i' \beta_4 + Z_s' \beta_5 + C_c' \beta_6 + \theta_s + \delta_t + \varepsilon_{isct} \tag{3.3}$$

where Y_{isct} is the standardized score in Spanish Language Test in t academic year of student i, that belongs to school s, class group c. Treat_s is a indicator variable equal to 1 if DytectiveU was available at some point between 2018-2019 and 2021-2022 academic year in school s and; Post_t is a indicator variable equal to 1 if the date is after 2017-2018 academic year. X_i is a vector of personal characteristics

including gender, early enrollment in the education system, parental college degree attainment and number of books at home. Z_s is a vector for school characteristics including school size and location and, C_c is a vector of class group including class group size. Note that the reduced number of control variables, in comparison to equation (3.1), is attributed to data unavailability from family, school, and school head records during the 2017-2019 academic years. We include school (θ_s) and year δ_t fixed effects and cluster standard errors at the school level.

This method controls for unobserved, time-invariant differences between the schools, which might otherwise bias the results. β_4 is the parameter of interest and captures the discontinuity in Spanish language performance caused by the deployment of the DytectiveU software, while controlling for both observed and unobserved inherent differences between students from schools that use the software earlier versus later, independent of the intervention.

Table 3.6 provides a comparative analysis of the effects of the DytectiveU program on standardized Spanish language test scores using both the DiD and OLS models, estimated from equations (3.3) and (3.1) respectively, and including the available controls for 2017-2018 academic year. The DiD model shows a significant positive effect of the DytectiveU software on the full sample (Column 1) and better estimated positive impact on the bottom 20% of the score distribution (Column 2),which was the target group of the DytectiveU software and where the distributional gains stem from, as shown in Table 3.3. The Post coefficient's lack of significance suggests no general time trends affecting the scores. The OLS model (Columns 3 and 4), presented for contextual comparison, aligns closely with the DiD results further confirming the validity of our estimates.

A third approach to assessing potential selection bias involves examining whether the implementation and intensity of DytectiveU usage are associated with prior academic performance. To investigate these two possibilities, we follow Ferrara et al. (2012) and perform two exercises. First, to test whether previous performance predicts the implementation of DytectiveU, we aggregate the data at the school level and we estimate the following specification:

Table 3.6: Identification Check#2: Effects on Spanish Language Standardized Test: DiD and OLS Analysis

	(1)	(2)	(3)	(4)
	DiD Model		**OLS Model**	
	Full Sample	Bottom 20%	Full Sample	Bottom 20%
Treat	0.7397***	0.1421***	0.1069**	0.0810***
	(0.036)	(0.035)	(0.051)	(0.025)
Post	-0.0263	-0.0454*		
	(0.034)	(0.026)		
Treat*Post	0.0905*	0.0819**		
	(0.049)	(0.037)		
R-squared	0.190	0.125	0.112	0.028
Sample Size	21,035	3,105	9,671	1,496
Number of Schools	270	259	224	210

Notes: The outcome variable is the standardized score in the Spanish language test. Columns (1) and (2) present estimates derived from equation (3.3), while Columns (3) and (4) report estimates based on equation (3.1). Columns (2) and (4) narrow the focus to the bottom 20% of the distribution in Spanish language performance. The analysis includes a comprehensive set of controls: students' characteristics like gender and pre-primary enrollment, parents' educational levels, parental investments indicated by the number of books at home, and school-related factors such as size, class size, and location. Standard errors, reported in parentheses, are robust and clustered at the school level. Statistical significance is denoted as follows: *** for the 1% level, ** for the 5% level, and * for the 10% level.

$$Treat_s^{2019} = \alpha + \beta_1 Performance_s^{2018} + X^{2018}{}_s\prime\beta_3 + Z^{2018}{}_s\prime\beta_3 + \varepsilon_s \tag{3.4}$$

where $Treat_s^{2019}$ is equal to 1 schools participating in the first call and 0 for school participating later.$Performance_s^{2018}$ if the average standardized score in 2018 Spanish test of students enrolled in s school and g grade level; $X^{2018}{}_s$ is a vector of personal/family characteristics, measured for s school and g grade level in 2017-2018 academic year; $Z^{2018}{}_s$ is a vector of school characteristics, measure for s school and g grade level; and ε_{sg} is the error term. Standard errors are robust and clustered at school level.

Second, to test whether previous performance predicts the use of DytectiveU we aggregate the data at school and grade level and use the following specification:

$$Coverage_{sg}^{2019} = \alpha + \beta_1 Performance_{sg}^{2018} + X^{2018}{}_{sg}\prime\beta_3 + Z^{2018}{}_{s}\prime\beta_3 + \varepsilon_{sg} \quad (3.5)$$

where $Coverage_{sg}^{2019}$ is the proportion of students using DytectiveU in 2018-2019 academic year enrolled in school s and grade level g. All other variables are defined as in equation (3.4).

The results of these exercises are displayed in Table 3.7. Columns (1) and (2) show that there is no correlation between a primary school's participation in the first call and their average performance in the Spanish standardized test of the previous academic year (2017-2018), even when controls are included. Similarly, Columns (3) and (4) reveal that there is no correlation between the proportion of students using DytectiveU and the average performance in the 2018 Spanish standardized test when adding controls. Thus, results reported in Table 3.7 suggest that higher performing schools do not appear to implement DytectiveU earlier or exert a higher use of DytectiveU.

We also assess the possibility of selection bias two performing falsification tests. First, we randomly assign treatment to schools and create a simulated Treat variable based on this random assignment. We then re-estimate the benchmark model, as outlined in equation (3.1), using this simulated variable, and record the estimates. We repeat this process 500 times. The resulting empirical cumulative distribution function and density of the estimated coefficients on the placebo *Treat* variable are depicted in Graphs (1) and (2) of Figure 3.2. As expected, the distribution of the estimated coefficients on the simulated DytectiveU software coverage assignment centers around zero. Furthermore, the lowest benchmark estimate from Column 1, Panel A, of Table 3.2 (indicated by a vertical line at +0.0926) falls outside of the range of coefficients estimated in the simulation exercise, providing further evidence for the robustness of the results.

The same exercise is repeated for DytectiveU coverage by randomly re-

Table 3.7: Identification Check #3: Possible Correlation Between Prior Academic Performance

	(1)	(2)	(3)	(4)
	Treated in 2019		Coverage	
Spanish Standardized Score	0.0665	0.1321	0.0736	0.0791
	(0.068)	(0.084)	(0.051)	(0.052)
Observations	262	261	314	307
R-squared	0.004	0.052	0.012	0.046
Controls	NO	YES	NO	YES
Number of Schools	262	261	87	86

Notes. The unit of observation in Columns (1) and (2) is school s and in Columns (3) and (4) is school s and grade level g. The outcome variable in Columns (1) and (2) is a dummy equals to 1 for students in the treatment schools, and in Columns (3) and (4) is the proportion of students logged in DytectiveU at school and grade level. Controls include students' gender and age, parents' educational level, number of books at home, school and class group size. Information on students' pre-primary enrollment, immigrant status, school's internet connection and teacher's years of experience is not available for 2017-2018 questionnaires. The sample is restricted to treated schools for estimates reported in Columns (3) and Columns (4). Standard errors (in parentheses) are robust and clustered at school level. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

distributing the proportion of students actively using DytectiveU across treated schools. Similar to the placebo test for the ITT estimates, the distribution of the estimated coefficients on the simulated DytectiveU coverage centers is around zero (Graph (4), Figure 3.2). These results provide additional assurance of the validity of the estimated treatment effects and suggest that our findings are not spurious.

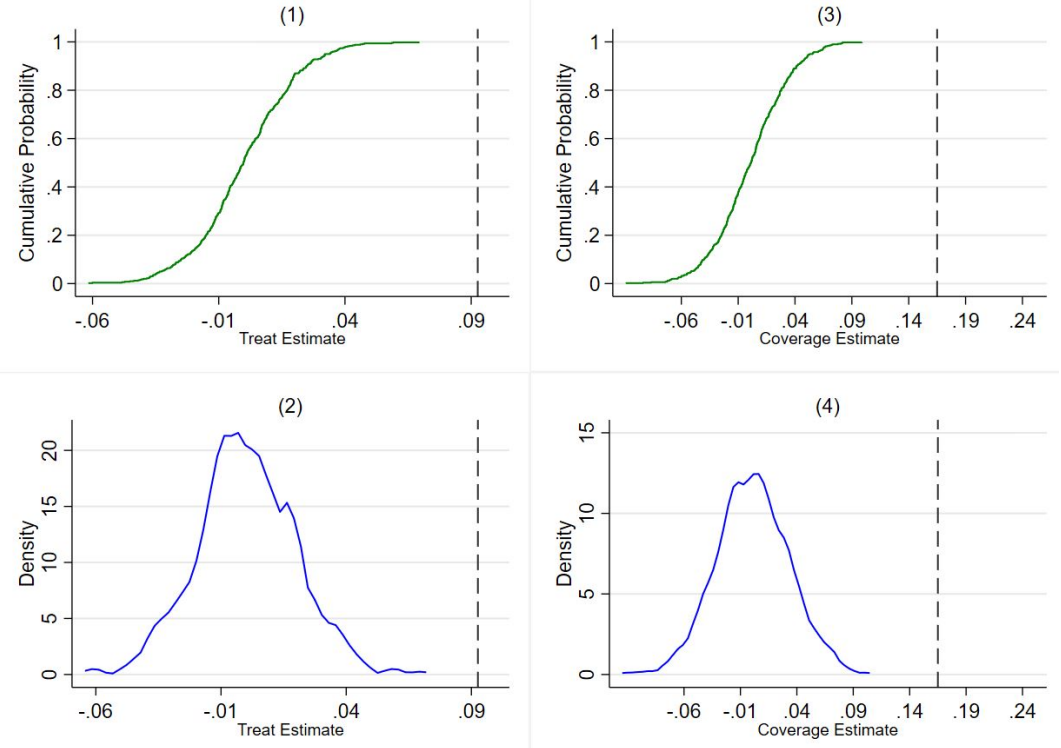
3.7 Discussion

3.7.1 Mechanisms

The estimates presented above may reflect a combination of the CAL program or enhanced teaching methodologies and increased focus on students who are struggling with reading and writing. In this section, we introduce multiple sets of supplementary evidence, each suggesting that the CAL software is the pivotal element driving the substantial improvements in test scores we find.

Improved Teaching Strategies

Figure 3.2: Placebo DytectiveU Implementation and Adoption in Distribution of Estimated Coefficients.



Notes. Top panel graphs report the cumulative distribution function of the estimated coefficients from 500 simulations using false Dytective U treatment assignment (Graph (1)) and coverage (Graph (3)). Bottom panel report the density distribution function of the estimated coefficients from 500 simulations using false Dytective U treatment assignment (Graph (2)) and coverage (Graph (4)).

We first provide evidence that the observed effects on student performance are not attributable to changes in teachers' understanding of learning difficulties as a result of adjustments to teaching strategies and increased attention towards struggling students. It may be plausible that teachers, by benefiting from the training program or by using the Dytective Test detection tool, could implement changes in instructional approaches. We re-estimated Equation (3.1), where the outcome variable is the standardized score on the 2018 Spanish test, and assigned the treatment to 39 primary schools that implemented the Dytective Test detection tool in the 2017-2018 academic year and benefited from the teacher and counselor training program. Note that in the 2017-2018 academic year, schools exclusively used the Dytective Test detection tool, and both teachers and school counselors underwent training sessions. Results from this exercise are shown in Table 3.8. For both

specifications, regardless of whether controls are included, our data show no differential change in students' performance between the treatment and control schools. This finding suggests that the intervention did not directly influence students' performance through changes in teaching strategies, but rather through the use of the CAL language software.

Table 3.8: Identification Check #4: Effects of Pilot Study on Spanish Language Standardized Test

	(1)	(2)
Treat	-0.0665	-0.0799
	(0.061)	(0.056)
Observations	25,108	12,418
R-squared	0.000	0.094
Controls	NO	YES
Number of Schools	299	298

Notes. Outcome variable: standardized score in 2018 Spanish test. The unit of observation is student i in school s, grade c and class group c. Treat is a dummy variable equals to 1 for students in the pilot schools, and 0 for schools that participate from 2018-2019. Columns (1) and (2) report results from estimating eq. 3.1 for students from schools that participated in 2017-2018 pilot study. Column (2) includes controls for students' gender and age, parents' educational level, number of books at home, school and group class size. Information on students' pre-primary enrollment, immigrant status and teacher's years of experience is not available for 2018 family and teacher questionnaires. Standard errors (in parentheses) are robust and clustered at school level. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

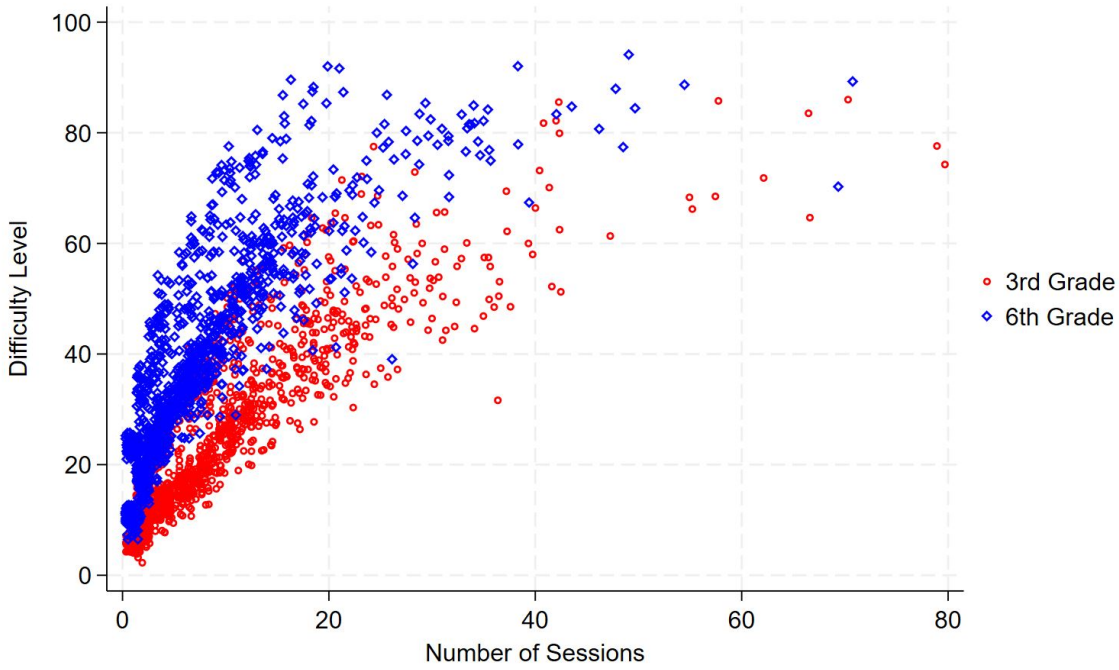
Personalized and Adaptive Instructional Material

We offer three sets of evidence demonstrating that the CAL software effectively mitigated the challenges faced by teachers in addressing the diverse abilities within group settings, by providing personalized and adaptive instructional materials tailored to each student's individual learning level.

We use CAL data to observe the dispersion of achieved difficulty levels across completed sessions by student grade. The main findings are presented in Figure 3.3, which shows the relationship between the acquired difficulty level—determined by the linguistic complexity addressed in the exercises—and the number of sessions

completed, for both third and sixth-grade students separately. Figure 3.3 highlights graphically two key attributes of DytectiveU software. The individual personalization feature of the instructional material can be inferred as students in the 3rd grade (red dots) typically engage with exercises at lower difficulty levels compared to 6th graders (blue diamonds), and; as students with the same number of completed sessions are seen engaging with exercises at varying difficulty levels, suggesting a tailored approach that targets different linguistic skills. The adaptive nature of the content is visually seen as a clear trend shows that as students progress through more sessions, the difficulty level of the exercises increases.

Figure 3.3: Dispersion In Achieved Level of Difficulty by Grade



Note: This figure shows the relationship between the number of completed sessions and the difficulty level of exercises assigned to students in the 3rd and 6th grades. The difficulty level, depicted on the y-axis as a normalized value ranging from 0 to 100, reflects the complexity of the linguistic elements addressed by the exercises (refer to Table B.3). The number of sessions, indicated on the x-axis, corresponds to the cumulative count of challenges each student has completed by the time of the standardized test.

Second, we also explored whether the improvements were consistent across grade, gender, and levels of maternal education by employing a linear interaction model. Our findings, as shown in Table B.14 in the appendix, reveal no significant variation in these respects. Thus, the use of DytectiveU seems to have pro-

vided equal benefits to students across a spectrum of socioeconomic backgrounds, suggesting that the software could effectively educate all students with equal proficiency.

Last, we explore whether the academic gains come from the intensive intensive margin (number of logged-in students) or the extensive margin (number of challenges per student). If the positive effects on test scores can be attributed to the software's potential for differentiated instructional content, we would expect to observe more significant gains through the extensive margin rather than the intensive margin, as the number of challenges completed by each student may be of lesser relevance. Students who complete fewer sessions could achieve similar gains as those with a higher number of sessions, given that the content fully adapts to each student's individual needs. Table B.15 of the Appendix presents the language score improvements attributable to the number of students who logged in and the number of sessions completed per student. Our analysis indicates that the variations in the extensive margin, reflected by the quantity of students logging in, account for an 18% of an standard deviation in language scores. Conversely, the intensive margin—measured by the number of challenges completed by each student—does not appear to contribute significantly to the score increments.

Consequently, the differentiated and adaptive nature of the content provided by DytectiveU seems to have been a pivotal factor in allowing all students to learn, thereby surpassing the constraints of grade-level pedagogy found in a typical classroom or an after-school tutoring group.

3.7.2 Cost-Effectiveness

Since we evaluate a CAL language program implemented in Spain, an inherent evaluation of cost-effectiveness is with CAL reading programs within the developed world. Escueta et al. (2020) synthesize and discuss the effectiveness of CAL programs within develop and developing countries and conclude that only a small fraction of CAL reading and spelling programs have been experimentally evaluated, while most of the research has focused on the evaluation of CAL math programs. Within the develop world, Escueta et al. (2020) indicate that, on average, CAL

reading programs had an impact of 0.15 standard deviations on a wide set of language related outcomes, including reading comprehension or spelling. Based on our ITT estimates, we observed that DytectiveU contributed to a gain of 0.09 standard deviations in Spanish language performance. It is important to note that the ITT estimates may represent a lower bound as the participation rate was around 50%. However, when accounting for actual program compliance, our dose-response estimates suggest a higher gain of around 0.19 standard deviations in Spanish language standardized test scores.

A second important comparison for policy-makers can be made with the productivity in terms time and cost of public primary schools, where the study participants were enrolled. In terms time, the DytectiveU CAL program was recommended to be used for over 1 hour and 20 minutes per week and was implemented 3 months before the standardized testing. In comparison, students in 3rd and 6th grade typically receive 6 and 5 hours of Spanish language instruction per week, respectively. When compared with the conventional learning gains typically observed in national and international assessments over the course of one academic year (Woessmann, 2016), the learning gains under typical teaching practices is equivalent to 0.12 standard deviations over the 3-month study period. In terms of cost, DytectiveU CAL program had a cost of 96,701.99 EUR. This includes the costs of the teacher and school counselor training, software development and IT support. Given that 4,928 students actively used DytectiveU in 2019-2019 academic year, the cost per student was 20.23 EUR. In 2018-2019 academic year, per-pupil academic year spending in public primary schools in the Region of Madrid was around 4,399 EUR (Ministerio de de Educación, 2022). While our findings suggest the DytectiveU CAL program can have positive effects on subjects beyond language, such as mathematics, it should be noted that public primary schools allocate their budgets to a variety of subjects, including music and physical education, where the reading demands may be less pronounced (Goldhaber et al., 2022).

This analysis suggest that CAL interventions as homework supplements could be more cost-effective than typical teaching practices, especially for struggling stu-

dents who show higher gains (as shown in Table 3.3). This implies that CAL programs can serve as a powerful tool for teachers to manage heterogeneous academic levels within a classroom. However, these insights should be take cautiously since the relative cost-effectiveness would depend on may contextual factors such as the student-teacher ratios or teaching cost.

3.8 Conclusions

We present pioneering evidence on the potential of CAL language programs like DytectiveU to improve students' literacy skills. Using administrative data on external blindly standardized test combined with survey data on families' and schools' characteristics, we show that the software was successful not only at increasing students' performance in Spanish language but also on subjects other than language, such as mathematics. These results are driven by students at the bottom of the language test-score distribution, who were the target of the program, and are usually left behind in traditional instruction. The success of DytectiveU can be attributed to its key software design features for delivering personalized and adaptive instructional content tailored to individual learning needs and preferences, underscoring the far-reaching potential of such technologies in education.

Our estimates are based on short-term academic outcomes. Further research may investigate whether the positive effects observed in the short-term are sustained over time, and whether they translate into improved long-term academic performance and other outcomes, such as labor or behavioral outcomes. Future research is also needed to dig more into the elements that underpin the effectiveness of CAL programs. While the importance of delivering personalized and adaptive content has been established, future studies should aim to disentangle these factors to individually assess their impacts. The role of feedback in the learning process and the influence of deployment modalities, such as at-home versus in-school usage, also require a more comprehensive investigation. These inquiries are essential for a more nuanced understanding of how distinct aspects of CAL software design contribute to educational gains.

Understanding the potentiality of CAL programs, such as DytectiveU, in increasing students' academic performance and closing gaps in education is crucial for designing cost-effective interventions in educational settings. Our research therefore has important policy implications. CAL programs, like DytectiveU, that provide personalized instruction without requiring teacher assistance, can offer a cost-effective solution for scaling up while avoiding the segregation of students, as they can be implemented as a homework supplement either in-class or at home. These programs can also be viewed as strategic tools to address the long-standing challenges teachers face in catering to diverse learning needs within a classroom.

Our paper documents a significant example of how a scalable and low-cost CAL language program can contribute to closing literacy gaps in educational settings and understanding the circumstances under which such technological tools are effective.

Chapter 4

Gender Gap In Response to Competitive Pressure

4.1 Introduction

Despite a global decrease in gender disparities in education and labor market participation, women continue to make educational choices leading to lower expected labor market earnings compared to men (Bertrand, 2020). Gender differences in performance and attitudes towards competitive pressure have been found to contribute significantly to the gender gap in educational choices and labor market outcomes, accounting for approximately 20% of the gender disparity in educational choices (Buser et al., 2014). While women tend to underperform compared to men in highly rewarded and more competitive tests, the opposite is true in less competitive settings or when the stakes are lower (Jurajda and Münich, 2011; Ors et al., 2013; Cai et al., 2019; Montolio and Taberner, 2021, among others). However, examining gender differences in response to increased competitive pressure remains a challenge, given the difficulty in isolating increased competition and stakes from other factors in real-life settings, such as gender differences in effort provision, and the self-selection of students into different types of examinations. I address this issue by using administrative records for the universe of students in Andalusia (Spain) from 2010 to 2019, and comparing gender differences in performance on both low-stakes exams during the last two years of high school and on high-stakes

exams from Spain's national university entrance examination.This approach allows for an examination of gender gaps in low- versus high-stakes settings, where in both settings performance is crucial for university access, thereby controlling ruling out potential gender differences in effort provision and minimizing sample selection bias.

There are several reasons as to why gender gaps may emerge from increasing competitive pressure. Gender differences in performance in competitive settings might not necessarily stem from skill differences, but could instead arise from behavioral disparities that become apparent only in specific competitive environments. Men's greater overconfidence and higher preference for performing in competitions seem to play an important role (Gneezy and Rustichini, 2004; Niederle and Vesterlund, 2007). Additionally, higher female aversion or tolerance to pressure, triggered by increased competitiveness, has been found to widen the gender gap in favor of men in educational settings (Jurajda and Münich, 2011; Cai et al., 2019; Morin, 2015, Iriberri and Rey-Biel, 2019). Recent literature has demonstrated the more detrimental effects of increased pressure on women's performance compared to men's in these settings. For instance, women are more negatively affected by time pressure (De Paola and Gioia, 2014), more likely to skip questions when penalized for incorrect answers (Pekkarinen, 2015; Saygin and Atwater, 2021; Iriberri and Rey-Biel, 2021), and tend to underperform when stakes are higher (Azmat et al., 2016; Montolio and Taberner, 2021).

The Spanish university admission system provides a useful framework for the analysis of gender performance gaps in competitive and high-stakes settings. In the region of Andalusia, as in the rest of Spain, approximately 95% of students taking the university entrance examination, the primary admission requirement for all universities, are graduates from upper secondary education. During this stage, students undergo evaluations with tests similar to those in the university entrance examination, in terms of duration, content, and choice. These high school tests account for 30% of a student's university access score. The remaining 70% comes from the university entrance examination, where a series of tests are conducted that mimic

those of upper secondary education in content, structure, and guidelines. The admission thresholds for each university program are calculated based on demand, and students are allocated to universities through a centralized algorithm using their university access scores. While the vast majority of students, around 90%, pass the university entrance examination, having a sufficiently high university access score is crucial for admission into the most competitive programs. Each academic year, the number of applicants exceeds the available places in public universities. For example, in the 2019-2020 academic year in Spain, though 389,652 students applied to university programs in public universities, only 245,513 places were offered (Ministerio de Universidades, 2020

The design of the Spanish university admission system, combined with student administrative records from both high school and the university entrance examination, enables two empirical strategies to identify gender differences in response to increased competitive pressure. First, I study gender disparities in performance by comparing how the same students perform in similar examinations under high-stakes and low-stakes settings, thereby identifying differences in gender responses to varying levels of pressure. Second, I use the proximity to the highest admission threshold as a means to introduce additional variation in competitive pressure within the university entrance examination setting. Based on the hypothesis that women may underperform under competitive pressure, it is expected that the gender gap in performance on the university entrance examination will be more pronounced for students whose scores are closer to passing the highest threshold, as indicated by their performance in high school.

In line with prior studies, I find strong evidence that women's performance relative to men's declines during high-stakes university entrance exams compared to low-stakes high school assessments. The analysis reveals that, on average, female performance suffers a decline of 0.31 standard deviations in the difference between low- versus high-stakes compared to male performance. This gender gap is largest for students applying to university programs in social and legal sciences, as well as for those applying to engineering and architecture, and smallest for those ap-

plying to arts and humanities. Examining the gender gap among students closer to the highest threshold, I find a difference of 0.7 standard deviations in the gender gap in favor of men between students within 0.015 points of the highest threshold and those within 0.04 to 0.06 points. This finding reinforces the role of competitive pressure, as women closer to the threshold experience a greater drop in their relative performance compared to women further away, thereby having less possibility of accessing the most demanded university programs. Based on conventional learning gains observed in international assessments, the reported gender gaps are equivalent to one academic year (Woessmann, 2016). In terms of implications for university placement, these estimates imply a reversal in the gender gap when allocating students based on their university entrance examination performance. Thus, the gender gap initially in favor of women was not only eliminated but actually turned in favor of men.

To further investigate the role of competitive pressure, I examine the reactions of women and men to performance shocks during the university entrance examination, defined as the deviation of a student's performance on a test in the university entrance examination relative to their performance in high school. I particularly exploit the structure of the university entrance examination, where multiple subjects are tested over three consecutive days. This setup allows to examine whether variation in pressure, driven by previous performance on an earlier test, arises gender differences. The university entrance examination comprises up to six subjects: four core compulsory subjects and two area-specific voluntary subjects, the latter carrying a higher weight in the university access score according to their relevance for the preferred university program. If competitive pressure contribute to gender disparities in performance, female students' responses to prior performance shocks will be more pronounced than those of male students, especially for subjects that are more significant for the university access score (voluntary field-specific subjects). The analysis shows that female relative performance on the next test is 0.24 standard deviations lower than that for males when I look at students closer to the highest threshold and performance on field-specific subjects, which weight more for

the college access score and take place on the same day. However, I do not find significant effects for the set of core subjects and for the groups of students' far away from the highest threshold. Performance shocks on field-specific subjects such as Biology and Chemistry appear to drive the gender differences in performance on the next test (Chemistry or Physics), sets of subjects taken by students that are applying to Health and Sciences university programs.

In examining gender differences in performance between high school tests and university entrance exams, I consider multiple mechanisms. The first potential mechanism is based on gender differences in effort: men might take high school less seriously as the subjects have lower weight compared to the university exam, and upper secondary education, being a two-year program, demands more sustained effort. Analyzing the impact of competitive pressure from the distance to the highest threshold, I find that both high-performing male and female students underperform when exceeding this threshold in high school, suggesting that gender disparities are not solely due to men's lower effort. Another potential channel is gender differences in unobserved ability. While university exams may require more cognitive skills, high school assessments might favor females due to non-cognitive factors like grit or behavior. However, findings show that female students within 0.04 to 0.08 points of the highest threshold have scores 0.16 standard deviations lower than males, and the gender score difference is 0.27 standard deviations for those within zero to 0.04 of the threshold. This indicates that unobserved ability differences do not solely account for gender disparities under increased pressure, particularly among those applying to competitive programs.

This paper contributes to the broad literature on gender differences in response to competitive pressure in educational settings. Several papers have looked at whether men and women respond differently to higher levels of pressure due to higher competition. Ors et al. (2013) compare the performance of students within a top Business School in France between the highly competitive admission examination and once admitted to the school showing that while males perform better than females at the admission test, women outperform men during the first year of

the course. Similarly, Morin (2015) exploits an exogenous increase in competition for university grades driven by an education reform in Ontario and finds that male average grades increase as well as the proportion of men graduating on time relative to women. Iriberri and Rey-Biel (2019) exploit a regional two-stage math contest and find that although female participants have higher maths grades at school, the gender gap reverses in the two stages of the contests. Focusing on the effects of pressure because of higher stakes at hand, Azmat et al. (2016) exploit the variation in the stakes of several tests in a high school in Spain and find that even though women outperform men in all tests, the gender gap disappears in tests taken during the last two years of the high school, which matters for the college access score. Montolio and Taberner (2021) examine gender gaps in performance at a university course when stakes of continuous assessment shift from low to high and find that the gender gap shrinks when the stakes decrease. While these studies explain gender differences in performance in response to higher competition or stakes, a narrower body of research has analysed such gender disparities as a result of both increased stakes and competitiveness. Jurajda and Münich (2011) find Czech women's exam results drop in high-stakes university entrance exams compared to less intense pilot tests. Pekkarinen (2015) show Finnish women's lower performance and admission rates in university economics and business exams versus their high school results. Cai et al. (2019) observe a steeper decline for Chinese women than men when comparing the high-stakes Gaokao with mock exams.

The uniqueness of the setting and the rich population-level student administrative records differentiates this paper from prior research. First, the low-stakes high-school examination accounts for the 60% of the university access score, this ensure that I am picking up gender differences in response to competitive pressure rather than differences in the provision of effort. As opposite to previous studies that examine differences in performance between actual and mock examinations, I am able to look at evaluations that actually matter for university admission. The analysis of effort provision indicates that the widening gender gap can be attributed to a decline in female performance in a high-stakes setting as opposed to an im-

provement in male performance. Second, since university entrance examination is the main requirement for university admission in Spain accounting with more than 90% pass rate, there is limited sample selection of students into the university entrance examination based on their performance during the last two years of high school. Third, I have population data for a ten-year period with individualised information on the score of each subject tested at the university entrance examination allowing to provide insights on the role competitive pressure on gender gaps across different fields, which have different demands and different labour returns, and to avoid sample selection bias that may affect previous studies using lab experimental methods in small samples or quasi-natural experiments in highly selected samples.

My findings will help shape policies around the design of standardised testing to address gender gaps in performance. Standardized tests, such as the Cito in the Netherlands, the SAT in the U.S., and the Baccalauréat in France, are widely used and play a crucial role in shaping students' educational paths. Yet, measuring student ability through standardized testing may be problematic. Test scores depend not only on cognitive ability, but also in other factors such as non-cognitive skills (Cunha and Heckman, 2007). Statistics from the 2015 to 2020 Integrated Information System on Universities in Spain (SIIU) indicate that while women account for about 52% of university students in Sciences in Spain, they are underrepresented in the most competitive bachelor's degrees with the greatest employability and future prospects.[1] To the extent that the ability to cope with competition is an important factor of gender differential in performance (Niederle and Vesterlund, 2007), policy makers should modify the stakes of the admission examinations or broadening the spectrum of entry criteria for university programs in order to increase the representation of well-qualified women in more competitive and higher-paying fields.

[1]The female share in the top programs in Sciences, such as the dual university program in Mathematics and Physics, is around 20%, while females are overrepresented in less competitive programs such as Chemistry or Geology, where female shares are 55% and 85.7% in Andalusia (Spain), respectively (Ministerio de Universidades, 2020). These differences at the college level may translate to gender differences in the labor market. Programs such as Mathematics or Physics account for unemployment rates between 2 and 4 percentage points lower than the average rate for university graduates, whereas Chemistry or Geology unemployment rates are around 4 percentage points higher and account for salaries around 15% lower than the average rate for college graduates (Instituto de Estadística , INE)

The remainder of the paper is structured as follows. Section 4.2 provides background on the Spanish upper secondary education system and university entrance examination. Section 4.3 presents the data used and descriptive statistics. Section 4.4 discusses the empirical strategy, section 4.5 shares the main results, and Section 4.6 explores possible mechanisms behind the increased gender differences in scores in low- versus high-stakes settings. Section 4.7 concludes.

4.2 Institutional Background

In Andalusia, as in the rest of Spain, most students that want to continue to university upon completion of compulsory secondary education (ages 12 - 15) choose the academic track and enroll in upper secondary education (ages 16 -17, ISCED 3), the so-called *Bachillerato*.[2] Upper secondary education is divided in three streams: Arts, Humanities and Social Sciences, and Sciences. During this stage, students are required to undertake four general core subjects, two field-specific core subjects (detailed in Table C.1 in the Appendix), and between a minimum of two and a maximum of three field-specific subjects (detailed in C.2 in the Appendix). Most high schools devote the entire two-year period to university entrance examination preparation. Students typically undergo tests mirroring those in the actual examination, featuring similar time constraints and choice options. As a result, the average scores from the last two years reflect performances in a low-stakes setting.

For university admission, students must pass the national university entrance examination, known as *Selectividad*, which is taken upon completion from upper secondary education (*Bachillerato*). Passing the university entrance examination is a compulsory prerequisite for all students seeking admission to any university, irrespective of the university program's field of knowledge or the type of institution, whether public or private. Students sitting for the university entrance examination must undertake at least four written tests: (1) Spanish, (2) a Foreign Language (English, French, Italian, German, or Portuguese), (3) History or Philosophy, and (4) one field-specific subject. In addition, they can choose to take up to two addi-

[2]International Standard Classification of Education (ISCED) adopted by the UNESCO General Conference in its 36th session in November 2011.

tional field-specific tests to increase their college access score, provided the subjects' fields correspond to those of the university program to which they later apply (see correspondence in Table C.3. Each voluntary field-specific test carries a weight of 0.1 or 0.2, contingent upon the relevance of the subject's syllabus to the preferred university program.

If a student uniquely sits for the compulsory tests at university entrance examination, the university access score will be graded from 5 to 10 points and computed as the weighted average of the last two years of high school (*Bachillerato*) (accounting for 60%) and the university entrance examination average score (accounting for the remaining 40%). Consequently, the university access score for student i will be calculated as follows:

$$\begin{aligned}\text{University Access Score}_i = & 0.6 \times \text{High-School Average Score}_i \\ & + 0.4 \times \text{University Entrance Examination (compulsory tests)}_i\end{aligned} \tag{4.1}$$

If student i also takes one or two voluntary field-specific tests, the college access score will be graded from 5 to 12 or 5 to 14 points, respectively:

$$\begin{aligned}\text{University Access Score}_i = & 0.6 \times \text{High-School Average Score}_i \\ & + 0.4 \times \text{University Entrance Examination (compulsory tests)}_i \\ & + A \times \text{Field-Specific Subject I}_i \\ & + B \times \text{Field-Specific Subject II}_i\end{aligned} \tag{4.2}$$

University program admission thresholds are calculated based on the demand for each program. They are published publicly every academic year and are defined by the university access score of the student who fills the last available slot. University programs are categorized into five main fields: Arts, Social and Legal

Sciences, Engineering and Architecture, Health Sciences, and Sciences. To be eligible for university enrollment, candidates must achieve a minimum average score of 4 out of 10 in the compulsory tests and, an average of at least 5 out of 10 combining weighted high school grades and compulsory test scores at university entrance examination (as calculated in equations 4.1 and 4.2).

While the vast majority of students pass the university entrance examination, approximately 90%, achieving a high enough university access score is crucial to compete for the most demanded programs. Each academic year, the number of applicants exceeds the available slots in public universities. For example, during the 2019-2020 academic year in Spain, 389,652 students applied to university programs in public universities, but only 245,513 spaces were available. This gap between supply and demand is particularly stark in fields like Health Sciences and Sciences, with more than 3 and 1.5 applicants per available spot, respectively (Ministerio de Universidades, 2020)). University programs in Health Sciences and Sciences also have the lowest admission ratios for top programs, both approximately 65%, while in other fields, admission rates for the first-choice options range from 75% to 85%.

4.3 Data and Sample

The data are based on population-level administrative records, including test scores and demographic information for the universe of students who sat for the university entrance examination in the region of Andalusia, Spain, from 2010 to 2019. The dataset includes information on upper secondary education - *Bachillerato* - average score as well as scores for each test on the different evaluated subjects at university entrance examination. The administrative records also contain information on the student's gender, birth date, and whether the university entrance examination tests belong to the ordinary call (June) or the extraordinary call (September). The sample comprises 287,555 individuals, of whom 240,961 sat for the ordinary call in June. Approximately 85% of the students who sit for the ordinary call take the two voluntary, field-specific, tests to be graded on a scale of 5 to 14 points, instead of 5 to

10 points, which would prevent them from being admitted to the most demanded programs.

Table C.4 in the Appendix reports score summary statistics for the full sample and by field of study, which are: Arts and Humanities, social and Legal Sciences, Health Sciences, Engineering and Architecture, and Sciences. I classify students by their field of study based on the set of subjects they choose to take in the university entrance examination. Specific sets of field-specific subjects carry a higher weight in the university access score if they align with the field of the subsequently chosen university program. Table C.3 in the Appendix shows the link between subject choices and the fields of study in university programs. Table C.4 in the Appendix shows a higher average performance in high school than in university entrance examination across all fields. Females constitute 55% of the university students. While female representation is 74% in Arts and Humanities, they are underrepresented in Engineering and Architecture, where they make up only 24% of students.

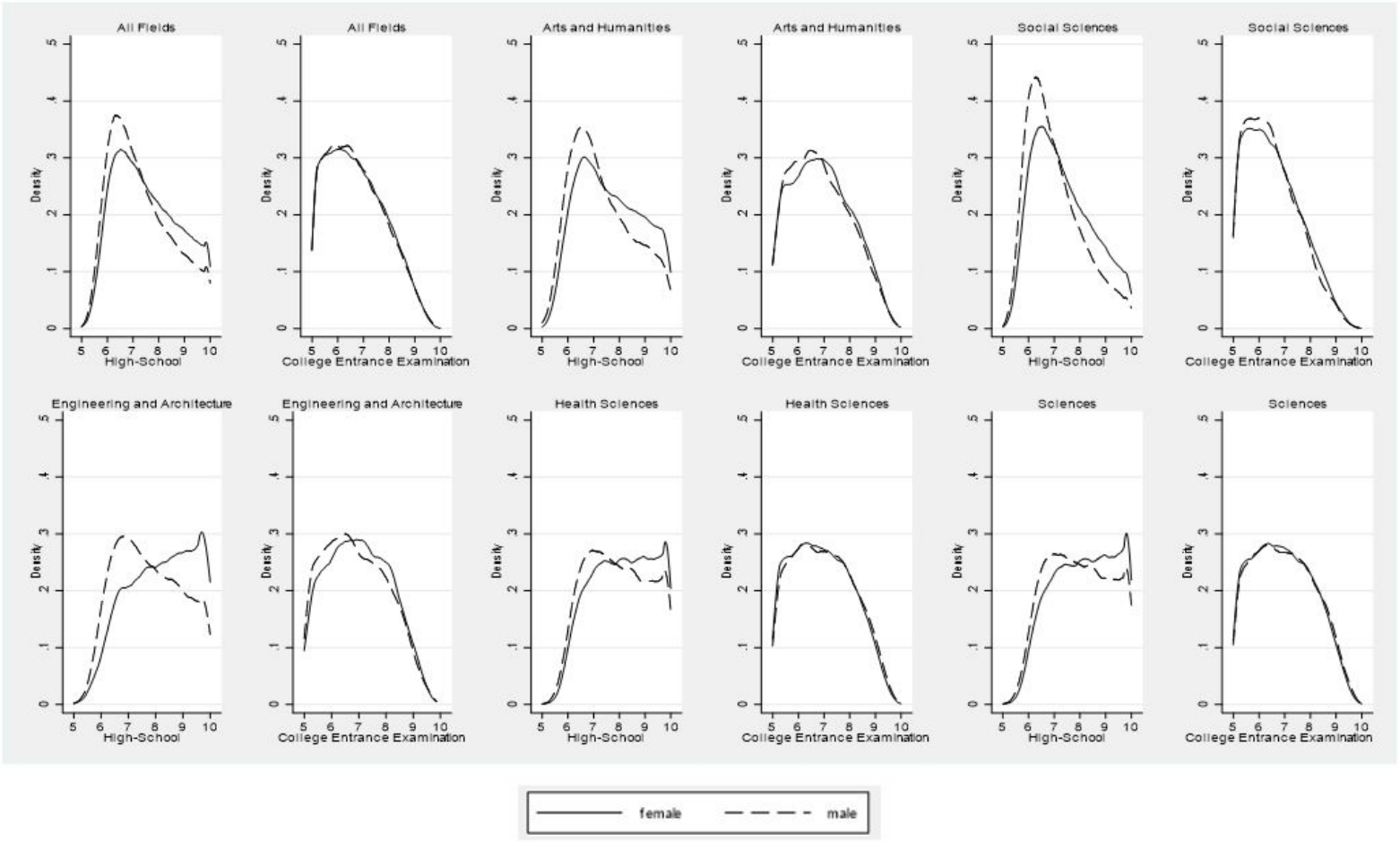
To identify students facing higher competitive pressure, I obtain data on the threshold for each university program in the region of Andalusia from 2010-2019. This data, sourced from the regional online portal, is merged with student administrative records by field and academic year. [3] Table C.5 in the Appendix reports higher admission thresholds for programs in Sciences and Health Sciences, and lower admission thresholds in the fields of Social Sciences and Law, or Engineering and Architecture. Data reports a 7% admission rate in the most competitive programs for the full sample. These rates drop to 5.77% in Sciences and rise up to 14% in Engineering and Architecture.

Figure 4.1 graphically shows the gender differences in the standardized score difference between high school and the university entrance examination for the full sample and by fields. The descriptive statistics indicate that on average, female

[3] In the regional university online portal, known as the *Distrito Único Andaluz*, one can find the historical data of admission thresholds for every university program within the public universities in Andalusia, see here: `https://www.juntadeandalucia.es/economiaconocimientoempresasyuniversidad/sguit/?q=grados&d=g_not_cor_anteriores_top.php`].

students experience a 0.20 standard deviation decline between the high school and university entrance examination average grade, whereas male students report a 0.15 standard deviation increase. These differences are even more pronounced among more competitive fields such as Sciences, where females experience a loss of 0.22 standard deviations, and males show a gain of 0.18 standard deviations.

Figure 4.1: Evidence of the gender gap in performance on high school and university entrance examination average raw scores for the full sample and by field from 2010 -2019.



Notes: Each panel plots the distribution of the total exam score for male and female students separately for the high-school and university entrance examination average score.

4.4 Empirical Strategy

The primary goal of this paper is to identify the causal effect of high- versus low-stakes settings on students scores by gender. Within a framework where high school is consider as a low-stakes setting and the university entrance examination as a high-stakes setting, I estimate the following equation using ordinary least squares (OLS):

$$S_{i,g}^{U} - S_{i,g}^{H} = \alpha + \beta Female_i + \delta\gamma_i + \theta_i + \varepsilon_i \tag{4.3}$$

Where U denotes the university entrance examination, H denotes high school, i denotes the individual, and g denotes gender. Therefore, $S_{i,g}^{U} - S_{i,g}^{H}$ is the standardized difference between the average score in the university entrance examination and the average score of the last two years of high school, which were evaluated following similar guidelines. $Female_i$ is a dummy variable to capture the gender, so β can be interpreted as the effect of being female on the difference in test scores between the university entrance examination and high school in standard deviations. Year fixed effects θ_i are included to monitor the differences of complexity of university entrance examination each year. γ_i denotes individuals characteristics (age and field), and ε_i is the error term.

4.5 Main Results

4.5.1 Gender Gap in Performance

Table 4.1 presents the estimated female coefficient, β, from the estimation of equation 4.3 for the full sample of students (Column 1) and desegregated for students applying for university programs in the fields of: Arts and Humanities (Column 2), Health sciences (Column 3), Social and Legal sciences (Column 4), Engineering and Architecture (Column 5), and Sciences (Column 5). As shown in Column 1, the baseline difference in score between university entrance examination and high school is, on average, 0.31 standard deviation lower for women compared to men across the full sample. This gender gap is more pronounced in Social and Legal Sciences and Engineering and Architecture, where the male advantage approximates

0.38 standard deviations. The smallest difference, accounting for 0.26 standard deviations, is found among students applying to university programs in the Arts and Humanities field. To provide a sense of the magnitude of our estimates, I compare them with the conventional learning gains observed in most national and international assessments over one academic year. These gains are usually between a quarter and one third of a standard deviation (Woessmann, 2016). Hence, the reported gap in favor of men ranging from 37% to 26% of a standard deviation is equivalent to 1.04 to 1.12 academic years of learning gains.

Table 4.1: Gender Gap In Performance (High Schools vs University Entrance Examination)

	(1) Full Sample	(2) Arts and Humanities	(3) Health Sciences	(4) Social and Legal Sciences	(5) Engineering and Architecture	(6) Sciences
Female	-0.3149***	-0.2600***	-0.3254***	-0.3723***	-0.3681***	-0.3329***
	(0.004)	(0.022)	(0.008)	(0.010)	(0.016)	(0.007)
Controls: Age, YearFE	X	X	X	X	X	X
Observations	240,905	10,341	47,891	46,161	17,443	60,568
R-squared	0.051	0.052	0.059	0.084	0.046	0.057

Notes.Each column is a separate regression with the standardized difference between the university entrance examination average score and the high school average score. Both the university entrance examination score and high school test scores were standardized to have a mean of 0 and standard deviation of 1 by student field of knowledge (Arts and Humanities, Health Sciences, Social and Legal Sciences, Engineering and Architecture and Sciences). The specification reported in Column 1 (full sample) includes field of study fixed effects. Robust standard errors are reported in brackets. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

4.5.2 Gender Gap for Students Closer to the Highest Threshold

In what follows, I delve deeper into how competitive pressure affects the gender gap in high versus low-stakes situation. In the following analysis, I focus only on students who are close to be eligible to the most competitive university program, and thereby scoring near the highest admission thresholds. These students are presumed to face greater competitive pressure based on the notion that a minor variation in their performance relative to peers who are less likely to gain admission to these top university programs could significantly affect their chances. To conduct these exercise, I obtain data from the regional university online portal, which includes de-

tailed information about the admission thresholds for all university programs across academic years.

Table 4.2 reports results from this exercise. The female coefficient estimate, β, from the estimation of equation 4.3 for the standardized difference in average scores between the university entrance examination and high school is shown for four subgroups of students: those with high-school scores within 0.015 points, 0.025 points, 0.03 to 0.04 points, and 0.04 to 0.06 points of the highest threshold by field and academic year. The evidence is consistent with the idea that female students underperform on the university entrance examination relative to high school when they are closer to be eligible to the most demanded programs; the gender gap in relative performance is largest among students within 0.015 points of the highest threshold and declines for students further away from the threshold. In particular, While for female students who score within 0.015 points of the highest threshold, the difference in scores is 0.21 standard deviations lower than that for males significant at the 1% level, for students within 0.04 to 0.06, the difference in scores is 0.14 standard deviations lower among females than that for males, significant at the 10% level. [4]

Table 4.2: Gender Gaps in Admission to Most Competitive University Programs

	(1) (+0.015,-0.015)	(2) (+0.025,-0.025)	(3) (0.03,+0.04) and (-0.04, -0.03)	(4) (+0.04,-0.06) and (-0.06, -0.04)
Female	-0.2104***	-0.1973***	-0.1649**	-0.1446*
	(0.074)	(0.062)	(0.074)	(0.076)
Controls: Age, Field, YearFE	X	X	X	X
Observations	578	841	532	622
R-squared	0.093	0.109	0.061	0.085

Note. Each column is a separate regression with the standardized difference between the high-school and university entrance examination average score as the dependent variable for students at a different distance from the highest threshold based on their high school performance. Only students that have chosen the most efficient combination of subjects, and therefore, shown intention to apply for the most demanded programs, have been selected. Robust standard errors are reported in parenthesis. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

[4] I conduct a placebo test using "fake" thresholds from two years prior to define the intervals. This two-year period was chosen to allow for some variability, as thresholds fluctuate only slightly between academic years. Since distance to these lagged thresholds should not imply increased pressure, the pattern observed in Table 4.2 should not persist. As demonstrated in Table C.6 of the Appendix, this is indeed the case, further supporting the validity of the estimates reported in Table 4.2.

4.5.3 Back of the Envelope

To provide a sense of the implications for placement into university programs that above estimates imply, I conduct a simple exercise assuming that the relative change in performance is due entirely to gender differences in the response to the low versus high stakes. To do so, I observe the proportion of students by gender who would have surpassed this highest threshold and been admitted to the most demanded program in their field based on their average performance in high school and their average scores on the university entrance examination.

Table 4.3 presents the overall fraction and, male and female fractions of students who are eligible for the most demanded university programs by field. This is determined by the proportion of students exceeding the highest thresholds based on their high school average scores and university entrance examination average scores. While based on their performance on high school women are 3 and almost 4 percentage points more likely to get admitted in the most demanded programs than their male counterparts in Health Sciences and Sciences respectively (Column 3), the likelihood reverses in both in these fields if allocating students with university entrance examination performance (Column 7). Given that on average, around 2% of students are eligible for the most demanded program in Health Sciences and Sciences, this translates to a relative decrease of around 113.77% in Health Sciences (0.0305 - (-0.0042) / 0.0305), and 110.08% in Sciences (0.0387 - (-0.0039) / 0.0387). On the other hand, whereas the likelihood of being accepted into a top program in Arts and Humanities and Engineering being a women is higher based on the performance at high school, this gap in favor of women disappears when observing the fraction of eligible students using their performance at university examination.

4.6 Mechanisms

4.6.1 Gender Gap in Response to Performance Shocks

Next, I examine whether male and female students react differently to performance shocks on a previous exam during the university entrance examination. As multiple subjects are tested consecutively across a three-day period during university

Table 4.3: Implication for University Placement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	High School				University Entrance Examination				
	Overall	Women	Men	Women-Men	Overall	Women	Men	Women-Men	Column (3) - Column (7)
Arts and Humanities	0.253	0.270	0.204	0.0571***	0.056	0.057	0.053	0.0022	0.054***
Health Sciences	0.223	0.232	0.207	0.0305***	0.018	0.016	0.022	-0.0042***	0.0347***
Social and Legal Sciences	0.216	0.249	0.166	0.0752***	0.045	0.046	0.043	0.0021	0.0731***
Engineering and Architecture	0.269	0.361	0.238	0.1195***	0.047	0.054	0.044	0.0082**	0.1111***
Sciences	0.223	0.239	0.202	0.0387***	0.018	0.016	0.021	-0.0039***	0.0426***

Notes. Columns (1) to (3) and (5) to (7) show the share of all, male and female students score above the thresholds for admission into the most demanded university program based on their performance at high school (Columns (1) to (3)) and university entrance examination (Columns (4) to (6)). Columns (4), (8) and (9) are obtained from a linear probability models with a dummy variable taking value 1 for individuals obtaining above the highest threshold. Specifications shown Columns (4), (8) and (9) include controls by age and test year fixed effect. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

entrance examination, I am able to exploit variation in pressure, which increases because of differences in performance on the previous exam. I also exploit the fact that the university entrance examination calendar, and therefore the order of the tests, changes every school year. This setup allows me to avoid possible bias driven by schedule effects and to distinguish whether the increased pressure effects due to an earlier performance shock persist even for those tests that take place the following day. Sievertsen et al. (2016) suggest that the time of day affects students' test performance because over the course of the day students become increasingly fatigued. On the other hand, previous research has also found that early school schedules are detrimental to academic performance due to young adult's different sleep and wake patterns (Wahistrom, 2002). For those sets of tests that take place on the same day I do not expect students to be able to adjust their preparation for the later exam within such a narrow time frame. Therefore, in this setting, there is limited scope for gender differences in effort provision to drive the observed patterns. However, for those sets of tests that take place on different days there is a wider scope for effort provision.

I hypothesize that performance on a previous test increases competitive pressure levels faced by students in next tests since they will need to compensate for underperforming in the previous test. I also expect a higher response to competitive pressure for those set of tests that take place the same day and for the voluntary set

of field-specific subjects, which are those that weigh more for the university access score since they are more relevant for the preferred university program. In particular, I examine how a student's performance on the later test is affected by shocks to their performance on the previous exam test. For this exercise, I select only students who have obtained the highest possible average score in high school, so I know that these students have received the highest possible score in every single subject (10 points out of 10). As subject evaluations during high school and university entrance examination are graded on the same scale (from 0 to 10 points), to proxy for performance shock, I use the deviation between an individual's score in a given test on the university entrance examination and the maximum score they could have obtained, which is the score that these highest-performing students have gotten during high school. I also select only students who have chosen the most efficient combination of subjects on university entrance examination for applying to the most competitive and demanded programs, which are with the highest threshold.

My empirical strategy relates a student's relative performance on the previous test to their relative performance on the later test. The regression specification is as follows:

$$S_{i,g,t}^{UNextTest} = \alpha + \beta_1 \text{Female}_i \times (S_{i,g,t}^{UPreviousTest}) + \beta_2 (S_{i,g,t}^{UPreviousTest}) + \beta_3 \text{Female}_i + \beta_4 \gamma_i + \theta_i + \delta_i + \varepsilon_i \tag{4.4}$$

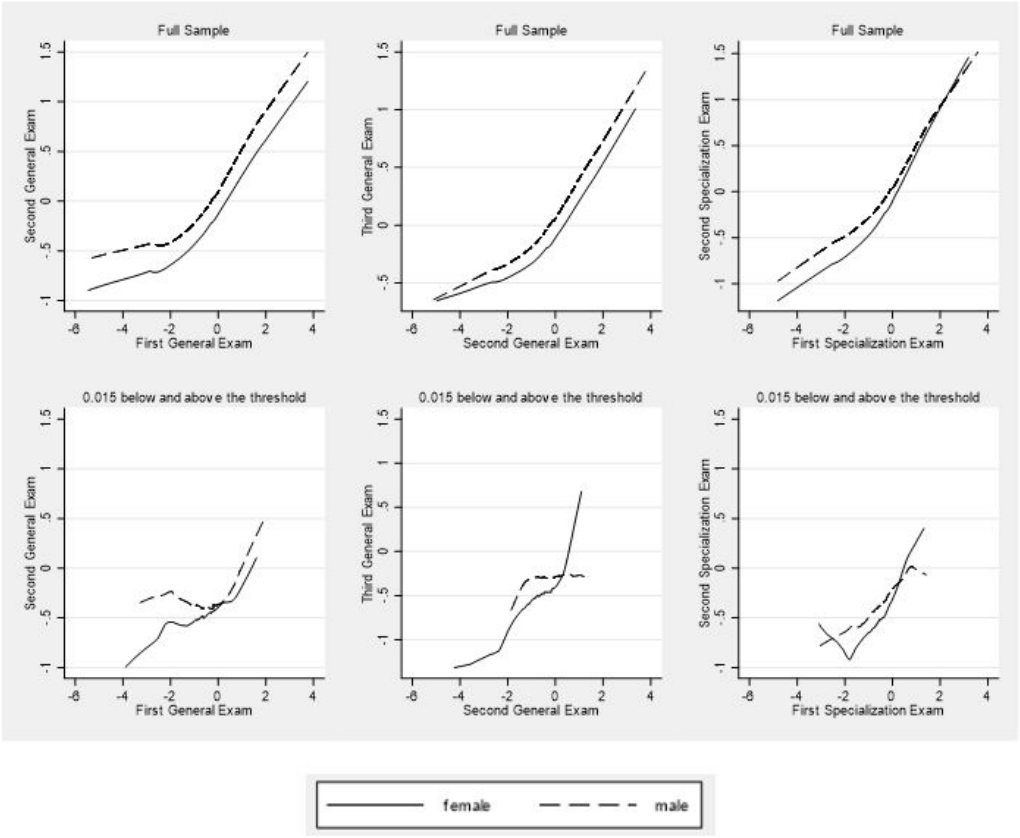
where the outcome $S_{i,g,t}^{UNextTest}$ is the standardized difference in student i's score on subject t on the university entrance examination that takes place second and its maximum possible score, and $S_{i,g,t}^{UPreviousTest}$ is the standardized difference in student i's score on subject t comprised in the university entrance examination that take place first and its maximum possible score. Controls θ_i are the same than those included in equation (4.3). I add test year (γ_i) and field (δ_i) fixed effects. The coefficient β_1 measures how deviation from the maximum score achieved in high school on the first test deferentially affects the relative performance of female students on the next test relative to male students.

Figure 4.2 presents the locally smoothed (LOWESS) graph of the uncondi-

tional relationship between the relative performance on the first test ($S_{i,g,t}^{UPreviousTest}$) and the relative performance on the next test ($S_{i,g,t}^{UNextTest}$) that take place the same day separately for men (solid line) and women (dashed line) and for the set of core subjects (compulsory) and field-specific subjects (voluntary). From Figure 4.2, Ican extract three main insights: (i) there is a positive relationship between relative performance on the previous test and the next test; (ii) this positive relationship appears to be more pronounce for female students relative to male students for the set of tests that weight more for the university access score (field-specific and voluntary tests), suggesting that female performance is more affected by performance shocks in previous tests than male performance for tests that matter more and; (iii) the relationship appears to be a lot stronger for women when students are closer the highest threshold based on the high-school performance, suggesting that female performance is more affected than male performance by performance shocks when they compete to access the most demanded programs.

Table 4.4 reports the impact of gender differences in negative performance shocks on the first core and compulsory test (Columns 1 to 3), second core and compulsory test (Columns 4 to 6), and the first voluntary and field-specific test (Columns 6 to 8) for the full sample of students achieving 10 points out of 10 in high school and by distance to the highest threshold. The coefficient of interest, β_1, indicates that in response to a 1 standard deviation improvement (decline) in relative scores in the previous test with respect to high school performance, the female relative performance on the next test is 0.24 standard deviations higher (lower) than that of males for those students closer to the highest threshold when Il ook at voluntary subjects, which weigh more for the university access score (Column 8). This estimate is significant at the 1% level. However, I do not find significant effects for the set of compulsory subjects, which weigh less for the university access score, and for the groups of students further away from the highest threshold. Interestingly, β_2 is positive and statistically significant for both compulsory and voluntary sets of tests and for the subgroups of students further away from the highest threshold, indicating that the relative performance on the previous test tends to be positively correlated

Figure 4.2: Relationship Between the Previous and Next Test on the Same Day



Notes. The panels plot the relationship between the standardized difference between the score of the test that take place first during the university entrance examination and its possible maximum against the next test separately for male (solid line) and female students (dashed line). Only students with the maximum score in high school and choosing the most efficient set of subjects for applying to the most demanded programs are sampled.

with the relative performance on the next test. Consistent with the graphical evidence presented in Figure 4.1, these findings suggest that negative performance on the previous test affects performance on the next test significantly more for female students relative to male students when it matters more for university admission.

In Table 4.5, I show which voluntary tests drive the gender differences in response to the performance shocks. Table 4.5 reveals that among students with the highest score in high school, in response to a 1 standard deviation improvement (or decline) in relative scores in Biology and Chemistry tests during the university entrance examination, the relative performance of female students in Chemistry (following Biology) and Physics (following Chemistry) is 0.22 and 0.34 standard

deviations higher (or lower) respectively than that of male students (as shown in Columns 6 and 7). These subject combinations are typically chosen by students applying to university programs in Health and Sciences fields, which are among the most demanded programs.university programs.

Table 4.4: Gender Differences in the Next Test in Response to Relative Performance Shock in the Previous Test on the Same Day for the Full Sample and by Distance to the Highest Threshold.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Standardized difference in university entrance test– possible maximum score on the next test								
	Full Sample	(-1,+1)	(-1.5,+1.5)	Full Sample	(-1,+1)	(-1.5,+1.5)	Full Sample	(-1,+1)	(-1.5,+1.5)
Relative performance on the first compulsory test x female	0.0102	0.045	0.0052						
	(0.026)	(0.05)	(0.043)						
Relative performance on the first compulsory test	0.0937***	0.053	0.1177***						
	(0.021)	(0.04)	(0.036)						
Relative performance on the second compulsory test x female				0.025	0.046	0.0004			
				(0.034)	(0.07)	(0.066)			
Relative performance on the second compulsory test				0.0861***	0.031	0.1707***			
				(0.027)	(0.06)	(0.055)			
Relative performance on the first voluntary test x female							0.032	0.2419***	-0.0387
							(0.033)	(0.092)	(0.061)
Relative performance on the first voluntary test							0.1245***	0.0945	0.2378***
							(0.025)	(0.065)	(0.05)
	(0.071)	(1.47)	(0.745)	(0.07)	(1.35)	(0.744)	(0.092)	(1.808)	(0.909)
Female Dummy	x	x	x	x	x	x	x	x	x
Controls: Age, FieldFE, YearFE	x	x	x	x	x	x	x	x	x
Observations	2608	656	978	2548	651	953	2678	697	1015
R-squared	0.215	0.107	0.102	0.116	0.013	0.068	0.259	0.122	0.191

Note. The relative performance is measured using the standardized difference in university entrance examination and high-school scores sampling only students who score 10 points out of 10 in every subject evaluated during high-school as well as tests that take place on the same day. Columns (1), (4) and (7) report the gender difference for the full sample of students. Columns (2), (5), and (8) report gender differences for students who score between -1 to 1 to the highest threshold based on their high-school performance. Columns (3), (6) and (9) report gender differences for students who score between -1.5 to 1.5 to the highest threshold based on their high-school performance. Robust standard errors are report in parenthesis. ***significant at 1%, **5%, *10%.

Table 4.5: Gender Differences in the Next Test in Response to Relative Performance Shock in the Previous Test on the Same Day for Field-Specific Voluntary Tests.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Standardized difference in university entrance test– possible maximum score on the next test						
	Geography	Economy	Mathematics of Sciences	Physics	Technical Drawing	Biology	Chemistry
Relative performance on the first voluntary test x female	0.0452	-0.0157	0.0372	-0.2334	0.0221	0.2188***	0.2905**
	(0.066)	(0.522)	(0.054)	(0.179)	(0.082)	(0.083)	(0.123)
Relative performance on the first voluntary test	-0.0592	-0.1483	0.1703***	0.2920***	0.0794*	0.1065*	0.2101***
	(0.054)	(0.370)	(0.044)	(0.092)	(0.046)	(0.062)	(0.075)
Female Dummy	x	x	x	x	x	x	x
Controls: Age, FieldFE YearFE	x	x	x	x	x	x	x
Observations	226	48	1,212	131	356	672	270
R-squared	0.043	0.102	0.088	0.242	0.125	0.092	0.355

Note. Each column is a separate regression with the standardized difference of each subject score on university entrance examination and high-school as dependent variable. I sample only students who score 10 points out of 10 in every subject evaluated during high school and who have selected the most relevant combination of subject for university entrance examination to apply for the most competitive programs. Robust standard errors are report in parenthesis. ***significant at 1%, **5%, *10%.

4.6.2 Gender Gaps in Effort Provision

An alternative possible explanation for the observed patterns is that men might not take high school (low-stakes) as seriously as women, considering each subject has a relatively lower weight in the final high-school average score and hence, they need to make a more sustained effort.

Another strategy to explore whether gender gaps in effort provision underlie the identified patterns, I employ a methodology that leverages the varying levels of stakes associated with proximity to the highest threshold within gender. This exercise will allow to disentangle whether the trends observed can be attributed to a decline in female performance or an enhancement in male performance, particularly in the high-stakes context of the university entrance examination. If the exercise reveals a significant decline in female performance, it would mitigate the possibility that the preliminary findings are primarily influenced by a lower level of effort from male students during their high school years.

Table 4.6 shows how the performance gap between female and male students varies as a function of the distance from the highest threshold. The dependent variable is the difference in standardized (within-gender) scores between university entrance examination and the high school average score. Column (1) reveals that female students exhibit a decline of 0.257 standard deviations (with a standard error of 0.096) in their university entrance examination performance compared to their high school average when they are within a range of zero to 0.015 points above the highest threshold. This is in contrast to female students who fall within the range of 0.04 to 0.06 points both below and above this threshold. Column (2) presents a similar, albeit non-significant, negative trend for male students who are zero to 0.015 points above the threshold, relative to their counterparts who are 0.04 to 0.06 points both below and above the threshold. The estimates for students positioned below or significantly beyond the threshold do not show statistical significance for either gender.

These results imply that the performance of both male and female students declines when they score above the highest threshold based on their high school

achievements. However, this decrease is more pronounced among female students. Importantly, the observed gender differences in performance between the university entrance examination and high school are not attributable to lower effort levels by high-performing male students during their high school years.

Table 4.6: Female Underperformance versus Male Underperformance.

	(1)	(2)	(3)
	Standardized Scores within Own Gender Distribution		
	Female	Male	Full Sample
(0, +0.015)	-0.2573***	-0.1660	-0.1436
	(0.096)	(0.120)	(0.117)
(-0.015, 0)	0.0334	0.0405	0.0657
	(0.089)	(0.102)	(0.095)
(-0.04, -0.02) and (0.02, 0.04)	-0.0530	0.0573	0.0631
	(0.065)	(0.075)	(0.071)
(-0.06, -0.04) and (0.04, 0.06)		Reference Group	
Controls: Age, FieldFE, YearFE	x	x	x
Observations	1,115	767	1,882
R-squared	0.079	0.074	0.072

Notes: Each column is a separate regression of the standardized (within-gender) difference between the Univeristy Entrance Examination and high-school average score for women only (columns 1) and men only (columns 2) on indicators of the distance from the highest threshold. Column 3 reports the difference in coefficients for the female sample and male samples. (0, 0.015) refers to a dummy variable indicating that a student scores zero to 0.015 points below the highest threshold, (0, +0.015) refers to a dummy variable indicating a student scored zero to 0.015 points above the highest threshold (-0.04, -0.02) and (0.02, 0.04) refers to a dummy variable indicating a student scored between 0.02 and 0.04 above and below the highest threshold. All the reported coefficients are relative to students who scored between 0.04 and 0.06 points from the highest threshold. Each specification includes individual-level controls such as age, fieldFE and academic yearFE. Only students that have chosen the most efficient combination of subjects, and therefore, shown intention to apply for the most demanded programs, have been selected. Robust standard errors are report in parenthesis. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

4.6.3 Gender Gaps in Unobserved Ability

Another potential mechanism that might explain the gender differences between mock exams during high school and the actual university entrance examination is disparities in latent ability between genders. The university entrance examination may comprise more cognitively demanding tests while high-school evaluations might consider other factors that positively affect female performance such as grit, behaviour, or attendance, among other non-cognitive factors. If high-performing male students possess greater unobserved ability than their female counterparts,

they would demonstrate a more significant advantage on the more demanding test.

Table 4.7 presents the results from an analysis similar to the one performed for Table 4.2, but here, students are further differentiated based on their high-school performance: those scoring zero to 0.04 points, and those scoring 0.04 to 0.08 points both above and below the highest threshold. My results show that the relative underperformance of females on the university entrance examination among students within zero to 0.04 points above the highest threshold is larger than for those high-performing students within 0.04 to 0.08 points above the highest threshold (Columns 1 and 3).

While the difference in scores between both evaluations for female students within 0.04 to 0.08 points above the highest threshold is 0.16 standard deviations lower than that for males, the gender difference in scores is 0.27 standard deviations for students within zero to 0.04 above the highest threshold. If unobserved ability was the primary factor, one might expect consistently larger gaps at all levels of high performance, not varying gaps based on proximity to the threshold. Thus, these findings suggest that factors other than just unobserved cognitive abilities might be influencing the gender gaps observed between high school evaluations and university entrance examination performance, particularly among students aiming for the most competitive programs.

Table 4.7: Gender Gap in Performance at Different Distances from the Highest Threshold.

	(1)	(2)	(3)	(4)
	Standardized Difference: Total			
	(0, 0.04)	(-0.04, 0)	(0.04, 0.08)	(-0.04, -0.08)
female	-0.2712***	-0.1756**	-0.1614**	-0.2010***
	(0.076)	(0.069)	(0.082)	(0.065)
Controls: Age, field, YearFE,	x	x	x	x
Observations	637	599	529	760
R-squared	0.110	0.102	0.092	0.057

Note. Each cell is separate regression with the standardized difference between the university entrance examination and high-school average score as the dependent variable for students at different points from the predicted threshold. Only students that have chosen the most efficient combination of subjects, and therefore, shown intention to apply for the most demanded programs, have been selected. Robust standard errors are report in parenthesis. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

4.7 Conclusions

In this paper, I explore the gender gap in academic performance under different levels of competitive pressure, focusing on the university entrance examination —a set of externally administered standardized tests crucial in shaping educational pathways, akin to the Cito exam in the Netherlands, the SAT in the United States, and the Baccalauréat in France. Using population level data from Andalusia, Spain, linking administrative records from high school examination and university entrance examination, I explore gender-based reactions to competitive pressures by comparing performance in these distinct contexts with varying stakes.

My findings reveal a pronounced gender gap in high-stakes environments such as the university entrance examination, as opposed to lower-stakes settings like high school evaluations. Remarkably, these differences are predominantly evident in tests related to health sciences—including the most demanded and competitive university programs. The gender differences in reaction to higher competitive pressure result in a reverse in the likelihood of females qualifying for science programs based on the high-stakes university entrance examination scores, compared to their low-stakes high school performance.

While this study does not exhaustively examine all the underlying mechanisms, it offers suggestive evidence that the patterns observed are not solely attributable to gender differences in effort or unobserved male ability during high school. Further research is needed to explore various factors, including the impact of stereotype threat, which is found to be specially relevant in explaining the later underrepresentation of women in STEM fields (Niederle and Vesterlund, 2010; Nollenberger et al., 2016;Iriberri and Rey-Biel, 2017, among others).

An additional line for future research would be to examine how these gender gaps translate into subsequent labor market outcomes. My results suggest that women's performance is more susceptible to prior achievements or failures in high-pressure situations, potentially influencing the future decision-making process and thereby the later likelihood of success. Despite a global decrease in gender disparities in education and labor market participation, women continue to make educa-

tional choices leading to lower expected labor market earnings compared to men (Bertrand, 2020).

Chapter 5

General Conclusions

In the first essay, I have studied the impact of a comprehensive educational policy reform in Spain on infant health outcomes, finding significant reductions in very low birth weight and preterm births. These outcomes are attributed to increased maternal labor supply and better family planning, facilitated by more general education in high school curricula. The second essay presents evidence on the effectiveness of the CAL language program in improving literacy skills, particularly for students at the lower end of the test-score distribution. This essay highlights the program's personalized and adaptive instructional design as key to its success. The third essay focuses on gender differences in academic performance under varying levels of competitive pressure, revealing a pronounced gender gap in high-stakes environments like university entrance examinations, especially for students applying to science or highly competitive programs.

A common element across these essays is the profound impact of education on various outcomes, such as health, literacy, and gender gaps. Each study underscores the significance of adapting educational policies and programs to address specific needs and challenges, whether it is integrating general education into curricula, employing technology for personalized learning, or understanding gender dynamics in educational settings.

While the essays in this thesis offer evidence on how education can help reduce inequalities across generations, learning levels, and genders, there is still a need for further research. In what follows, I will outline potential areas for future investiga-

tion in each of the key topics addressed in this thesis: health endowments, literacy skills, and gender gaps.

In the first essay, the first essay shows how integrating more general education into high school curricula leads to improved health outcomes at birth. However, little is yet known about which specific elements of the new comprehensive curricula (such as reading and writing skills, mathematical understanding, or particular areas of study) are driving improvements in infant health outcomes, information that would provide targeted guidance for policymaking. In addition, since my sample is limited to children born to mothers up to the age of 33, future research could explore the broader implications of educational curricula on maternal labor, health, and social outcomes across the entirety of their reproductive lives. Digging into the long-term health effects on children stemming from changes in their mothers' educational curriculum is also an interesting area of study, as I only observe the health outcomes at birth. Looking at the effects of curricula changes on other outcomes, such as prenatal care visits, mental health issues like anxiety or depression, and maternal earnings, could provide deeper insights into the pathways through which maternal education influences infant health at birth.

The second essay, which evaluates the effectiveness of a CAL language program, opens avenues for additional research. Future studies could examine the persistence of the program's potential to foster long-lasting improvements in academic performance and other areas, such as labor and behavioral outcomes. Additional research is needed to explore the key elements contributing to the success of CAL programs, especially the distinct effects of personalized and adaptive content, the importance of feedback in the learning process, and the outcomes associated with various usage modes, including in-school and at-home applications.

In the third essay, I explore gender disparities in academic performance across different degrees of competitive pressure, underscoring the need for additional research into the mechanisms that may account for the widening of these gaps in favor of men. A significant area for further investigation is the array of factors that create these gender gaps, focusing particularly on how competitive pressure intersects

with stereotype threat—a major element in understanding women's underrepresentation in STEM fields. A key question arises: to what extent does the observed underperformance of women in high-stakes science university entrance exams stem from their response to heightened competitive pressure and stereotype threat? Another confounding factor that has been only partially addressed in the third essay is the differences in evaluation (blind versus not blind) characteristics between high- and low-stakes settings, which may account for part of the widening of the gender gap through teacher grading bias. Furthermore, it is essential to study how these gender gaps in academic settings translate into subsequent labor market outcomes, particularly examining the impact of women's performance under pressure on their future decision-making and success, to gain a deeper understanding of the enduring gender inequalities in education and the workforce.

Appendix A

Chapter 1 Appendix

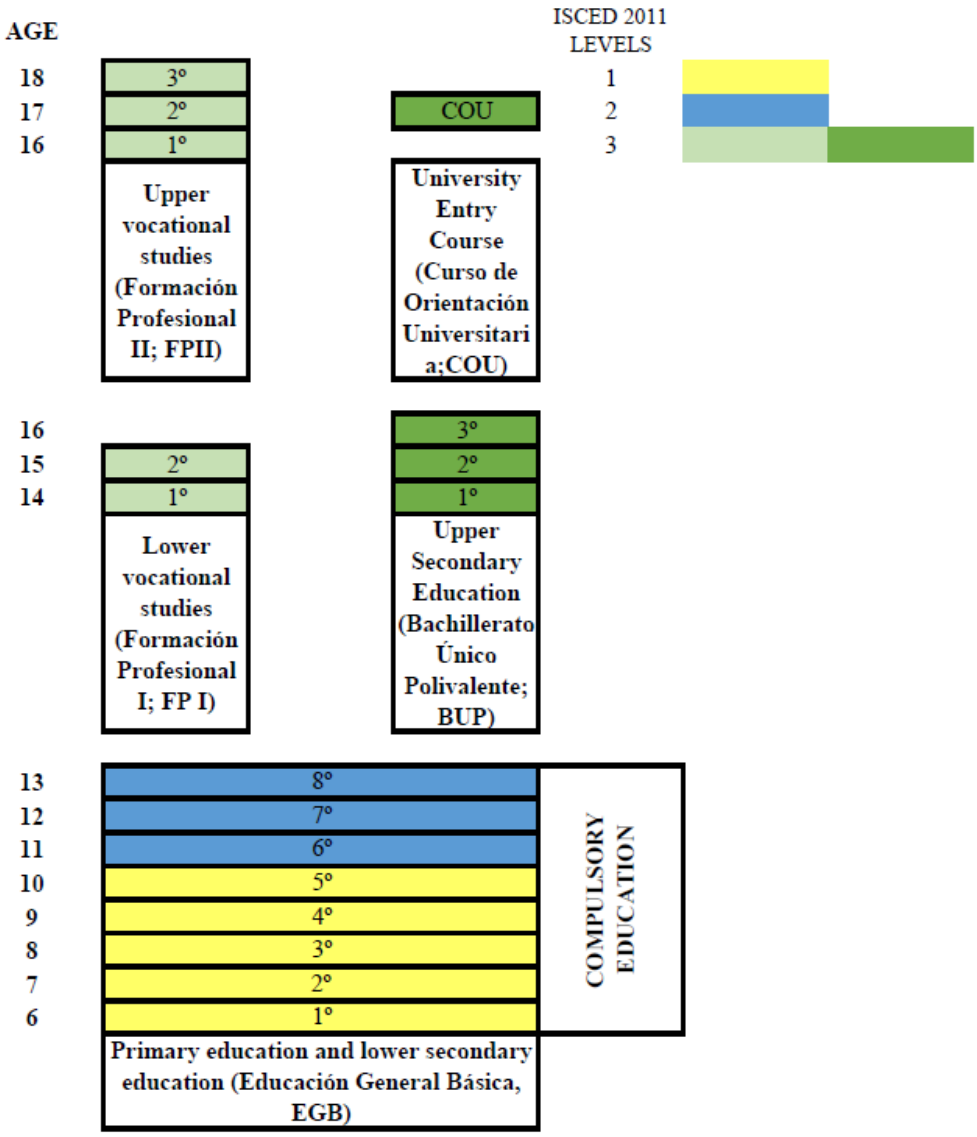


Figure A.1: Main Pathways of the Spanish Education System Before the LOGSE
Notes: The artistic education program (Music and Dance and Dramatic Arts) is not included. Source: Spanish Education Ministry.

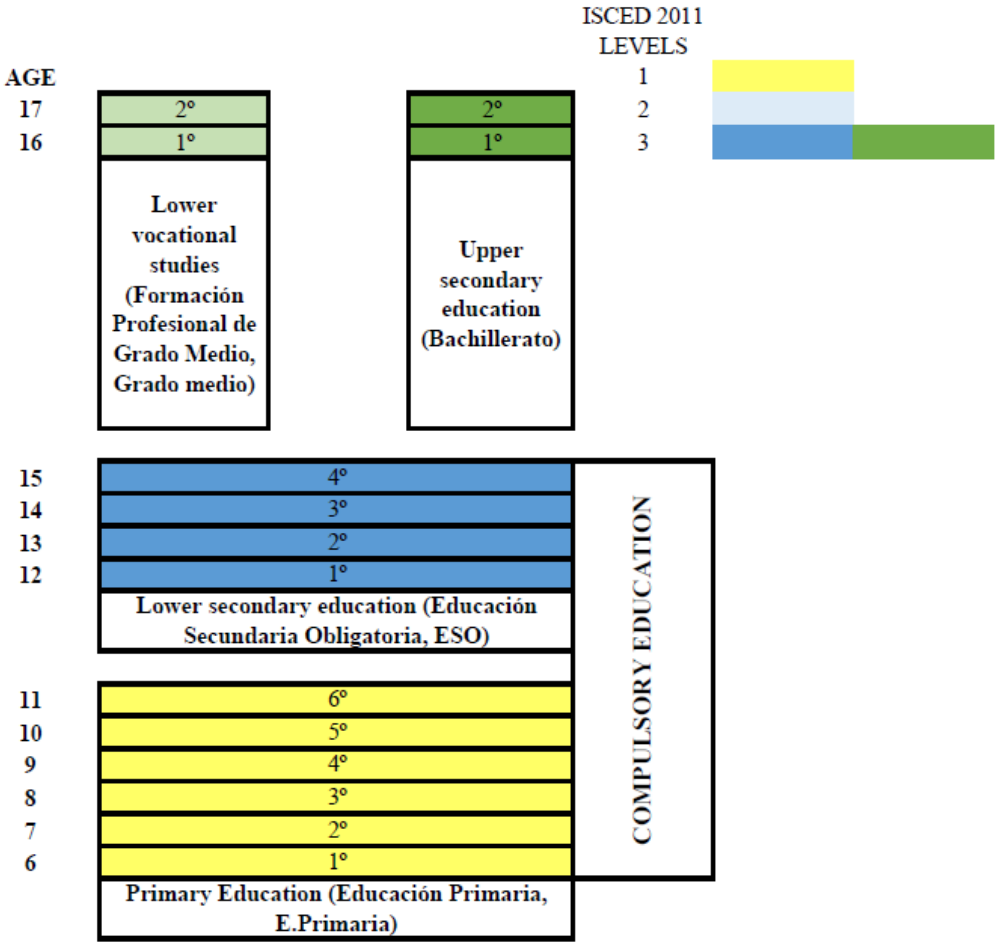


Figure A.2: Main Pathways of the Spanish Education System After the LOGSE.
Notes: The artistic education program (Music and Dance and Dramatic Arts) is not included. Source: Spanish Education Ministry.

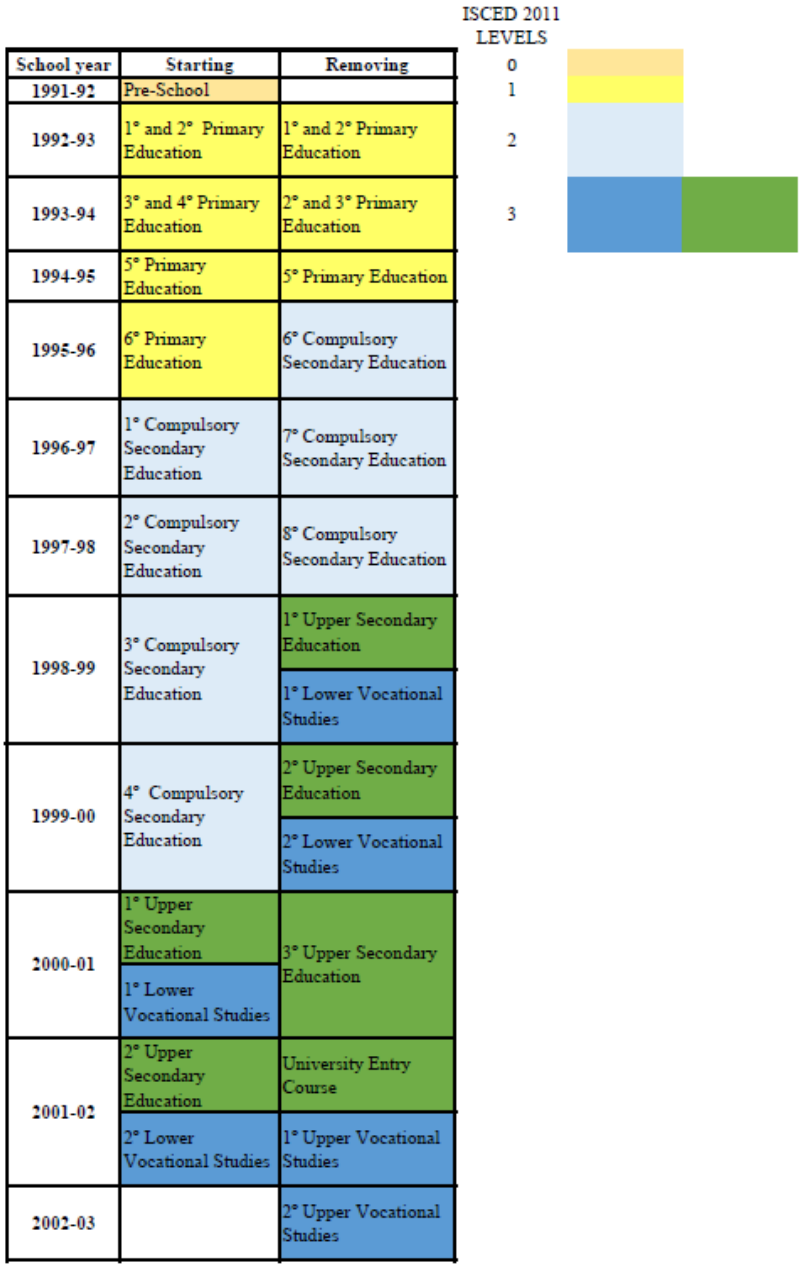


Figure A.3: National Calendar of LOGSE Implementation

Notes: Updated by R.D. 173/1998; Artistic Education Program not included (Music and Dance and Dramatic Arts). Source: Education Ministry.

Table A.1: Educational Curricula Before and After the Reform At Ages 14 and 15

Age	Pre-Reform System		Post-Reform System
	Vocational Education	Academic Education	
14			
	5 Occupation-Specific Subjects	Math	Math
	Apprenticeships	Social Sciences	Social Sciences
	Spanish Language	Spanish Language	Spanish Language and Literature
	Humanistic Education	Natural Sciences	Natural Sciences
	Foreign Language	Foreign Language	Foreign Language
	Physical Education	Physical Education	Physical Education
	Civic, Social and Political Education	Drawing	Education in Civic and Ethical Values
	Religious Education	Religious Education	Religious Education (voluntary)
		Music and Arts	Music and Arts
			Technology and Digitalisation
			1 Optional Subject (Second Foreign Language or Classical Culture)
15			
	5 Occupation-Specific Subjects	Math	Math
	Apprenticeships	Social Sciences	Social Sciences
	Spanish Language	Spanish Language and Literature	Spanish Language and Literature
	Foreign Language	Foreign Language	Foreign Language
	Physical Education	Physical Education	Physical Education
	Civic, Social and Political Education	Latin	Education in Civic and Ethical Values
	Religious Education	Religious Education	Religious Education or Study Time (voluntary)
	Humanistic Education	1 Occupation-Specific Subject	3 Academic Field-Specific Subjects (Biology and Geology, Physics and Chemistry, Music and Arts, Technology, Second Foreign Language, Classical Culture)
		Natural Sciences	

Notes. Educational curricula corresponding to lower vocational studies (*FP I*, ISCED 2011 level 3) and the first two years of upper secondary education (*BUP* ISCED level 3) of the previous educational system (*LGE*) and the last two years of lower secondary education (*ESO* ISCED 2011 level 3) of the new educational system. Source: Laws (Real Decreto 160/1975, Real Decreto 707/1976, Real Decreto 1007/1991).

Table A.2: Health at Birth Summary Statistics

	Mean	SD	Min	Max	Definition
Weight	3202.492	513.35	42	6580	Weight at birth in grams of the first-born.
Low Weight	0.072	0.26	0	1	Dummy variable equal to 1 if the weight at birth of the first-born is under 2500 grams; 0 otherwise.
Very Low Weight	0.008	0.09	0	1	Dummy variable equal to 1 if the weight at birth of the first-born is lower than 1500 grams; 0 otherwise.
Preterm	0.128	0.33	0	1	Dummy variable equal to 1 if the first-born is born under 38 weeks of gestation; 0 otherwise.
Very Preterm	0.012	0.11	0	1	Dummy variable equal to 1 if the first-born is born under 33 weeks of gestation; 0 otherwise.
Weeks of Gestation	39.158	1.90	19	46	Number of weeks of gestation of the first-born.
Fetal Death	0.001	0.03	0	1	Dummy variable equal to 1 if the first-born is born dead; 0 otherwise.
Survive 24h after Birth	0.999	0.04	0	1	Dummy variable equal to 1 if the first-born survives the first 24 hours after the birth; 0 otherwise.
N	**1,521,770**				

Notes. Sample: Female Spanish nationals aged 25-33 at their first birth and born between 1975 and 1985. Source: Prepared by the authors from the 2000-2018 Vital Statistics (INE) childbirth microdata for the 1975-1985 birth year cohorts.

Table A.3: Maternal Background Characteristics Summary Statistics

	Mean	SD	Definition
Student	0.009	0.095	Dummy variable equal to 1 if the mother is a student at any kind stage of education, 0 otherwise.
Working Mother	0.860	0.345	Dummy variable equal to 1 if the mother is not a homemaker, 0 otherwise.
Married	0.644	0.479	Dummy variable equal to 1 if the mother is married, 0 otherwise.
Marriage Age	26.597	2.896	Mother's age at first marriage.
Qualified Job	0.454	0.498	Dummy variable equal to 1 if the mother works a highly trained job (managerial position, scientific or academic profession, administrative or office worker, qualified personnel in primary sector, qualified personnel in secondary sector and construction), 0 otherwise.
Non-Qualified Job	0.272	0.445	Dummy variable equal to 1 if the mother is employed in a non-skilled job (Catering, protection and sales workers, plant and machinery operators and assemblers, or elementary occupations)
Mate Qualified Job	0.434	0.496	Dummy variable equal to 1 if the mother's mate works a highly trained job (managerial position, scientific or academic profession, administrative or office worker, qualified personnel in primary sector, qualified personnel in secondary sector and construction), 0 otherwise.
Mate Non-Qualified Job	0.398	0.489	Dummy variable equal to 1 if the mother's mate works a non-skilled job (Catering, personal, protection and sales workers, plant and machinery operators and assemblers, or elementary occupations)

Notes. Sample: Female Spanish nationals aged 17-33 at their first birth and born between 1975 and 1985. Source: Prepared by the authors from the 2000-2018 Vital Statistics (INE) childbirth microdata for the 1975-1985 birth year cohorts.

Table A.4: Education Summary Statistics

	Mean	SD	Definition
Panel A: High School Enrollment			
No Degree	0.010	0.10	Dummy variable equal to 1 if the highest educational degree is lower than primary school; 0 otherwise
Comprehensive Education	0.331	0.47	Dummy variable equal to 1 if the highest educational degree is compulsory education; 0 otherwise.
Academic Education	0.453	0.50	Dummy variable equal to 1 if the highest educational degree is academic secondary (post-compulsory) education or college; 0 otherwise.
Vocational Education	0.157	0.36	Dummy variable equals to 1 if the highest educational degree is vocational secondary education; 0 otherwise.
N	**109,339**		
Panel B: Degree Completion			
No Degree	0.011	0.10	Dummy variable equal to 1 if the highest educational degree is lower than primary school; 0 otherwise
High School Degree	0.337	0.22	Dummy variable equal to 1 if the highest educational degree is lower or upper secondary education; 0 otherwise.
College Degree	0.361	0.23	Dummy variable equal to 1 if the highest educational degree is college; 0 otherwise.
Vocational Degree	0.254	0.19	Dummy variable equal to 1 if the highest educational degree is lower or upper vocational education; 0 otherwise.
Age at Highest Qualification	20.209	18.13	Age in years at completion of highest qualification.
N	**85,348**		

Notes. Panel A presents summary statistics on high school enrollment among female Spanish nationals aged 17-25 and born between 1975 and 1985. Panel B presents summary statistics on degree completion for sample of women within the age range of 25 to 33. Source: 1991-2018 Spanish LFS.

Table A.5: Adult Health Summary Statistics

	Mean	SD	Min	Max	Definition
Lung Cancer	0.039	0.487	0	21	Number of female hospitalizations due to lung cancer for each patient's province of residence and cohort (1975-1985)
Diabetes	2.281	4.129	0	59	Number of female hospitalizations due to diabetes for each patient's province of residence and cohort (1975-1985)
Cirrhosis	0.314	1.323	0	28	Number of female hospitalizations due to cirrhosis for each patient's province of residence and cohort (1975-1985)
Hypertension	0.151	1.210	0	41	Number of female hospitalizations due to hypertension for each patient's province of residence and cohort (1975-1985)
N	**11,388**				

Notes. Sample: Female Spanish nationals aged 25-31 and born between 1975 and 1985. Source: 2004-2015 MSBD.

Table A.6: Identification Check #1: The LOGSE Exposure Index and Macroeconomic Outcomes

	(1) GDP per capita	(2) Female Employment Rate	(3) Female Labor Participation Rate
Index	-0.4598	0.0090	0.0068
	(0.291)	(0.008)	(0.010)
Mean	10.543	0.263	0.364
Std. Dev.	3.33	0.06	0.06
Observations	535	535	535
R-squared	0.963	0.904	0.880

Notes. Table reports OLS coefficients. Standard errors are in parentheses. Each specification includes controls for cohort and province of residence. Data from the Autonomous Cities of Ceuta and Melilla is not included. Data are from 1991 to 2001, Spanish Statistical Office. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

Table A.7: Identification Check #2: Education Outcomes and LOGSE Entry

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year of LOGSE Implementation							
No Degree rate of 1975 cohort	28.6246				33.1545*			
	(22.371)				(18.763)			
Compulsory Secondary Education rate of 1975 cohort		4.3962				3.6060		
		(3.047)				(3.296)		
Academic Secondary Education or Higher rate of 1975 cohort			-3.8084				-3.8101	
			(2.452)				(2.401)	
Vocational Education rate of 1975 cohort				-0.1318				2.1913
				(3.018)				(3.179)
Macroeconomic Controls	N	N	N	N	Y	Y	Y	Y
Observations	50	50	50	50	50	50	50	50
R-squared	0.034	0.048	0.062	0.000	0.203	0.191	0.224	0.170

Notes. Standard errors are in parentheses. The estimates are obtained from estimating eq. 2.3 on a sample of female Spanish nationals aged 17-31 and born between 1975 and 1985. All variables are measured at the provincial level for the cohort of 1975. Data are from the 1991-2018 Spanish LFS. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

Table A.8: Identification Check #2: Prior Health Birth Outcomes and LOGSE Entry

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Year of LOGSE Implementation															
Panel A: Weight at Birth Measures																
Weight	0.0035								0.0011							
	(0.006)								(0.007)							
Low Weight		49.1793								38.6231						
		(33.683)								(33.457)						
Very Low Weight			87.3747								83.7812					
			(121.557)								(119.881)					
Panel B: Gestational Age																
Late Preterm				12.5678								4.0966				
				(9.124)								(10.456)				
Very Preterm					-23.5020								-8.0243			
					(29.556)								(23.975)			
Weeks						-0.7408								-0.0826		
						(1.581)								(1.583)		
Panel C: Mortality At Birth																
Fetal Death							165.7732								194.1798	
							(164.345)								(128.048)	
Survive 24h After Birth								-146.8872								-138.6498
								(140.708)								(109.080)
Macroeconomic Controls	N	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y
Observations	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50
R-square	0.007	0.074	0.020	0.037	0.006	0.005	0.025	0.026	0.165	0.208	0.182	0.168	0.165	0.164	0.198	0.187

Notes. Standard errors are in parentheses. The estimates are obtained from estimating eq. 2.3 on a sample of female Spanish nationals aged 25-31 and born between 1975 and 1985. All variables are measured at the provincial level for the cohort of 1975. Data are from the 1995-2013 Vital Statistics (INE) childbirth microdata. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

Table A.9: Identification Check #3: Placebo Check for Spurious Correlations Between School Enrollment and Degree Completion Differences Prior (1984-1989) to the LOGSE

	(1)	(2)	(3)	(4)	(5)
Panel A: School Enrollment					
	No Degree	Comprehensive Education	Academic Education	Vocational Education	
Lag Index	0.001	0.026	0.009	0.008	
	(0.008)	(0.027)	(0.031)	(0.020)	
Observations	148,370	148,370	148,370	148,370	
R-squared	0.006	0.012	0.024	0.014	
Panel B: Degree Completion					
	Age at Highest Qualification	No Degree	Secondary Degree	Vocational Degree	College Degree
Lag Index	0.582*	0.002	-0.036	0.026	0.037
	(0.320)	(0.010)	(0.031)	(0.024)	(0.027)
Romano Wolf P-Value	0.999				
Observations	106,171	126,764	126,764	126,764	126,764
R-squared	0.029	0.009	0.017	0.016	0.020

Notes. Standard errors are in parentheses. The estimates are obtained from estimating eq. 2.1 on a sample of female Spanish nationals aged 17-33 and born between 1970 and 1975. All specifications include a constant and main controls for birth year and province of residence. Standard errors are clustered at province level for each specification. Romano-Wolf p-values based on 1,000 studentized bootstrap replications. Data are from the 1987-2018 Spanish LFS. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

Table A.10: Identification Check #4: Placebo Check for Spurious Correlations Between Health at Birth Outcomes Differences Prior (1984-1989) to the LOGSE

	(1)	(2)	(3)
Panel A: Weight at Birth Measures			
	Weight	Low Weight	Very Low Weight
Lag Index	0.3735	0.0035	0.0016
	(6.007)	(0.003)	(0.001)
Obs	903,569	903,569	903,569
R-squared	0.002	0.000	0.000
Panel B: Gestational Age			
	Weeks	Late Preterm	Very Preterm
Lag Index	-0.0142	-0.0023	0.0019
	(0.024)	(0.004)	(0.001)
Obs	808,176	808,176	808,176
R-squared	0.002	0.001	0.000
Panel C: Mortality at Birth			
	Fetal Death	Survive 24h after Birth	
Lag Index	0.0002	0.0000	
	(0.000)	(0.000)	
Obs	949,274	949,274	
R-squared	0.000	0.000	

Notes. Standard errors are in parentheses. The estimates are obtained from estimating eq. 2.1 on a sample of female Spanish nationals aged 25-33 and born between 1970 and 1975. All specifications include a constant and main controls for birth year and province of residence. Standard errors are clustered at province level for each specification. Data are from the 1995-2013 Vital Statistics (INE) childbirth microdata. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

Table A.11: Reform Effect on Fertility Patterns

	(1) Birth Rate	(2) Motherhood Entry Age
Index	-0.0012	-0.0630
	(0.004)	(0.043)
1975's Cohort Mean	0.069	29.518
Std.Dev	0.03	2.36
Obs	4,995	1,521,770
R-squared	0.543	0.010

Notes. Standard errors are in parentheses. The estimate is obtained from estimating eq. 2.1 on a sample of female Spanish nationals aged 25-33 and born between 1975 and 1985. Birth Rate (Column 1) is calculated as the number of first births divided by the number of women born in Spain per mother age and province. All specifications include a constant and main controls for birth year and province of residence. Motherhood Entry Age (Column 2) is mother's age at first birth. Standard errors are clustered at province level for each specification. Data are from the 1975-1985 and 2000-2018 Vital Statistics (INE) childbirth microdata. ***Significant at the 1% level, **Significant at the 5% level, * Significant at the 10% level.

Appendix B

Chapter 2 Appendix

Table B.1: Literacy Interventions

(1) Reference	(2) Country	(3) Intervention	(4) Method	(5) Outcome	(6) Effect Direction	(7) Effect Size	(8) Sample Size	(9) Cost
Banerjee and Duflo, 2016	India	(1) Summer Camp: volunteers teach a curriculum adapted to the academic level of children based on their initial ability (2) Change in teaching approach: (1) School received learning material; (2) Teachers were trained to teach personalised curriculum; (3) teachers and volunteers received teaching material + training	RCT	ASER language and math tests	Math: Positive Effects Language: Positive Effects	(1) Summer Camp: Maths: +0.086 s.d. Language:+0.074 s.d. (2) Change in Treatment Approach: Maths: +0.11 s.d. Language: +0.13 s.d.	Around 15,000 students (from grade 1 to 5)	Not specified
Banerjee et al., 2007	India	(1) Remedial education program. Women work with struggling students. (2) CAL math program	RCT	School Test Scores (Math and Language)	Math: Positive Effects Language: Null Effects	(1) Remedial education program: Maths: First Year: +0.14s.d.; Second Year: +0.28s.d. Language: Null Effects (2) CAL math program: Maths. First Year: +0.35s.d.; Second Year: +0.47s.d. Language: Null Effects	(1) Remedial education program: First Year: 12,855 students; Second Year: 21,936 students (Grades 3 and 4) (2) CAL math program: First Year: 5,732 students; Second Year: 5,523 students (Grade 4)	15USD/year per student
Beg et al., 2019	Pakistan	Brief, expert-led, curriculum based videos integrated (1) into the classroom experience or (2) tablets.	RCT	School tests (Maths and Science)	(1) Integrated into class curriculum: Positive Effects (2) Provided through tablets: Negative Effects	(1) Integrated into class curriculum: + 0.3 s.d. (2) Provided through tablets: -0.4 s.d.	Around 3,000 students	9USD per student

Continued on Next Page…

Table B.1 – continued from previous page

Reference	Country	Intervention	Method	Outcome	Effect Direction	Effect Size	Sample Size	Cost
Bouguen, 2016	France	Teacher training program on reading skills to adapt instruction to the specific need of each student	Value-Added Model	Specialized Tests (vocabulary, letter recognition, comprehension, sounds recognition, segmentation, pseudo-reading, lexical reading)	Vocabulary: Null Effects Letter Recognition: Positive Effects Comprehension: Null Effects Sounds Recognition: Positive Effects Segmentation: Positive Effects Pseudo-reading: Positive Effects Lexical reading: Positive Effects	Vocabulary: Null Effects Letter Recognition: +0.135s.d. Comprehension: Null Effects Sounds Recognition: +0.179s.d. Segmentation:+ 0.244s.d. Pseudo-reading: +0.447s.d. Lexical reading: + 0.135s.d.	Around 1,300 students	211,8$ per student OECD exchange rate, 0,903)
Carlana and Ferrara, 2021	Italy	Online individual tutoring during lock-down	RCT	(1) Academic performance: standardized test scores (2) Aspirations, socio-emotional skills, and psychological well-being	Positive Effects	(1) Academic performance: +0.26 s.d. (2) Aspiration (+0.15 s.d.); perseverance (+0.12 s.d.) and; psychological well-being index (+0.17 s.d.)	Around 1,000 students	59.2$ per student (exhange rate of 2021)
Carneiro et al., 2022	Ecuador	Teacher coaching program that focused on teacher quality and teacher–child interactions	RCT	(1) Peabody Picture Vocabulary Test (PPVT) (2) Woodcock–Johnson battery of achievement tests	Null Effects	Null Effects	Around 1,800 students (grade 1)	Not specified
Duflo et al., 2011	Kenya	Hire an additional teacher and split the classroom into two groups based on their initial abilities.	RCT	Standardized Test Scores (Math and Language)	Math: Positive Effects Language: Positive Effects	Reading Assessment (PIRA) age-standardised test Key Stage 1 reading level	Around 10,000 students grade 1)	Not specified

Continued on Next Page…

Table B.1 – continued from previous page

Reference	Country	Intervention	Method	Outcome	Effect Direction	Effect Size	Sample Size	Cost
Fryer and Howard-Noveck, 2020	EEUU	After-school tutoring program	RCT	(1) The New York state ELA and math achievement tests (McGraw Hill) (2) School attendance	Null Effects	Null effects	Around 1,700 (grades 6 to 8)	2,500$/year per student
	EEUU	Incentives to motivate reading through a summer reading program (called Project READS) through arms: (1) weekly mailed books; (2) weekly mailed books + incentives to read and; (3) post-testing books.	RCT	(1) The Gates-MacGinitie reading test (GMRT) for vocabulary and reading (2) The Massachusetts Comprehensive Assessment System (MCAS) standardized tests	(1) Mailed books: Positive Effects (2) Maliked books + incentives (well matched books): Null Effects	(1) Mailed books: GMRT:Null MCAS: Null (2) Maliked books + incentives (well matched books): GMRT: +0.20 s.d. MCAS:+0.38 s.d.	Around 400 students	Not specified
Jacob, 2017	EEUU	Evidence-Based Literacy Instruction (EBLI) that aims to provide teachers with several instructional strategies.	RCT	Standardized reading and math scores from the Measures of Academic Progress (MAP)	Null Effects	Null Effects	Around 1,500 students (grades 2-5)	Not specified
Johnson et al., 2019	UK	Teaching assistants are trained to deliver a tightly structured package of materials, both with and without the use of ICT.	RCT	(1) Reading Assessment (PIRA) age-standardised test (2) Key Stage 1 reading level	Positive Effects	(1) ICT PIRA: Between +0.18 s.d. in the short-run and and not significant one year later Key Satege 1: Null (2) Non ICT PIRA: Between +0.27 s.d. in the short run and not significant one year later Key Stage 1: +6 p.p in expected reading level	Around 2,000 students (grades 1 and 2)	31.9$ per student
Kerwin and Thornton, 2021	Uganda	Early-primary teacher training literacy program	RCT	Early Grade Reading Assessment (EGRA)	Positive Effects	Between +0.45 s.d. and +0.64 s.d.	Around 1,400 students (grades 1 to 3)	19.88$ per student

Continued on Next Page…

Table B.1 – continued from previous page

Reference	Country	Intervention	Method	Outcome	Effect Direction	Effect Size	Sample Size	Cost
Keslair et al., 2012	UK	Adaptive Curriculum for Special Educational Needs (SEN)	DiD	Key Stage 2 Maths, English and Science	Null Effects	Null Effects	Around 3,000,000 students (around 80,000 students with special needs)	2,115$ per students
Lavecchia et al., 2020	Canada	Tutoring program (coaching, tutoring assistance, and post-secondary financial aid)	DiD	Longer-term impacts on (1) employment, (2) earnings and (3) tuition expenditures	Positive Effects	(1) Employed at age 28: between +14–16 \% (2) Annual Earnings at age 28: +19\% (3) Annual tuition expenditures between +47 to +100 \%	Around 46,000 students (grades 9–12)	13,400$ per student
Lavy et al., 2022	Israel	High school educational remedial program	Propensity Score Matching Methods	Longer-term impacts (1) completed years of college schooling, (2) annual earnings (3) months employed	Positive Effects	(1) Year of college schooling: +10 p.p. (2) Annual earnings: +4 p.p. (3) Months employed: +1.5 p.p.	Around 1,000 students (grades 7-12)	1,100$ per student
Loyalka et al., 2019	China	Teacher professional development (PD) program	RCT	Student Standardized Mathematics Tests	Null Effects	Null Effects	Around 33,000 students	22.6$ per student
Machin and McNally, 2008	UK	Introduction of a highly structured literacy hour	DiD + Propensity Score Matching Methods	(1) Key Stage 2 English standardized test. (2) Percentile reading score on the Key Stage 2 English standardized test. (3) Percentage of students achieving level 4 or above on the Key Stage 2 English standardized test.	Positive Effects	(1) Key Stage 2 English standardized test: Between +0.06 and +0.08 s.d. (2) A 2–3 p.p. improvement in the reading and English skills (3) Increase of the percentage of students achieving level 4 or above by about 3 p.p.	Around 1,600,000 students	46,9$/year per students (2008 exchange rate of 0,544)
Machin et al., 2007	UK	Application of ICT to teaching and learning in schools	IV	Share of students achieving level 4 or above in National tests: (1) English, (2) Science and (3) Maths	(1) English: Positive Effects (2) Science: Positive Effects (3) Maths: Null Effects	(1) English: +2 p.p. increase in the proportion of pupils achieving level 4 or above in English. (2) Science: +1.4 p.p. in the proportion of pupils achieving level 4 or above. (3) Maths: Null Effects	591 Local Education Authorities	112$ Primary schools and 128$ Secondary schools (exchange rate of 0.50 of 2007)

Continued on Next Page…

Table B.1 – continued from previous page

Reference	Country	Intervention	Method	Outcome	Effect Direction	Effect Size	Sample Size	Cost
Machin et al., 2018	UK	Practice changes shifted reading instruction to focus on 'synthetic phonics'.	DiD	(1) Teacher assessed standardised score in Communication, Language and Literacy. (2) Key Stage 1 reading (3) Externally assessed standardised test score in reading.	Positive Effects	Between zero and +0.30 s.d.	Around 300,000 students (Students ages 5, 7 and 11)	67,000£ per local authority
Özek, 2021	EEUU	Remedial courses in middle school	RDD	(1) ELA test scores (2) Instruction time (3) College credit-bearing courses (4) College selectivity (5) Persistence beyond first and second years (6) Degree attainment	Positive Effects	(1) ELA test scores: +0.11s.d. (2) Instruction time: +hour each day (3) College credits: +4.9 p.p. (4) Very competitive college: +4.6 p.p. (5) Persistence in college: +4.6 p.p. (first year) and +4.7 p.p. (second year) (6) Degree attainment: +3.7 p.p.	Around 25,000 students (grade 6 to 8)	Not specified

Table B.2: CAL Language Programs

(1) **Reference**	(2) **Country**	(3) **CAL Program**	(4) **Approach**	(5) **Method**	(6) **Effect Size**	(7) **Sample Size**	(8) **Teacher Role**	(9) **Dynamically Adaptive**	(10) **Personalized**	(11) **Fast Feedback**	(12) **Cost**
Bai et al., 2016	China	Reading with Orthographic and Segmented Speech (ROSS) programs	(1) Version 1: Curriculum Substitute (2) Version 2: Homework supplement	RCT	(1) Version 1: +0.16 s.d. in language (2) Version 2: no effect in language	Around 6,000 students	Limited teacher-assisted	No	No	Yes	Not specified
Bai et al., 2023	China	CAL remedial tutoring program	Homework supplement	RCT	Std language test scores (English): between +0.48 s.d. Std maths test scores: Null	Around 1,600 students	Limited teacher-assisted	No	No	Yes	12.01/14.32 USD per student
Borman et al., 2009	USA	Fast ForWord - computer-based language and reading training program	Classroom curriculum substitute	RCT (ITT) + IV (Dose-response)	2nd Grade: Null effects 7th Grade: Null effects on language and 0.21 standard deviation increase in reading	415 students	Limited teacher-assisted	No	No	Yes	Not specified
Campuzano et al., 2009	EEUU	Destination Reading - Course 1	Homework supplement	RCT	Null Effects	Around 3,000	Limited teacher-assisted	Yes	No	Yes	78 USD/ year per student
Campuzano et al., 2009	EEUU	Headprouts (CAL Instructional reading and writing program)	Homework supplement	RCT	+0.01 s.d.	Around 3,000	Limited teacher-assisted	Yes	Yes	Yes	146 USD/ year per student
Campuzano et al., 2009	EEUU	Programmed Logic for Automatic Teaching Operations (PLATO Focus)	Curriculum Substitute	RCT	Null Effects	Around 3,000	Highly teacher-assisted	Yes	No	Yes	351 USD/ year per student
Campuzano et al., 2009	EEUU	Waterford Early Reading Program - Levels 1-3	Homework supplement	RCT	Null Effects	Around 3,000	Limited teacher-assisted	Yes	Yes	Yes	223 USD /year per student
Campuzano et al., 2009	EEUU	CAL tutoring program (LeapTrack)	Curriculum Substitute	RCT	+0.09 s.d.	Around 3,000	Not specified	Yes	Yes	Yes	217 USD/year per student
Campuzano et al., 2009	EEUU	CAL tutoring program (Academy of Reading)	Curriculum Substitute	RCT	Null Effects	Around 3,000	Not specified	Yes	No	Yes	154 USD/year per student

Table B.2 – continued from previous page

Reference	Country	CAL Program	Approach	Method	Effect Size	Sample Size	Teacher Role	Dynamically Adaptive	Personalized	Fast Feedback	Cost
Carrillo et al., 2011	Ecuador	Personalized Complementary and Interconnected Learning (APCI) program (Más Tecnología)	Classroom curriculum substitute	RCT	Language: Null Maths: + 0.30 s.d.	Around 500 students	Highly teacher-assisted	No	Yes	Yes	Not specified
Deault et al., 2009	Canada	ABRACA- DABRA (web-based literacy program)	Classroom curriculum substitute	RCT	(1) Synthetic group:+ 0.41 s.d. on listening comprehension; null on vocabulary: and +0.35 s.d. on reading comprehension; (2) Analytic group: Null	144 students	Highly teacher-assisted	No	No	Yes	Not specified
Faber and Visscher, 2018	Holland	Snappet - digitized assignment tool focused on spelling	Homework supplement	RCT	Null Effects	1,605 students	Highly teacher-assisted	Yes	Yes	Yes	Not specified
Lai et al., 2016	China	CAL remedial tutoring program	Homework supplement	RCT	Standardised language test: +0.20sd Standardised math test:+0.15 s.d.	Around 3,000 students	Limited teacher-assisted	No	No	Yes	7.6 USD per student
Muralidharan et al., 2019	India	Mindspark (Technology-led instructional program)	Homework supplement	RCT (ITT) + IV (Dose-response)	(1) Language Tests: IIT: +0.23 s.d. IV: +0.39 s.d. (2) Math Tests: IIT: +0.37 s.d. IV: +0.6 s.d.	619 students	Highly teacher-assisted	Yes	Yes	Yes	150$ USD/ year per student
Rouse and Krueger, 2004	USA	Fast ForWord - computer-based language and reading training program	Classroom curriculum substitute	RCT (ITT) + IV (Dose-response)	Null Effects	485 students	Limited teacher-assisted	No	Yes	No	770$ USD/ year per student

Table B.2 – continued from previous page

Reference	Country	CAL Program	Approach	Method	Effect Size	Sample Size	Teacher Role	Dynamically Adaptive	Personalized	Fast Feedback	Cost
Wijekumar et al., 2012	EEUU	Intelligent Tutoring for.ture Strategy (ITSS)	Classroom curriculum substitute	RCT	(1) GSRT: +0.10 s.d. (2) Experimenter-designed measures of reading comprehension (main quality idea): + 0.49 s.d.	2, 643 students	N/A (Lab Experiment)	Yes	Yes	Yes	Not specified

Table B.3: Subtypes of Exercises Depending on the Linguistic Element

Linguistic Element	**Example of Exercise**
Most Basic Elements	
	The user needs to find the different symbol among the ones displaced.
[1] Non-symmetric symbols	*Find the different symbol <#> vs. <*>.*
[2] Symmetrical symbols	*Find the different symbol <[> vs. <]>.*
[3] Non-symmetrical alphanumeric signs	*Find the different letter <E> vs. <F>.*
[4] Symmetrical alphanumeric signs	*Find the different letter <q> vs. <p>.*
	The user sees a set of digits for some seconds
	and them s/he is asked to memorize them
	and write them down.
[5] Digits	*Look and then write <8 3 6>.*
	The user hears the name of a letter and s/he is asked to identify it from among the distractors
	within a time frame, using a Whac-A-Mole-style game interaction
[6] Vowel letters	*Click on ¡a¿.*
[7] Consonant letters	*Click on ¡g¿.*
[8] Mirror letters	*Click on ¡n¿.* It targets similar looking letters.
	These are letters that have mirror features i.e. 'u' and 'u'.
[9] Rotated letters	*Click on ¡p¿.*
	Letters with rotation features are (<b, d, p , q>)
[10] All letters.	*Look and then write ¡a i p¿.*
	The user hears the sound (phoneme) and has to map it with the letter displayed.
[11] Vowel sounds	*Click on [a].*
[12] Plosives sounds	*Click on [b].* Sounds of letters <b c d g k p qu t v w>.
[13] Nasals sounds	*Click on [m].* Sounds of letters <m n ñ>.
[14] Laterals and rhotics sounds	*Click on [l].* Sounds of letters <l ll r rr>.
[15] Fricatives and affricates	*Click on [tS].* Sounds of letters <c ch f g j s x y z>.
	Targeting consonant vowel (CV) structure.
[16] Syllables with CV structure	*Click on <da>.*
[17] CV syllable memorization	*Look and then write <da ba ca>.*
[18] Syllables with VC structure	Click on <al>.
[19] VC syllable memorization	*Look and then write <al en el>.*
[20] Syllables with CVC structure	*Click on <sen>.*
[21] CVC syllable memorization	*Look and then write <dar sis mar>.*
[22] Syllables with CCV structure	*Click on <bla>.*
[23] CCV syllable memorization	*Look and then write <bla pla cre>.*
[24] Syllables with CVV structure	*Click on <cei>.*
[25] CVV syllable memorization	*Look and then write <pai fai cei>.*
[26] Syllables with CCVC structure	*Click on <tran>.*
Complex Elements	
[27] Simple words	*Click on <gato> (cat).* Words with two or less syllables.
[28] Complex words	*Click on <elefante> (elephant).* Words with three or more syllables.
[29] Words with transparent spelling	*Write the word <tigre> (tiger).* Words with transparent orthography, where the word
	is written as it sounds.
[30] Words with arbitrary spelling	*Write the word <búho> (owl).* The word has an opaque orthography, that is, there is
	no regular correspondence between letter and sound.
[31] Deletion of letter/s in words	**patto → pato (duck).* The user needs to perform an operation (deletion, insertion,
	substitution or reorder) to create a correct word.
[32] Insertion of letter/s in words	**ambre → hambre (hunger).*
[33] Substitution of letter/s in words	**crase → clase (class).*
[34] Order of letter/s in words	**u g a a → agua (water).*
[35] Order of syllable in words	**ra do mo → morado (purple).*
[36] Simple word pairs	*<cara vara para> (face stick for).* The user is asked to match the same words.
[37] Complex word pairs	<comida *cómodo cómico*> *(food confort funny)*
[38] Non-words	*Click on <modeme>*
[39] Write non-words	*Write <toti>.* A non-word is a word a does not exits but it is phonetically possible.
[40] Non-word pairs	*<nuita tuira cuira>*
Most Difficult Elements	
[41] Suffixes	pastel + [ición — ería — edor] → pastelería *baker + [ious — y — ive] → bakery*
[42] Prefixes	[dis — re — tras] + hacer → rehacer. *[trans — re — anti] + do → redo.*
[43] Compound words	toma — corta — porta + uñas → cortauñas.
[44] Word segmentation	*apartirde → a partir de, *beginningwith → beginning with*
[45] Semantic errors	Click on the error: El río lleva muchas *pierdas. *The river carries many *loss* instad of *rocks*.
	The word '*pierdas' in Spainsh means 'loss' while 'piedras' -the correct one- means 'rocks'.
[46] Syntax errors	Click on the error: *Todo eran iguales aquí. The word '*Todo' should be 'Todos.'
	The user is asked to click on the error, in this case it is a syntax -or grammatical- error.
	*(*Everyones was the same here).*
[47] Transparent spelling errors	Click on the error: Consejos de *utilisación. *Advice of *ushe*
[48] Arbitrary spelling errors	Click on the error: Ayer *izo mucho calor. *It *wass very hot yesterday*
[49] Accent mark	Click on <hábito> vs. <habito> and <habitó>. The Accent mark differentiates meanings of
	words in Spanish, <hábito> *(habit)*, <habito> *(live)* and <habitó> *(lived).*

Notes. The most basic elements, including the easiest exercises, target symbols, letters, sounds, and syllables. Complex elements, encompassing medium-difficulty exercises, focus on words and non-words. Exercises that belong to the most difficult linguistic elements target word parts (morphemes) and sentences.

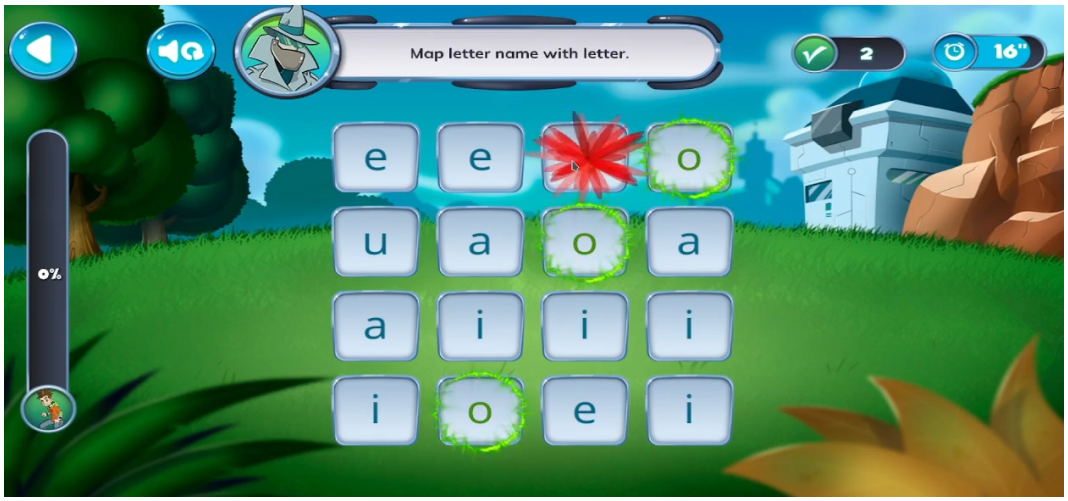
Table B.4: Cognitive Abilities and Performance Measures Used For The Personalized Exercises of Each Participant in the DytectiveU Computer-Assisted Learning Program.

Panel A: Cognitive Abilities	
Language Skills	Alphabetic Awareness
	Phonological Awareness
	Syllabic Awareness
	Lexical Awareness
	Morphological Awareness
	Syntactic Awareness
	Semantic Awareness
	Orthographic Awareness
Executive Functions	Activation of Attention
	Sustained Attention
	Simultaneous Attention
Perceptual Processes	Auditory Discrimination and Categorization
	Visual Discrimination and Categorization
Working Memory	Visual (alphabetical)
	Auditory (phonology)
	Sequential (auditory)
	Sequential (visual)
Panel B: Performance Measures	
	Reading Comprehension
	Reading Speed
	Natural Spelling
	Arbitrary Spelling
	Writing Speed
	Error Recognition
	Error Correction

Figure B.1: Example of User Profile Personalization

Note: This figure provides an illustration of user profile customization. Users have the capability to personalize their profiles according to the coins they have accumulated.

Figure B.2: Example Feedback in a Given Exercise



Note: This figure shows the manner in which feedback is delivered. In this exercise, correct selections are highlighted in green, while incorrect choices are indicated in red.

Figure B.3: Examples of Exercises In The DytectiveU Computer Assisted Learning Program

Notes. This figure shows three examples of exercises. On the left, the player has to select within a time limit the symbol that is different than the others (Visual Discrimination and Categorization). In the second exercise (middle), the player is asked to recognize the incorrect letter in a word and substitute it with the correct one (Phonological Awareness, Lexical Awareness, Reading Comprehension, and Arbitrary Spelling). In the third exercise (right), the player needs to identify and select all pairs of pseudowords (Phonological Awareness and Visual Discrimination and Categorization).

Table B.5: Overview of Main Datasets

Data Source	Data Availability (academic years)	Unit of Observation	Main Variables
Student Register	2017 - 2018 and 2018 - 2019	Student	Standardized Test Scores (Spanish and Mathematics); Gender; School Location; Inmigrant Status; School and Class Size.
Family Questionnaire	2018 - 2018 and 2018 - 2019	Student	Parents' educational level, parental investments (number of books and digitial devices at home), pre-primary enrollment,
Teacher and School Head Questionnaires	2019 - 2018 and 2018 - 2019	School and Group Class	Teacher Work Experience, Internet Connection
DytectiveU Dataset	2018 - 2019	School and Grade Level	Number of Student Logged in the DytectiveU; Results of Dytective Test (Risk or No Risk of Dyslexia); Date of the First Test and Last Test

Notes. The linkage between the student register and family questionnaire is done at student ID. The teacher and school head questionnaires are linked at school ID and class ID levels. The DytectiveU Dataset is linked to the rest of the datasets at School ID and Grade Level.

Table B.6: Definition of Main Variables

Variables	Definition
Panel A: Student Register	
Standardized Score in Spanish	The standardized scores in Spanish for the 2018-2019 standardized test in the Region of Madrid with respect to the mean and standard deviation.
Standardized Score in Maths	The standardized scores in Maths for the 2018-2019 standardized test in the Region of Madrid with respect to the mean and standard deviation
Female	Dummy Variable 1 - Student self-identifies as a female student. 0 - Student self-identifies as a male student.
3rd Grade	Dummy Variable 1 - The student is in 3rd Grade 0 - The student is in 6th Grade
School Size	Total number of students in 3rd Grade and 6th per school
Class Size	Total number of students per class
Panel B: Questionnaire Data	

Continued on Next Page...

Table B.6: Definition of Main Variables

Variables	**Definition**
Student Started After 3yo	Dummy Variable 1 - Student was enrolled in the education system by the age of 3. 0 - Student was not enrolled in the education system.
Inmigrant	Dummy Variable 1 - Student was not born in Spain 0 - Students was born in Spain
College Mother	Dummy Variable 1 - Mother has a college degree. 0 - Mother does not have a college degree.
College Father	Dummy Variable 1 - Father has a college degree. 0 - Father does not have a college degree.
Less than 50 Books at Home	Dummy Variable 1 - The student has at home less than 50 books. 0 - The student has at home more than 50 books.
Between 50 and 100 Books at Home	Dummy Variable 1 - The student has at home more than 50 books and less than 100 books at home. 0 - Otherwise
More than 100 Books at Home	Dummy Variable 1 - The student has at home more than 100 books at home. 0 - The student has less than 100 books at home.
More than 5 Digital Devices	Dummy Variable 1 - The student has at home more than 5 digital devices. 0 - The student does not have at home more 5 than digital devices.

Continued on Next Page...

Table B.6: Definition of Main Variables

Variables	**Definition**
Teacher More than 10 Years of Experience	Dummy Variable 1 - The teacher has more than 10 years of experience. 0 - The teacher has less than 10 years of experience.
Severe Internet Inconvenience	Dummy Variable 1 - School head reports severe internet inconvenience. 0 - Otherwise.
Moderate Internet Inconvenience	Dummy Variable 1 - School head reports moderate internet inconvenience. 0 - Otherwise.
Mild Internet Inconvenience	Dummy Variable 1 - School head reports mild internet inconvenience. 0 - Otherwise.
No Internet Inconvenience	Dummy Variable 1 - School head reports no internet inconvenience. 0 - Otherwise.
School Capital	Dummy Variable 1 - School is located in the Capital (Madrid city). 0 - Otherwise.
School East	Dummy Variable 1 - School is located in the East area of the Region of Madrid 0 - Otherwise.
School North	Dummy Variable 1 - School is located in the North area of the Region of Madrid 0 - Otherwise.
School West	Dummy Variable 1 - School is located in the West area of the Region of Madrid 0 - Otherwise.

Continued on Next Page...

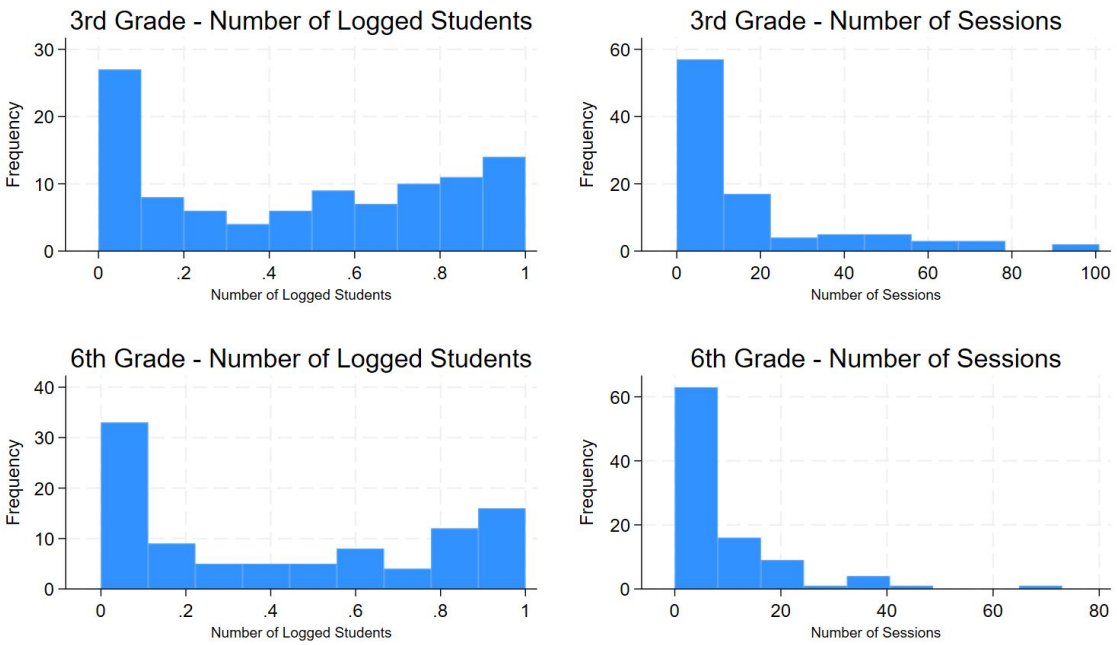
Table B.6: Definition of Main Variables

Variables	**Definition**
School South	Dummy Variable 1 - School is located in the South area of the Region of Madrid 0 - Otherwise.
Panel D: CAL Data	
Coverage	Proportion of students actively using DytectiveU before the standardized testing out of the total number who have taken the standardized tests at the school and grade levels in 2018 - 2019 academic year.
Number of Challenges	Number of completed challenges (sessions of 20 minutes) by the time of the Spanish standardized test.
Difficulty Level	Normalized value on a scale from 0 (lowest) to 100 (highest), representing the sum of linguistic levels achieved by the time of the Spanish standardized test (see Table B.3 of the Appendix).

Table B.7: Treatment and Control Groups

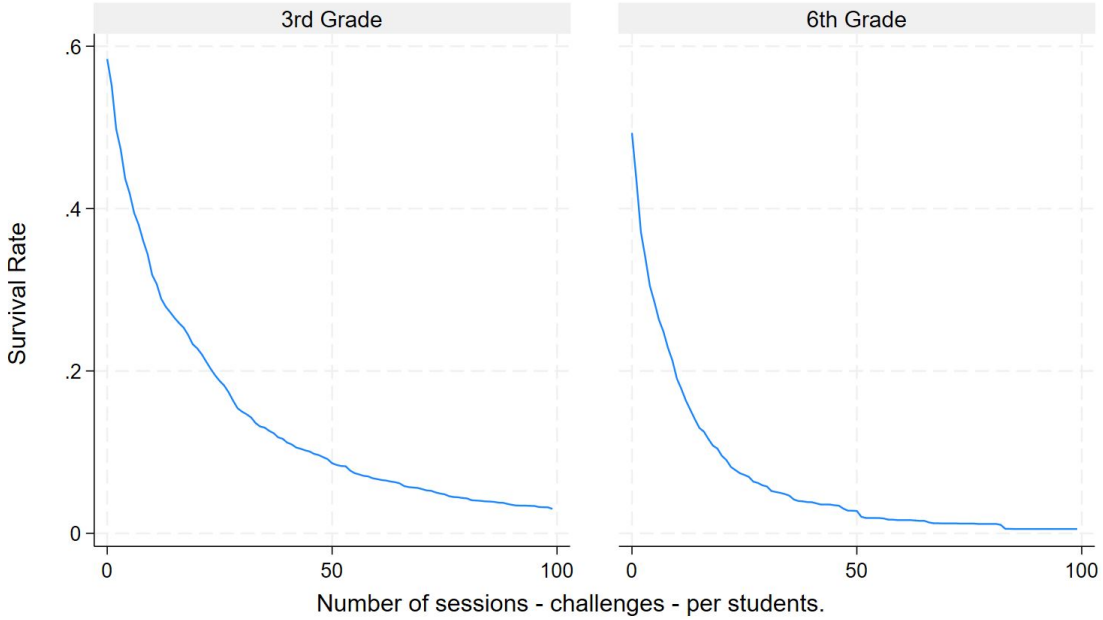
Groups	School Coverage	Entry
Treatment Group	91 schools	2018-2019
Control Group 1	167 schools	2020-2021
Control Group 2	173 schools	2021-2022

Figure B.4: Distribution of Take-up Among Treated Schools By School Grade



Notes. This figure shows the distribution of the number of logged students and mean number of sessions - challenges - by grade among treated schools.

Figure B.5: Student Retention Across Sessions



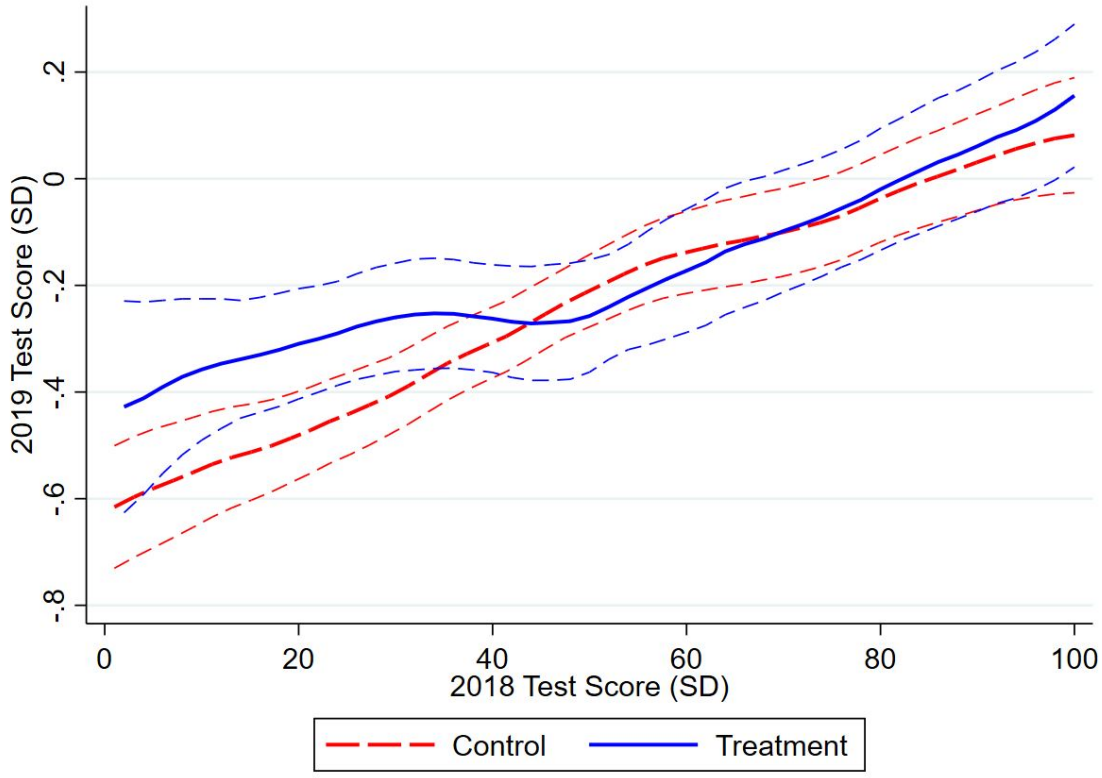
Notes. This figure shows the retention rates of students over successive sessions by school grade, illustrating the proportion of students who continue to participate in challenges.

Table B.8: Effects on Spanish Language Standardized Test: Excluding Students with Missing Questionnaire Data

	(1)
Panel A. Intent-To-Treat Estimates	
Treat	0.1008*
	(0.057)
R-squared	0.002
Panel B. Dose-Response Estimates	
Coverage	0.2144***
	(0.082)
R-squared	0.004
Sample Size	8,810
Number of Schools	214
Controls	NO

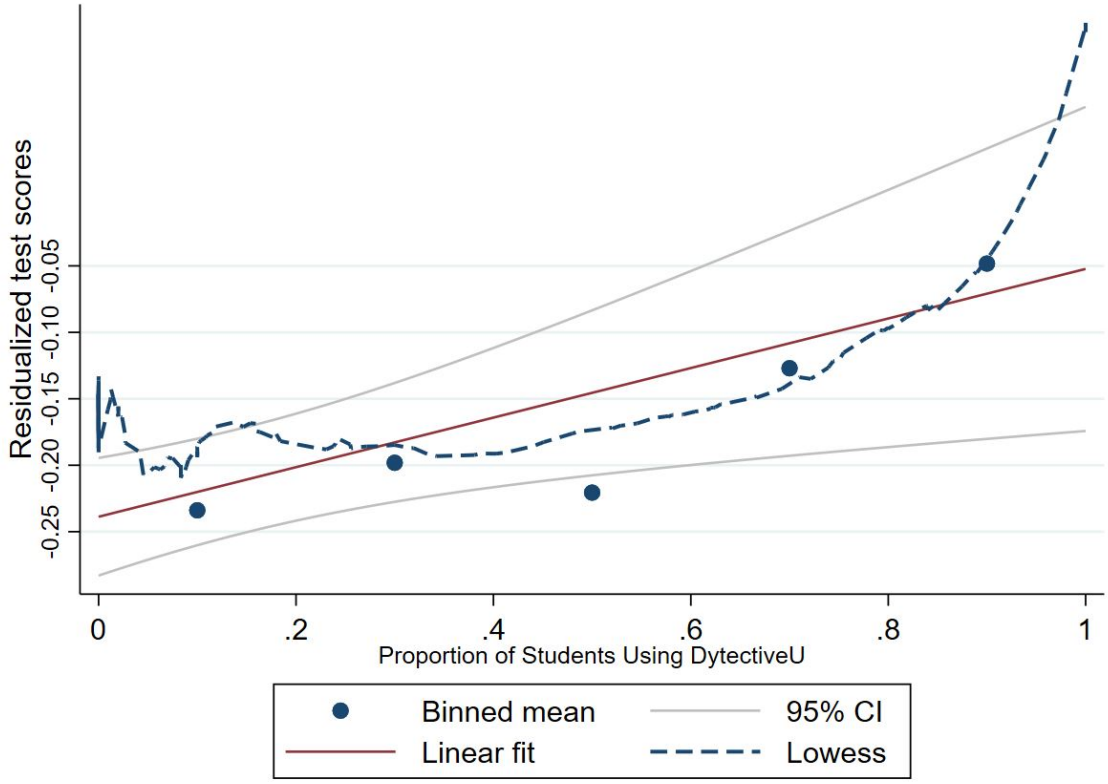
Notes. Outcome variable: standardized score in 2019 Spanish test. The unit of observation is student i in school s, grade c and class group c. Intent-To-Treat Estimates from eq. 3.1 are shown in Panel A. Treat is an intention to treat dummy equals to 1 for students in the treatment schools. Dose-response estimates from eq. 3.2 are shown in Panel B. Proportion of students using DyectiveU the fraction of students logged in DytectiveU by grade and school. Sample is restricted to students that report the 2018-2019 families', teachers' and school head's questionnaires. Standard errors (in parentheses) are robust and clustered at school level. ***Significant at 1% level, **Significant at 5% level, *Significant at 10% level.

Figure B.6: Non-Parametric Investigation of Treatment Effects By Pre-Intervention Performance Percentiles



Notes. The figure presents kernel-weighted local mean smoothed plots which relate 2019 test scores in Spanish to percentiles in the 2018 Spanish test scores, separately for the treatment and control groups, alongside 95confidence intervals.

Figure B.7: Dose-Response Relationship



Notes. This presents the relationship between the 2019 Spanish standardized score in Spanish test and the proportion of students actively using DytectiveU among the treated schools. It presents the mean standardized score in bins of proportion of students along with a linear fit and a lowess smoothed non-parametric plot.

Table B.9: Quadratic Dose-response Relationship

	(1)	(2)
Coverage	0.0736	0.3360
	(0.364)	(0.358)
Coverage(square)	0.1915	-0.1805
	(0.415)	(0.411)
R-squared	0.005	0.117
Sample Size	22,430	9,151
Number of Schools	269	221
Controls	NO	YES

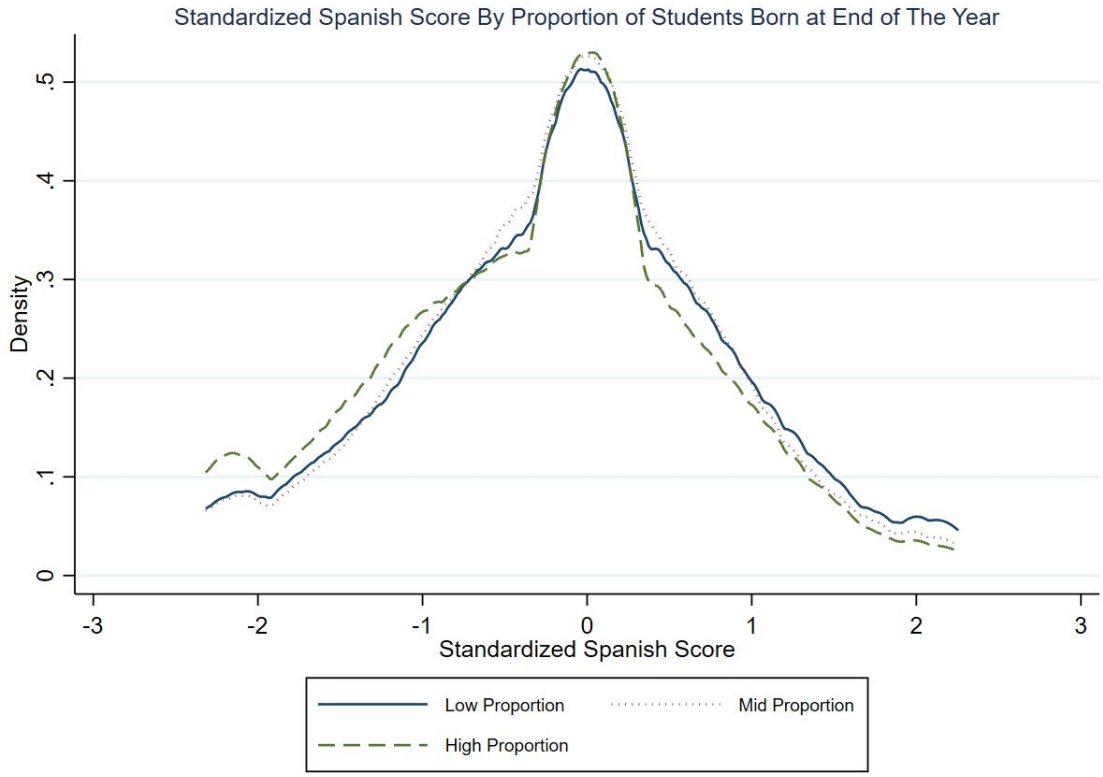
Notes. Outcome variable: 2019 standardized score in Spanish test. The unit of observation is student i in school s, grade level g and class group c.All regressions controls for students' characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection and teacher's years of experience. Standard errors (in parentheses) are robust and clustered at school level. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

Table B.10: Distributional Effects on Mathematics Standardized Test

	(1) .05Q	(2) .1Q	(3) .25Q	(4) .5Q	(5) .75Q	(6) .9Q	(7) .95Q
Panel A. Intent-To-Treat Estimates							
Treat	0.0844*	0.0400	0.0586	0.0607	0.0692	0.0857	0.1252
	(0.047)	(0.051)	(0.052)	(0.055)	(0.070)	(0.080)	(0.114)
R-square	0.081	0.093	0.098	0.099	0.100	0.097	0.090
P-value Parente-Santos Silva test	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel B. Dose-Response Estimates							
Coverage	0.1557**	0.0811	0.1706**	0.1787*	0.1593	0.2262	0.3568**
	(0.075)	(0.076)	(0.087)	(0.094)	(0.121)	(0.139)	(0.153)
R-square	0.084	0.094	0.100	0.101	0.102	0.098	0.092
P-value Parente-Santos Silva test	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Number of Students	8,836	8,836	8,836	8,836	8,836	8,836	8,836
Controls	YES	YES	YES	YES	YES	YES	YES

Notes. Outcome variable: standardized score in 2019 mathematics test. The unit of observation is student i in school s, grade level g and class group c. ITT estimates from eq. 3.1 are shown in Panel A. Treat is a dummy equals to 1 for students in the treatment schools. Dose-response estimates from eq. 3.2 are shown in Panel B. Proportion of students using DyectiveU is the fraction of students logged in DytectiveU by grade and school. All regressions include controls for students' characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection and teacher's years of experience Sample: students in 3rd and 6th grades from schools that implemented DytectiveU between 2018-2019 to 2021-2022. Standard errors (in parentheses) are robust and clustered at school level. ***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

Figure B.8: Pre-Intervention Distribution of The Standardized Score in the 2018 Spanish Test by Proportion of Students Born At The End Of The Year



Notes. The graph plots the distribution of the 2018 Spanish standardized score for grade 3rd and 6th and by schools with low, mid and high proportion of students born between November and December. Schools are classified as low proportion if they are in the first tertile, mid if in the second and high if in the third

Table B.11: Robustness Check #1. Effects on Spanish Performance - Remove Outliers: Deleted Top and Bottom 10% of Proportion of Active DytectiveU Users

	(1)	(2)
Panel A. Intent-To-Treat Estimates		
Treat	0.1183**	0.1122**
	(0.059)	(0.056)
R-squared	0.003	0.118
Panel B. Dose-Response Estimates		
Coverage	0.2464**	0.2434***
	(0.110)	(0.092)
R-squared	0.003	0.118
Sample Size	20,518	8,326
Number of Schools	253	208
Controls	NO	YES

Notes. Outcome variable: standardized score in 2019 Spanish test. The unit of observation is student *i* in school *s*, grade *c* and class group *c*. Intent-To-Treat Estimates from eq. 3.1 are shown in Panel A. Treat is a dummy variable equals to 1 for students in the treatment schools. Dose-response estimates from eq. 3.2 are shown in Panel B. Proportion of students using DyectiveU is the fraction of students logged in DytectiveU by grade and school. Controls include students' characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection and teacher's years of experience.***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

Table B.12: Robustness Check #2: The Effects of Imputing Mean Values to Missing Family Data on Students' Spanish Performance

	(1)	(2)
Panel A. Intent-To-Treat Estimates		
Treat	0.1207**	0.1069***
	(0.050)	(0.040)
R-squared	0.003	0.107
Panel B. Dose-Response Estimates		
Coverage	0.2323***	0.2048***
	(0.076)	(0.062)
R-squared	0.005	0.108
Sample Size	22,430	22,430
Number of Schools	269	269
Controls	NO	YES

Notes. Outcome variable: standardized score in 2019 Spanish test. The unit of observation is student i in school s, grade c and class group c. Intent-To-Treat Estimates from eq. 3.1 are shown in Panel A. Treat is a dummy variable equals to 1 for students in the treatment schools. Dose-response estimates from eq. 3.2 are shown in Panel B. Proportion of students using DyectiveU is the fraction of students logged in DytectiveU by grade and school. Controls include students' characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection and teacher's years of experience.***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

Table B.13: Robustness Check #3: The Effects of Assigning Missing Value to Outcomes on Students' Spanish Performance

	(1)	(2)
Panel A. Intent-To-Treat Estimates		
Treat	0.1019*	0.1013**
	(0.056)	(0.051)
R-squared	0.002	0.117
Panel B. Dose-Response Estimates		
Coverage	0.2158***	0.1924**
	(0.081)	(0.076)
R-squared	0.004	0.117
Sample Size	9,151	9,151
Number of Schools	221	221
Controls	NO	YES

Notes. Outcome variable: standardized score in 2019 Spanish test. The unit of observation is student *i* in school *s*, grade *c* and class group *c*. Intent-To-Treat Estimates from eq. 3.1 are shown in Panel A. Treat is a dummy variable equals to 1 for students in the treatment schools. Dose-response estimates from eq. 3.2 are shown in Panel B. Proportion of students using DyectiveU is the fraction of students logged in DytectiveU by grade and school. Controls include students' characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection and teacher's years of experience.***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

Table B.14: Heterogeneity in treatment effect by grade, gender and maternal educational attainment

	(1)	(2)	(3)	(4)	(5)	(6)
	Grade		Gender		Mother College	
A. Intent-To-Treat Estimates						
Treat	0.1400**	0.1314**	0.1015*	0.0686	0.0948*	0.1182**
	(0.064)	(0.060)	(0.054)	(0.058)	(0.053)	(0.053)
Covariate	-0.0562	0.2166	0.3963***	0.3747	0.4090***	0.3351
	(0.065)	(0.382)	(0.027)	(0.256)	(0.040)	(0.285)
Interaction	0.0552	0.0559	-0.0215	-0.0665	0.0206	0.0380
	(0.075)	(0.088)	(0.032)	(0.042)	(0.052)	(0.049)
R-Square	0.003	0.119	0.040	0.118	0.048	0.118
Panel B. Intensity Estimates						
Coverage	0.2405**	0.1761*	0.2101**	0.2007**	0.1867**	0.2097***
	(0.100)	(0.106)	(0.082)	(0.082)	(0.076)	(0.075)
Covariate	-0.0159	0.2096	0.3801***	0.5518**	0.4284***	0.3439
	(0.035)	(0.389)	(0.016)	(0.247)	(0.030)	(0.285)
Interaction	-0.0441	0.0300	0.0104	-0.0093	-0.0375	-0.0378
	(0.111)	(0.118)	(0.048)	(0.066)	(0.084)	(0.087)
R-Square	0.004	0.120	0.041	0.119	0.049	0.119
Observations	22,603	9,541	22,599	9,541	10,690	9,541
Controls	NO	YES	NO	YES	NO	YES

Notes. Outcome variable: standardized score in 2019 Spanish test. The unit of observation is student *i* in school *s*, grade *c* and class group *c*. Intent-To-Treat Estimates from eq. 3.1 are shown in Panel A. Treat is a dummy variable equals to 1 for students in the treatment schools. Dose-response estimates from eq. 3.2 are shown in Panel B. Proportion of students using DyectiveU is the fraction of students logged in DytectiveU by grade and school. Controls include students' characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection and teacher's years of experience.***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

Table B.15: Intensive and Extensive Margins

	(1)	(2)	(3)	(4)
	Number of Students		Number of Challenges	
Coverage	0.2154***	0.1825**	0.0001	0.0005
	(0.075)	(0.076)	(0.002)	(0.002)
Observations	22,603	8,810	14,832	5,906
R-squared	0.004	0.119	0.000	0.123
Controls	NO	YES	NO	YES

Notes. Outcome variable: standardized score in 2019 Spanish test. The unit of observation is student i in school s, grade c and class group c. Controls include students' characteristics (gender, age, pre-primary enrollment, immigrant status), parents' educational level, parental investments (number of books at home and digital devices), school and class size, school location, school's reported internet connection and teacher's years of experience.***Significant at 1% level, **Significant at 5% level, * Significant at 10% level.

Appendix C

Chapter 3 Appendix

Table C.1: Core Subjects in Upper Secondary Education (Bachillerato)

		Core Subjects		
		Sciences	**Humanities and Social Sciences**	**Arts**
First Year	General Core Subjects	Philosophy Spanish First Foreign Language Mathematics I	Philosophy Spanish First Foreign Language Latin I (Humanities) Applied Mathematics (Social Sciences)	Philosophy Spanish First Foreign Language Foundations of Art
	Specific Core Subjects	Biology and Geology Technical Drawing I Chemistry and Physics	Economy Greek I Contemporary World History Universal Literature	Culture Audio-visual I Contemporary World History Universal Literature
Second Year	General Core Subjects	History of Spain Spanish First Foreign Language Mathematics II	History of Spain Spanish First Foreign Language Latin II (Humanities) Applied Mathematics II (Social Sciences)	History of Spain Spanish First Foreign Language Foundations of Art
	Specific Core Subjects	Biology Technical Drawing II Physics Geology Chemistry	Economy Geography Greek II History of Art History of Philosophy	Performing Arts Audio-visual Culture Design

Table C.2: Field Specific Subjects in Upper Secondary Education (*Bachillerato*)

	Specific Subjects
First Year	Physical Education (compulsory) Musical Analysis I Pathological Anatomy Scientific Culture Artistic Drawing I Technical Drawing I Musical Language Christian Religion Second Foreign Language I Industrial Technology I Information and Communication Technology Visual and Plastic Arts Non-taken core subject
Second Year	Musical Analysis II Environmental Sciences Artistic Drawing II Technical Drawing II Administration Management Philosophy History Music History Image and Sound Phycology Christian Religion Second Foreign Language II Graphic Expression Techniques Industrial Technology II Information and Communication Technology II Non-taken core subject

Table C.3: Fields of Study and Subject Combinations

Field of Study	Subject Combinations
Arts and Humanities	- Latin II, History of Philosophy, Fundamentals of Art
	- Latin II, History of Philosophy, Greek
	- Latin II, History of Philosophy, Art History
	- Latin II, History of Philosophy, Second Foreign Language
	- Latin II, Fundamentals of Art, Greek
	- Latin II, Fundamentals of Art, Art History
	- Latin II, Fundamentals of Art, Second Foreign Language
	- Latin II, Greek, Art History
	- Latin II, Greek, Second Foreign Language
	- Latin II, Art History, Second Foreign Language
	- Fundamentals of Art, History of Philosophy, Greek
	- Fundamentals of Art, History of Philosophy, Art History
	- Fundamentals of Art, History of Philosophy, Second Foreign Language
	- Fundamentals of Art, Greek, Art History
	- Fundamentals of Art, Greek, Second Foreign Language
	- Fundamentals of Art, Art History, Second Foreign Language
Health Sciences	- Mathematics II, Biology, Chemistry
Social and Legal Sciences	- Latin II, History of Philosophy, Social Sciences Mathematics
	- Latin II, History of Philosophy, Audiovisual Culture
	- Latin II, History of Philosophy, Design
	- Latin II, History of Philosophy, Business Economics
	- Latin II, History of Philosophy, Geography
	- Latin II, History of Philosophy, Greek
	- Latin II, Social Sciences Mathematics, Audiovisual Culture
	- Latin II, Social Sciences Mathematics, Design
	- Latin II, Social Sciences Mathematics, Business Economics
	- Latin II, Social Sciences Mathematics, Geography
	- Latin II, Social Sciences Mathematics, Greek II
	- Latin II, Audiovisual Culture, Design
	- Latin II, Audiovisual Culture, Business Economics
	- Latin II, Audiovisual Culture, Geography
	- Latin II, Audiovisual Culture, Greek
	- Latin II, Design, Business Economics
	- Latin II, Design, Geography
	- Latin II, Design, Greek
	- Latin II, Business Economics, Geography
	- Latin II, Business Economics, Greek
	- Latin II, Geography, Greek
	- Social Sciences Mathematics, History of Philosophy, Audiovisual Culture
	- Social Sciences Mathematics, History of Philosophy, Design
	- Social Sciences Mathematics, History of Philosophy, Business Economics
	- Social Sciences Mathematics, History of Philosophy, Geography
	- Social Sciences Mathematics, History of Philosophy, Greek
	- Social Sciences Mathematics, Audiovisual Culture, Design
	- Social Sciences Mathematics, Audiovisual Culture, Business Economics
	- Social Sciences Mathematics, Audiovisual Culture, Geography
	- Social Sciences Mathematics, Audiovisual Culture, Greek
	- Social Sciences Mathematics, Design, Business Economics
	- Social Sciences Mathematics, Design, Geography
	- Social Sciences Mathematics, Design, Greek
	- Social Sciences Mathematics, Business Economics, Geography
	- Social Sciences Mathematics, Business Economics, Greek
	- Social Sciences Mathematics, Geography, Greek
Engineering and Architecture	- Mathematics, Technical Drawing, Physics
Sciences	- Mathematics, Biology, Technical Drawing
	- Mathematics, Biology, Physics
	- Mathematics, Biology, Chemistry
	- Mathematics, Technical Drawing, Physics
	- Mathematics, Technical Drawing, Chemistry
	- Mathematics, Physics, Chemistry

Notes. The subject combinations listed in this table are specifically tailored to align with different fields of study. In university applications, these field-specific subjects are assigned a higher weight when students apply to university programs that correspond to their chosen field.

Table C.4: Summary Statistics

	Full Sample	Arts and Humanities	Social and Legal Sciences	Health Sciences	Engineering and Architecture	Sciences
% Female Students	55%	74%	59%	66%	25%	58%
High-School Average Score	7.452	7.593	7.234	8.067	7.884	8.078
College Entrance Examination Average Score	6.082	6.423	5.879	6.388	6.412	6.416
Observations	287,550	11,816	54,508	52,710	19,391	66,404

Notes. Summary Statistics on a sample of student who sit for the University Entrance Examination ordinary call (June) in Andalusia (Spain) between 2010 and 2019. High-School and University Entrance Examination tests are graded in the same scale, from 0 to 10 points. Fields identified uniquely for those students who sit for the voluntary area-specific tests.

Table C.5: Highest Thresholds Summary Statistics

	Full Sample	Arts and Humanities	Social and Legal Sciences	Health Sciences	Engineering and Architecture	Sciences
Highest Threshold	12.22	12.26	11.69	12.71	12.07	12.77
Fraction of Students above the highest Threshold	7.63%	7.92%	6.57%	6.74%	14.38%	5.77%
College Access Score	10.47	10.77	9.92	10.59	10.46	10.92

Notes. The data for the university access thresholds are obtained the regional online portal. Data available here: `https://www.juntadeandalucia.es/economiaconocimientoempresasyuniversidad/sguit/?q=grados&d=g_not_cor_anteriores_top.php`. The fraction of students scoring above each of the highest threshold were calculated based on the distribution of university access score.

Table C.6: Placebo Test: Gender Gaps in Admissions to the Most Competitive University Programs - Based on the Highest Threshold from 2 Years Prior

	(1) (+0.015,-0.015)	(2) (+0.025,-0.025)	(3) (+0.03,+0.04) and (-0.04, -0.03)	(4) (-0.04,-0.06) and (-0.06, -0.04)
female	-0.2124***	-0.2878***	-0.2987***	-0.2624***
	(0.073)	(0.061)	(0.083)	(0.068)
Controls: Age, Field, YearFE	X	X	X	X
Observations	588	841	486	685
R-squared	0.064	0.070	0.137	0.069

Notes: Each column represents a separate regression, with the standardized difference between high school and university entrance examination average scores as the dependent variable. This is calculated for students at varying distances from the highest threshold set two years prior, based on their high school performance. The analysis includes only those students who chose the most efficient combination of subjects, thereby demonstrating an intention to apply for the most demanded programs. Robust standard errors are reported in parentheses. Significance levels are indicated at 1%, 5%, and 10%.

Bibliography

Aizer, A. (2011). Poverty, Violence, and Health: The Impact of Domestic Violence During Pregnancy on Newborn Health. *Journal of Human Resources*, 46(3):518–538.

Aizer, A. and Currie, J. (2014). The intergenerational transmission of inequality: Maternal disadvantage and health at birth. *Science*, 344(6186):856–861.

Almond, D., Chay, K., and Lee, D. (2004). The Costs of Low Birth Weight. Technical Report w10552, National Bureau of Economic Research, Cambridge, MA.

Almond, D. and Currie, J. (2011). Killing Me Softly: The Fetal Origins Hypothesis. *Journal of Economic Perspectives*, 25(3):153–172.

Almond, D., Currie, J., and Duque, V. (2018). Childhood Circumstances and Adult Outcomes: Act II. *Journal of Economic Literature*, 56(4):1360–1446.

Almond, D. and Mazumder, B. (2011). Health Capital and the Prenatal Environment: The Effect of Ramadan Observance During Pregnancy. *American Economic Journal: Applied Economics*, 3(4):56–85.

Amuedo-Dorantes, C., Arenas-Arroyo, E., and Sevilla, A. (2018). Immigration enforcement and economic resources of children with likely unauthorized parents. *Journal of Public Economics*, 158:63–78.

Anderberg, D., Bagger, J., Bhaskar, V., and Wilson, T. (2019). Marriage Market Equilibrium, Qualifications, and Ability. *SSRN Electronic Journal*.

Angrist, J. D. and Keueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics*, 60(1):47.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). GENDER DIFFERENCES IN RESPONSE TO BIG STAKES: Gender Differences in Response to Big Stakes. *Journal of the European Economic Association*, 14(6):1372–1400.

Bai, Y., Mo, D., Zhang, L., Boswell, M., and Rozelle, S. (2016). The impact of integrating ICT with teaching: Evidence from a randomized controlled trial in rural schools in China. *Computers & Education*, 96:1–14.

Bai, Y., Tang, B., Wang, B., Mo, D., Zhang, L., Rozelle, S., Auden, E., and Mandell, B. (2023). Impact of online computer assisted learning on education: Experimental evidence from economically vulnerable areas of China. *Economics of Education Review*, 94:102385.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*, 122(3):1235–1264.

Banerjee, A. V. and Duflo, E. (2016). Structured Study Time, Self-Efficacy, and Tutoring. Unpublished.

Barfield, W. D. (2018). Public Health Implications of Very Preterm Birth. *Clinics in Perinatology*, 45(3):565–577.

Basu, A., Jones, A. M., and Dias, P. R. (2018). Heterogeneity in the impact of type of schooling on adult health and lifestyle. *Journal of Health Economics*, 57:1–14.

Becker, G. S. (1965). A Theory of the Allocation of Time. *The Economic Journal*, 75(299):493.

Beg, S., Lucas, A., Halim, W., and Saif, U. (2019). Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not. Technical Report w25704, National Bureau of Economic Research, Cambridge, MA.

Behrman, J. R. and Rosenzweig, M. R. (2004). Returns to Birthweight. *Review of Economics and Statistics*, 86(2):586–601.

Bellés-Obrero, C. and Duchini, E. (2021). Who benefits from general knowledge? *Economics of Education Review*, 85:102122.

Bertrand, M. (2020). Gender in the Twenty-First Century. *AEA Papers and Proceedings*, 110:1–24.

Bertrand, M., Mogstad, M., and Mountjoy, J. (2020). Improving Educational Pathways to Social Mobility: Evidence from Norway's "Reform 94". *Journal of Labor Economics*, page 713009.

Bitler, M. P. and Currie, J. (2005). Does WIC work? The effects of WIC on pregnancy and birth outcomes. *Journal of Policy Analysis and Management*, 24(1):73–91.

Björklund, A. and Salvanes, K. G. (2011). Education and Family Background. In *Handbook of the Economics of Education*, volume 3, pages 201–247. Elsevier.

Black, S., Devereaux, P., and Salvanes, K. (2004). Fast Times at Ridgemont High? The Effect of Compulsory Schooling Laws on Teenage Births. Technical Report w10911, National Bureau of Economic Research, Cambridge, MA.

Black, S. and Devereux, P. (2010). Recent Developments in Intergenerational Mobility. Technical Report w15889, National Bureau of Economic Research, Cambridge, MA.

Borman, G. D., Benson, J. G., and Overman, L. (2009). A Randomized Field Trial of the Fast ForWord Language Computer-Based Training Program. *Educational Evaluation and Policy Analysis*, 31(1):82–106.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting Event Study Designs: Robust and Efficient Estimation. Publisher: arXiv Version Number: 4.

Bouguen, A. (2016). Adjusting content to individual student needs: Further evidence from an in-service teacher training program. *Economics of Education Review*, 50:90–112.

Braghieri, L., Levy, R., and Makarin, A. (2022). Social Media and Mental Health. *American Economic Review*, 112(11):3660–3693.

Brunori, P., Ferreira, F. H. G., and Peragine, V. (2013). *Inequality of Opportunity, Income Inequality and Economic Mobility: Some International Comparisons*. Policy Research Working Papers. The World Bank.

Bulman, G. and Fairlie, R. (2016). Technology and Education. In *Handbook of the Economics of Education*, volume 5, pages 239–280. Elsevier.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, Competitiveness, and Career Choices*. *The Quarterly Journal of Economics*, 129(3):1409–1447.

Cai, X., Lu, Y., Pan, J., and Zhong, S. (2019). Gender Gap under Pressure: Evidence from China's National College Entrance Examination. *The Review of Economics and Statistics*, 101(2):249–263.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2021). Difference-in-Differences with a Continuous Treatment. Publisher: arXiv Version Number: 2.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Campuzano, L., Dynarski, M., Agodini, R., and Rall, K. (2009). Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts. NCEE 2009-4041. Technical report, U.S. Department of Education.

Carlana, M. and Ferrara, E. L. (2021). Apart But Connected: Online Tutoring and Student Outcomes During the COVID-19 Pandemic. Discussion Paper DP15761, CEPR.

Carneiro, P., Cruz-Aguayo, Y., Intriago, R., Ponce, J., Schady, N., and Schodt, S. (2022). When promising interventions fail: Personalized coaching for teachers in a middle-income country. *Journal of Public Economics Plus*, 3:100012.

Carneiro, P., Meghir, C., and Parey, M. (2013). Maternal education, home environments, and the development of children and adolescents. *Journal of the European Economic Association*, 11:123–160.

Carrell, S., Fullerton, R., and West, J. (2009). Does Your Cohort Matter? Measuring Peer Effects in College Achievement. *Journal of Labor Economics*, 27(3):439–464.

Carrillo, M. S., Alegría, J., Miranda, P., and Sánchez Pérez, N. (2011). Evaluación de la dislexia en la escuela primaria: Prevalencia en español. *Escritos de Psicología / Psychological Writing*, 4(2):35–44.

Chou, S.-Y., Liu, J.-T., Grossman, M., and Joyce, T. (2010). Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan. *American Economic Journal: Applied Economics*, 2(1):33–61.

Commission, T. E. (2022). National Education Systems. Spain. Educational support and guidance. Guidance and counselling in early childhood and school education.

Conti, G., Heckman, J. J., Lopes, H. F., and Rémi, P. (2010). Constructing economically justified aggregates: An application to the early origins of health. *University of Chicago, Department of Economics*.

Crawford, C., Dearden, L., and Meghir, C. (2007). When You Are Born Matters: The Impact of Date of Birth on Child Cognitive utcomes in England. Technical report, The Institute for Fiscal Studies, London, UK.

Cunha, F. and Heckman, J. (2007). The Technology of Skill Formation. *American Economic Review*, 97(2):31–47.

Currie, J. (2009). Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development. *Journal of Economic Literature*, 47(1):87–122.

Currie, J. (2011). Inequality at Birth: Some Causes and Consequences. *American Economic Review*, 101(3):1–22.

Currie, J. and Moretti, E. (2003). Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings. *The Quarterly Journal of Economics*, 118(4):1495–1532.

Currie, J. and Neidell, M. (2005). Air Pollution and Infant Health: What Can We Learn from California's Recent Experience? *The Quarterly Journal of Economics*, 120(3):1003–1030. Publisher: Oxford University Press.

Currie, J., Stabile, M., Manivong, P., and Roos, L. L. (2010). Child Health and Young Adult Outcomes. *Journal of Human Resources*, 45(3):517–548.

Cygan-Rehm, K. and Maeder, M. (2013). The effect of education on fertility: Evidence from a compulsory schooling reform. *Labour Economics*, 25:35–48.

De Chaisemartin, C. and D'Haultfoeuille, X. (2022). Difference-in-Differences Estimators of Intertemporal Treatment Effects. Technical Report w29873, National Bureau of Economic Research, Cambridge, MA.

de Educación, M. (2022). Estadística de las enseñanzas no universitarias. alumnado con necesidad específica de apoyo educativo curso 2020-2021. Technical report, Subdirección General de Estadística y Estudios, Madrid.

de Estadística (INE), I. N. (2015). Survey on the Labour Market Insertion of University Graduates 2014.

De Paola, M. and Gioia, F. (2014). Who Performs Better Under Time Pressure? Results from a Field Experiment. *SSRN Electronic Journal*.

de Universidades, M. (2020). Datos y cifras del Sistema Universitario Español. Publicación 2019-2020. Technical report, Secretaría General Técnica del Ministerio de Universidades, Madrid.

Deault, L., Savage, R., and Abrami, P. (2009). Inattention and Response to the ABRACADABRA Web-Based Literacy Intervention. *Journal of Research on Educational Effectiveness*, 2(3):250–286.

Dhuey, E. and Lipscomb, S. (2010). Disabled or young? Relative age and special education diagnoses in schools. *Economics of Education Review*, 29(5):857–872.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5):1739–1774.

Eren, O., Depew, B., and Barnes, S. (2017). Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics*, 153:9–31.

Escueta, M., Nickow, A. J., Oreopoulos, P., and Quan, V. (2020). Upgrading Education with Technology: Insights from Experimental Research. *Journal of Economic Literature*, 58(4):897–996.

Faber, J. M. and Visscher, A. J. (2018). The effects of a digital formative assessment tool on spelling achievement: Results of a randomized experiment. *Computers & Education*, 122:1–8.

Felgueroso, F., Gutiérrez-Domènech, M., and Jiménez-Martín, S. (2014). Dropout trends and educational reforms: the role of the LOGSE in Spain. *IZA Journal of Labor Policy*, 3(1):9.

Ferrara, E. L., Chong, A., and Duryea, S. (2012). Soap Operas and Fertility: Evidence from Brazil. *American Economic Journal: Applied Economics*, 4(4):1–31.

Fertig, A. R. and Watson, T. (2009). Minimum drinking age laws and infant health outcomes. *Journal of Health Economics*, 28(3):737–747.

Fischer, M., Gerdtham, U.-G., Heckley, G., Karlsson, M., Kjellsson, G., and Nilsson, T. (2021). Education and Health: Long-Run Effects of Peers, Tracking and Years. *Economic Policy*, page eiaa027.

Fryer, R. G. and Howard-Noveck, M. (2020). High-Dosage Tutoring and Reading Achievement: Evidence from New York City. *Journal of Labor Economics*, 38(2):421–452.

Galama, T., Lleras-Muney, A., and Kippersluis, H. v. (2018). The Effect of Education on Health and Mortality: A Review of Experimental and Quasi-Experimental Evidence. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.

Garlick, R. (2018). Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment. *American Economic Journal: Applied Economics*, 10(3):345–369.

Gaviria, A. and Raphael, S. (2001). School-Based Peer Effects and Juvenile Behavior. *Review of Economics and Statistics*, 83(2):257–268.

Geruso, M. and Royer, H. (2018). The Impact of Education on Family Formation: Quasi-Experimental Evidence from the UK. Technical Report w24332, National Bureau of Economic Research, Cambridge, MA.

Gneezy, U. and Rustichini, A. (2004). Gender and Competition at a Young Age. *American Economic Review*, 94(2):377–381.

Goldhaber, D., J. Kane, T., McEachin, A., Morton, E., Patterson, T., and O. Staiger, D. (2022). The Consequences of Remote and Hybrid Instruction During the Pandemic. Technical report, Center for Education Policy Research, Harvard University, Cambridge, MA.

Goldin, C. (2001). The Human-Capital Century and American Leadership: Virtues of the Past. *The Journal of Economic History*, 61(2):263–292.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, 80(2):223–255.

Grytten, J., Skau, I., and Sørensen, R. J. (2014). Educated mothers, healthy infants. The impact of a school reform on the birth weight of Norwegian infants 1967–2005. *Social Science & Medicine*, 105:84–92.

Gust, S., Hanushek, E. A., and Woessmann, L. (2022). Global Universal Basic Skills: Current Deficits and Implications for World Development. Technical report, Research on Improving Systems of Education (RISE).

Hall, C. (2012). The Effects of Reducing Tracking in Upper Secondary School: Evidence from a Large-Scale Pilot Scheme. *The Journal of Human Resources*, 47(1):237–269. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System].

Hanushek, E. A., Schwerdt, G., Woessmann, L., and Zhang, L. (2017). General Education, Vocational Education, and Labor-Market Outcomes over the Lifecycle. *Journal of Human Resources*, 52(1):48–87.

Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences*, 104(33):13250–13255.

Hoynes, H., Page, M., and Stevens, A. H. (2011). Can targeted transfers improve birth outcomes? *Journal of Public Economics*, 95(7-8):813–827.

Iriberri, N. and Rey-Biel, P. (2017). Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision. *Journal of Economic Behavior & Organization*, 135:99–111.

Iriberri, N. and Rey-Biel, P. (2019). Competitive Pressure Widens the Gender Gap in Performance: Evidence from a Two-stage Competition in Mathematics. *The Economic Journal*, 129(620):1863–1893.

Iriberri, N. and Rey-Biel, P. (2021). Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review*, 131:103603.

Jacob, B. (2017). When evidence is not enough: Findings from a randomized evaluation of Evidence-Based Literacy Instruction (EBLI). *Labour Economics*, 45:5–16.

Jacob, B. A. and Lefgren, L. (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *Review of Economics and Statistics*, 86(1):226–244.

Johnson, H., McNally, S., Rolfe, H., Ruiz-Valenzuela, J., Savage, R., Vousden, J., and Wood, C. (2019). Teaching assistants, computers and classroom management. *Labour Economics*, 58:21–36.

Jurajda, and Münich, D. (2011). Gender Gap in Performance under Competitive Pressure: Admissions to Czech Universities. *American Economic Review*, 101(3):514–518.

Kerwin, J. T. and Thornton, R. L. (2021). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. *The Review of Economics and Statistics*, 103(2):251–264.

Keslair, F., Maurin, E., and McNally, S. (2012). Every child matters? An evaluation of "Special Educational Needs" programmes in England. *Economics of Education Review*, 31(6):932–948.

Lai, F., Zhang, L., Bai, Y., Liu, C., Shi, Y., Chang, F., and Rozelle, S. (2016). More is not always better: evidence from a randomised experiment of computer-assisted learning in rural minority schools in Qinghai. *Journal of Development Effectiveness*, 8(4):449–472.

Lavecchia, A. M., Oreopoulos, P., and Brown, R. S. (2020). Long-Run Effects from Comprehensive Student Support: Evidence from Pathways to Education. *American Economic Review: Insights*, 2(2):209–224.

Lavy, V., Kott, A., and Rachkovski, G. (2022). Does Remedial Education in Late Childhood Pay Off After All? Long-Run Consequences for University Schooling, Labor Market Outcomes, and Intergenerational Mobility. *Journal of Labor Economics*, 40(1):239–282.

Lindeboom, M., Llena-Nozal, A., and van der Klaauw, B. (2009). Parental education and child health: Evidence from a schooling reform. *Journal of Health Economics*, 28(1):109–131.

Lindo, J. M. (2011). Parental job loss and infant health. *Journal of Health Economics*, 30(5):869–879.

Loyalka, P., Popova, A., Li, G., and Shi, Z. (2019). Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program. *American Economic Journal: Applied Economics*, 11(3):128–154.

Ludwig, D. S. and Currie, J. (2010). The association between pregnancy weight gain and birthweight: a within-family comparison. *The Lancet*, 376(9745):984–990.

Machin, S. and McNally, S. (2008). The literacy hour. *Journal of Public Economics*, 92(5-6):1441–1462.

Machin, S., McNally, S., and Silva, O. (2007). New Technology in Schools: Is There a Payoff? *The Economic Journal*, 117(522):1145–1167.

Machin, S., McNally, S., and Viarengo, M. (2018). Changing How Literacy Is Taught: Evidence on Synthetic Phonics. *American Economic Journal: Economic Policy*, 10(2):217–241.

McCrary, J. and Royer, H. (2011). The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth. *American Economic Review*, 101(1):158–195.

Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., Qiao, Y., Boswell, M., and Rozelle, S. (2014). Integrating computer-assisted learning into a regular curriculum: evidence from a randomised experiment in rural schools in Shaanxi. *Journal of Development Effectiveness*, 6(3):300–323.

Montolio, D. and Taberner, P. A. (2021). Gender differences under test pressure and their impact on academic performance: A quasi-experimental design. *Journal of Economic Behavior & Organization*, 191:1065–1090.

Morin, L.-P. (2015). Do Men and Women Respond Differently to Competition? Evidence from a Major Education Reform. *Journal of Labor Economics*, 33(2):443–491.

Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review*, 109(4):1426–1460.

Niederle, M. and Vesterlund, L. (2007). Do Women Shy Away From Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Niederle, M. and Vesterlund, L. (2010). Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives*, 24(2):129–144.

Nollenberger, N., Rodríguez-Planas, N., and Sevilla, A. (2016). The Math Gender Gap: The Role of Culture. *American Economic Review*, 106(5):257–261.

Noonan, K., Reichman, N. E., Corman, H., and Dave, D. (2007). Prenatal drug use and the production of infant health. *Health Economics*, 16(4):361–384.

Onda, M. and Seyler, E. (2020). English learners reclassification and academic achievement: Evidence from Minnesota. *Economics of Education Review*, 79:102043.

Oosterbeek, H. and Webbink, D. (2007). Wage effects of an extra year of basic vocational education. *Economics of Education Review*, 26(4):408–419.

Oreopoulos, P., Page, M. E., and Stevens, A. H. (2006). The Intergenerational Effects of Compulsory Schooling. *Journal of Labor Economics*, 24(4):729–760.

Ors, E., Palomino, F., and Peyrache, E. (2013). Performance Gender Gap: Does Competition Matter? *Journal of Labor Economics*, 31(3):443–499.

Palme, M. and Simeonova, E. (2015). Does women's education affect breast cancer risk and survival? Evidence from a population based social experiment in education. *Journal of Health Economics*, 42:115–124.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization*, 115:94–110.

Perin, J., Mulick, A., Yeung, D., Villavicencio, F., Lopez, G., Strong, K. L., Prieto-Merino, D., Cousens, S., Black, R. E., and Liu, L. (2022). Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet Child & Adolescent Health*, 6(2):106–115.

Rello, L., Baeza-Yates, R., Ali, A., Bigham, J. P., and Serra, M. (2020). Predicting risk of dyslexia with an online gamified test. *PLOS ONE*, 15(12):e0241687.

Rello, L., Baeza-Yates, R., and Llisterri, J. (2017a). A resource of errors written in Spanish by people with dyslexia and its linguistic, phonetic and visual analysis. *Language Resources and Evaluation*, 51(2):379–408.

Rello, L., Bayarri, C., Otal, Y., and Pielot, M. (2014). A computer-based method to improve the spelling of children with dyslexia. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility - ASSETS '14*, pages 153–160, Rochester, New York, USA. ACM Press.

Rello, L., Macias, A., Herrera, M., De Ros, C., Romero, E., and Bigham, J. P. (2017b). DytectiveU: A Game to Train the Difficulties and the Strengths of Children with Dyslexia. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 319–320, Baltimore Maryland USA. ACM.

Robalino, J. D. and Macy, M. (2018). Peer effects on adolescent smoking: Are popular teens more influential? *PLOS ONE*, 13(7):e0189360.

Roschelle, J., Feng, M., Murphy, R. F., and Mason, C. A. (2016). Online Mathematics Homework Increases Student Achievement. *AERA Open*, 2(4):233285841667396.

Rosenzweig, M. R. and Schultz, T. P. (1989). Schooling, Information and Nonmarket Productivity: Contraceptive Use and Its Effectiveness. *International Economic Review*, 30(2):457.

Rouse, C. E. and Krueger, A. B. (2004). Putting computerized instruction to the test: a randomized evaluation of a "scientifically based" reading program. *Economics of Education Review*, 23(4):323–338.

Sacerdote, B. (2011). Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far? In *Handbook of the Economics of Education*, volume 3, pages 249–277. Elsevier.

Saygin, P. O. and Atwater, A. (2021). Gender differences in leaving questions blank on high-stakes standardized tests. *Economics of Education Review*, 84:102162.

Schleicher, A. (2019). PISA 2018: Insights and Interpretations. Technical report, OCDE Publishing.

Servicio de Estudios Estadísticos (1994). *Estadística de la enseñanza en España 1991/92: niveles de Preescolar, General Básica y EE. Medias*. Centro de Publicaciones, Ministerio de Educación y Ciencia, Madrid. OCLC: 1348840540.

Sievertsen, H. H., Gino, F., and Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences*, 113(10):2621–2624.

Silliman, M. and Virtanen, H. (2022). Labor Market Returns to Vocational Secondary Education. *American Economic Journal: Applied Economics*, 14(1):197–224.

Slavin, R. E., Lake, C., Davis, S., and Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, 6(1):1–26.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Thomas, D., Strauss, J., and Henriques, M.-H. (1991). How Does Mother's Education Affect Child Height? *The Journal of Human Resources*, 26(2):183.

UNICEF-WHO (2019). UNICEF-WHO Low birthweight estimates: Levels and trends 2000–2015. Technical report, United Nations Children's Fund (UNICEF), World Health Organization (WHO), Geneva.

Vignoles, A. (2016). What is the economic value of literacy and numeracy? *IZA World of Labor*.

Wahistrom, K. (2002). Changing Times: Findings From the First Longitudinal Study of Later High School Start Times. *NASSP Bulletin*, 86(633):3–21.

Wijekumar, K., Meyer, B. J. F., Lei, P.-W., Lin, Y.-C., Johnson, L. A., Spielvogel, J. A., Shurmatz, K. M., Ray, M., and Cook, M. (2014). Multisite Randomized Controlled Trial Examining Intelligent Tutoring of Structure Strategy for Fifth-Grade Readers. *Journal of Research on Educational Effectiveness*, 7(4):331–357.

Wijekumar, K. K., Meyer, B. J. F., and Lei, P. (2012). Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve non-fiction reading comprehension. *Educational Technology Research and Development*, 60(6):987–1013.

Woessmann, L. (2016). The Importance of School Systems: Evidence from International Differences in Student Achievement. *Journal of Economic Perspectives*, 30(3):3–32.

Özek, U. (2021). The effects of middle school remediation on postsecondary success: Regression discontinuity evidence from Florida. *Journal of Public Economics*, 203:104518.