The economics of the school curriculum

Citation for published version (APA):

Diris, R. E. M. (2012). *The economics of the school curriculum*. [Doctoral Thesis, Maastricht University]. Datawyse / Universitaire Pers Maastricht. <https://doi.org/10.26481/dis.20121205rd>

Document status and date:

Published: 01/01/2012

DOI:

[10.26481/dis.20121205rd](https://doi.org/10.26481/dis.20121205rd)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

The Economics of the School Curriculum

©2012 Ron Diris, Maastricht

Cover: Answers Marked on Test. Copyright held by Debbie Parker — flickr.com

ISBN: 978 94 6159 195 1

Printed by Datawyse bv | Universitaire Pers Maastricht

The Economics of the School Curriculum

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. dr. L.L.G. Soete
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op woensdag 5 december 2012 om 14.00 uur

door

Ron Emiel Marie Diris

Promotor:

Prof. dr. L. Borghans

Beoordelingscommissie:

Prof. dr. T. Dohmen (voorzitter)

Prof. dr. A. De Grip

Prof. dr. H. Oosterbeek (Universiteit van Amsterdam)

I thank the Maastricht Research School of Economics of Technology and Organizations (METEOR) for financial support.

To my parents, my brother and Iryna

Acknowledgements

I could never have written this thesis without both the academic and moral support of some very great people. Lex, you were an excellent supervisor. You gave me a lot of freedom to consider the subjects that I found most interesting, and gave me the opportunity to work out my ideas independently. You taught me how any economic problem can be modeled in a way that captures its essence, among many other things. Your comments were always valuable and to the point. I also appreciate that you took the time to consider my work during late evenings or weekends in periods when your schedule was at its busiest.

Next, I would like to thank the members of my assessment committee, Thomas Dohmen, Andries de Grip and Hessel Oosterbeek. This thesis has also benefited from valuable feedback and comments from a very large range of people, which, aside from the committee members, include Bart Golsteyn, Tim Kautz and Trudie Schils.

I would also like to express my thanks to Professor Jim Heckman for granting me the opportunity to stay at the University of Chicago as a visiting scholar. I would like to thank him and those in and around his team for their hospitality and valuable cooperations. I learned so much during my stay in Chicago, and had a great time there.

Throughout my PhD, I was lucky enough to work with very nice people in my own office, and the one across the hall. Norbert, I miss those intermezzo's of bad 80s and 90s music during tough working days. Ehsan, thanks for your patience in answering countless questions about LaTeX. Both of you have become great friends. Lennart, I hope that your new office mates share your preference for switching off the light during work and I am sorry I did not. Caicai, I will miss making jokes about your tea cup. Bart, Tyas and Ulf, I always enjoyed dropping by your office and talking about serious and less serious issues. Sorry for always stealing tea.

Of course, I cannot forget the crazy “jungens” from the Accounting department, Jonas, Patrick and Thomas. I do not know who is in charge of hiring PhD-students there, but they should keep this strategy going! I had great fun during lunches, talks in the office and cultural trips to Germany. I would also like to thank many other nice colleagues at Maastricht University, including Nadine Funcke, Bas Marban, Mona Offermanns, Cyril Rutten, Paul Smeets and Simone Wies, for co-creating such a pleasant working atmosphere. Silvana, Fleur and Sylvia, thanks for your help in dealing with numerous administrative tasks, and for the candies at the Secretariat that often provided much needed sugar boosts. I also thank my old high school classmates, Bart, Benno, Dennis, Joost, Rob and Rob, for their continuing friendship through all these years.

Ik wil graag afsluiten met de mensen die het allerdichtst bij me staan. Allereerst mijn ouders, die me altijd de vrijheid hebben gegeven om te doen wat ik leuk vond, en altijd voor me klaarstaan wanneer dat nodig is. Bart, het is een hele eer om deze thesis straks naast die van jou te zetten in de Diris-boekenkast. Bedankt voor je continue steun. Of het nu ging om econometrische problemen, administratieve zaken, sollicitatiebrieven, of het doorlezen van mijn papers, nooit was je te beroerd om te helpen en zeer uitgebreide feedback te geven. Ik ben je daar heel erg dankbaar voor. Iryna, ik kan je nooit genoeg bedanken in een paar regels tekst. Met je positieve energie, je gevoel voor humor, je steun en je liefde maak je voor mij leuke tijden nog leuker en moeilijke tijden makkelijk. *Diakuii!*

Leuven, October 2012

Contents

List of Figures	xi
List of Tables	xiii
1 Introduction	1
2 The Theory of the School Curriculum	7
2.1 Changes in the school starting age	8
2.1.1 The Skill Technology Function	9
2.1.2 Costs and benefits	12
2.2 Allocating Instruction Time	17
2.2.1 Long-run effects	17
2.2.2 Spillover effects	19
2.3 Tracking	22
2.4 Grade Retention	24
2.5 Figures	27
3 An Economic Analysis of the Optimal School Starting Age	31
3.1 Introduction	31
3.2 Literature	32
3.2.1 Economic literature	32
3.2.2 Psychological literature	34
3.3 Data	35
3.3.1 Descriptive Statistics	37
3.4 Methodology	38
3.4.1 Within-country analysis	38

CONTENTS

3.4.1.1	Correction method	39
3.4.2	Between-country analysis	41
3.4.3	Cost-benefit analysis	42
3.5	Results	43
3.5.1	Within-country analysis	43
3.5.2	Between-country analysis	44
3.5.3	Cost-benefit analysis	44
3.6	Robustness	46
3.6.1	Peer effects	46
3.6.2	Tracking	48
3.6.3	Inequality	48
3.6.4	Month of birth	49
3.7	Conclusion	49
3.8	Figures and tables	51
4	Allocating Instruction Time: How Language Instruction Can Affect Multiple Skills	61
4.1	Introduction	61
4.2	Dutch educational system	63
4.3	Literature	64
4.3.1	Instruction time	64
4.3.2	IQ literature	66
4.4	Data	67
4.4.1	The PRIMA-dataset	67
4.4.2	Descriptive statistics	69
4.5	Methodology	70
4.6	Results	73
4.6.1	Language	73
4.6.2	IQ	74
4.6.3	Effects on other subjects	75
4.6.4	Magnitudes	76
4.6.5	Effect of week of testing	77
4.6.6	Heterogeneity	77
4.6.7	Panel data	79

4.7	Robustness	81
4.7.1	Other changes over time	81
4.7.1.1	School-level changes	81
4.7.1.2	Class-level changes	82
4.7.1.3	Cross-sectional changes	83
4.7.2	Trends	84
4.7.3	Volatility after change	86
4.7.4	Non-cognitive skills	86
4.8	Conclusion	87
4.9	Figures and tables	89
5	The Labor Market Effects of Tracking: A Regression Discontinuity Approach	103
5.1	Introduction	103
5.2	Dutch educational system	105
5.2.1	Track assignment	105
5.2.2	Post-secondary education	106
5.3	Literature	106
5.4	Data	108
5.4.1	The sample	108
5.4.2	Graphical analysis	109
5.5	Methodology	110
5.5.1	The regression discontinuity design	110
5.5.2	Predictive power	111
5.5.3	The IV-model	112
5.5.4	The sample multiplier	113
5.5.5	Bandwidth	114
5.6	Results	115
5.6.1	Educational attainment	116
5.6.2	Earnings	116
5.7	Robustness	117
5.7.1	Sensitivity to bandwidth	117
5.7.2	Control function	118
5.7.3	Cito vs. Entrance Test	119

CONTENTS

5.7.4	Peer effects	120
5.8	Conclusion	121
5.9	Figures and tables	123
6	The Impact of Early Tracking on Later-life Outcomes: An Instrumental Variable Approach	143
6.1	Introduction	143
6.2	Data	145
6.3	Methodology	146
6.3.1	Defining early tracking	146
6.3.2	Endogeneity	147
6.3.3	The IV model	147
6.4	Results	149
6.4.1	Educational attainment	150
6.4.2	Labor market outcomes	152
6.4.3	Magnitudes	153
6.5	Robustness	154
6.6	Conclusion	156
6.7	Figures and tables	158
7	The Impact of Grade Retention: An IV approach to Estimating Heterogeneous Treatment Effects	169
7.1	Introduction	169
7.2	Literature	170
7.3	Data	172
7.4	Methodology	173
7.4.1	Comparison group	173
7.4.2	Selection	174
7.4.3	Heterogeneity	175
7.5	Results	177
7.5.1	Overall results	177
7.5.2	Heterogeneity by percentile	178
7.5.3	Heterogeneity by demographics	179
7.5.4	Effects on confidence	179

7.5.5	Early versus late retention	180
7.6	Robustness	181
7.6.1	Curriculum bias	181
7.6.1.1	Effects by difficulty level	181
7.6.1.2	Problem solving	181
7.6.2	Exogeneity of month of birth	182
7.7	Conclusion	183
7.8	Figures and tables	184
8	Summary and conclusion	197
8.1	Summary	197
8.2	Policy implications	200
8.3	Reservations and future research	202
	References	205
9	Nederlandse samenvatting	209
9.1	De economie van het school curriculum	209
9.2	Beleidsimplicaties	214
9.3	Reservaties en toekomstig onderzoek	215
	Curriculum Vitae	217

List of Figures

2.1	The Skill Technology Function	27
2.2	Effect of track on outcomes: efficient selection	28
2.3	Effect of track on outcomes: inefficient selection	29
3.1	Birth cohort vs. grade: no correction	51
3.2	Birth cohort vs. grade: with correction	51
3.3	Mean reading scores	52
3.4	Cost-benefit analysis	53
4.1	Mean levels of language instruction time	89
4.2	Mean levels of math instruction time	89
4.3	Mean language scores by panel	90
4.4	Mean IQ scores by panel	91
5.1	Track assignment by test score	123
5.2	Density of forcing variable	124
5.3	Fraction of students assigned to T1 (<i>vmbo</i>) by test score	124
5.5	Fraction of students assigned to T1a (<i>lbo</i>) by test score	125
5.4	Fraction of students assigned to T3 (<i>vwo</i>) by test score	125
5.6	Mean earnings by test score: 2007-2004	126
5.7	Mean years of schooling by test score	127
5.8	Linear prediction of earnings pattern	127
5.9	Mean squared errors (1977)	128
5.10	Probability of treatment: Cito	129
5.11	Probability of treatment: Entrance Test	129

LIST OF FIGURES

6.1 Ultimate track assignment per advice group 158

6.2 Ultimate educational attainment per advice group 159

6.3 Distribution of early tracking instrument: T1/T2 160

6.4 Distribution of early tracking instrument: T2/T3 160

6.5 Value of instrument by mean test score in municipality 161

7.1 Frequency of retention by month of birth 184

7.2 Frequency of Grade Retention by month of birth 184

7.3 Frequency of Late Starts by month of birth 185

7.4 Effect of Grade Retention and Late Starts on science scores by percentile . . . 185

7.5 Effect of Grade Retention and Late Starts on math scores by percentile 186

7.6 Effect of Grade Retention and Late Starts on reading scores by percentile . . . 186

List of Tables

2.1	Difference in Cost-benefit structure	13
3.1	The Birth Month Effect (BME) for 4th grade reading	54
3.2	Overall coefficients of birth month on score	55
3.3	Relative magnitude of Birth Month Effect	56
3.4	Effect of school starting age on Birth Month Effect	57
3.5	Sensitivity of cost-benefit analysis	58
3.6	Effect of month of birth on confidence	59
3.7	Effect of school starting age on inequality	59
4.1	Summary statistics	92
4.2	Value of dummy instruments	93
4.3	Effect of ability indicators on instruction time for language	93
4.4	First stage results (F-values)	94
4.5	Effect of instruction time (IT) for language on test scores in language	94
4.6	Effect of instruction time (IT) for language on IQ test scores	95
4.7	Effect of instruction time (IT) for language on math and environmental studies	96
4.8	Effect of instruction time on ranking: language scores (IV)	97
4.9	Effect of instruction time on ranking: IQ (IV)	98
4.10	Effect of instruction time on ranking: math and environmental studies (IV)	99
4.11	Mean differences of school characteristics	100
4.12	Sensitivity to extracting trends (grade 6)	101
5.1	Summary statistics	130
5.2	Thresholds along test score distribution	131
5.3	Strength of instrumentation: F-values	131

LIST OF TABLES

5.4	OLS estimates of track assignment on earnings and years of schooling	132
5.5	IV estimates of Track 1 assignment on earnings and years of schooling	133
5.6	IV estimates of Track 3 assignment on earnings and years of schooling	134
5.7	IV estimates of Track 1a assignment on earnings	135
5.8	Sensitivity of results to bandwidth	136
5.9	IV estimates of track assignment when dropping tracks	137
5.10	IV estimates of track assignment when using track-specific trends	138
5.11	IV estimates of track assignment when also instrumenting control function . . .	139
5.12	IV estimates of track 1 assignment when adding controls	140
5.13	Cito score versus Entrance Test score as forcing variable	141
6.1	Comparison of variables for different age of tracking	162
6.2	First stage results (F-values)	163
6.3	OLS estimates of early tracking	164
6.4	IV estimates of early tracking: educational attainment	165
6.5	IV estimates of early tracking: labor market effects	166
6.6	Robustness to alternative specifications	167
7.1	Degree of retention by country	187
7.2	First stage results: F-values	187
7.3	Effect of Grade Retention and Late Starts: OLS	188
7.4	Effect of Grade Retention and Late Starts: OLS with controls	188
7.5	Effect of Grade Retention and Late Starts: IV	189
7.6	Effect of Grade Retention and Late Starts: IV with controls	189
7.7	IV estimates of Grade Retention and Late Starts: per decile	190
7.8	IV estimates of Grade Retention and Late Start: per percentile	191
7.9	IV estimates of Grade Retention and Late Start: Overall Score (percentiles) . .	192
7.10	IV estimates of Grade Retention and Late Starts by demographic group	193
7.11	IV estimates of Grade Retention and Late Start on confidence-related variables	194
7.12	IV estimates of Early (ISCED 1) versus Late (ISCED 2) Grade Retention . . .	195
7.13	IV estimates of Grade Retention and Late Start by difficulty level of questions .	195
7.14	IV estimates of Grade Retention and Late Start: problem solving (percentiles) .	196

1

Introduction

There is a growing understanding that improving the formation of human capital can alleviate many economic and social challenges of today. Both cognitive and non-cognitive skills substantially contribute to success in later life as measured by educational attainment, crime, earnings, health and life satisfaction.¹ Moreover, the return on skills is increasing over time. Over the last 25 years, the real wages of the high-skilled are increasing, while those of low-skilled workers are stagnating or even falling. Increasing the share of highly educated individuals will also raise the productivity of the unskilled and makes skilled workers less scarce (Heckman, 2000). As such, improving skills has the potential to address inequality as well.

The development of socially productive skills² occurs throughout life and in many different environments. Learning is not the sole domain of formal academic institutions but also occurs in the family environment and among peers, and continues in the work place. Nonetheless, schools are a core producer of skills. The importance of education is large and present across many domains. Previous research has identified causal impacts of schooling on earnings, health and crime.³ Moreover, education not only produces academic skills, but enhances soft skills as well. Heckman et al. (2006) show that more years of schooling lead to increases in self-esteem and locus of control, which measures the degree to which people perceive important matters in life as under their own control and is a strong predictor of success in life in multiple domains. Hence, schooling produces a wide range of capabilities and produces many favorable outcomes. However, the exact processes within the educational system that lead to these favorable impacts

¹For an overview, see Almlund et al. (2011).

²Referring to skills that are relevant for important later-life outcomes, such as educational attainment and earnings.

³See e.g. Angrist and Krueger (1991); Lochner (2011); Lochner and Moretti (2004).

1. INTRODUCTION

still remain somewhat of a black box. Getting a better insight into these mechanisms can help in designing educational systems that will be more optimal in developing socially productive skills. As such, the return on schooling can be further increased, thereby furthering economic progress and reducing multiple social challenges that societies face today.

This thesis examines and analyzes how certain aspects of the educational system influence skill development and important later-life outcomes, such as educational attainment and earnings. The school curriculum can be designed in very different ways, for example with respect to the school starting age it employs or the age at which students are put in different tracks. The exact design has strong implications for the development of skills for students that pass through that system. This will be reflected in their scores on achievement tests when they are still in education, as well as in outcomes in later life. If we can design the curriculum in a way that increases the efficiency of skill development, these later-life outcomes will increase as well. We still know little about how to optimize the learning process. Curriculum designs are based on tradition, common sense or developmental theory. They are seldom based on strong empirical evidence. This is mainly because strong evidence is lacking in most areas. There is no clear understanding on what ages are best for learning which skills, which order of teaching is most efficient, in what peer groups students learn best, or how we should deal with students that lag behind the rest of the class.

Although economic research in this field is expanding, the analysis of questions like these has largely been the domain of educationalists. Economic theory can enrich the discussion on these topics. For example, discussion on the optimal school starting age has largely neglected the human capital argument that postponing formal education implies foregone time and earnings. We show that incorporating this argument leads to completely different conclusions regarding the effects of employing different starting ages. Moreover, the complex and dynamic nature of skill development inhibits multiple economic concepts.

Skill development is a process characterized by strong interactions. Skills acquired in the past determine how fast someone learns new skills. Interactions can be present over a wide range of subjects. For example, it is obvious that learning the alphabet is a prerequisite for learning to write, but well-developed reading skills can also enhance the productivity of skill development in subjects such as mathematics and geography. Skills developed in the past make future learning more productive and increase the return on future investments in the child. These concepts are known as self-productivity and complementarity. The dynamic nature of skill acquisition

underlines the importance of measuring effects at different points in time. Initial effects can develop in very different ways, depending on the degree of self-productivity and complementarity over time. This also implies that remedial policies at later ages often fail, or are very costly, because they are building on a less well-developed set of skills. Investments have higher payoffs when they are located early in life, because their returns contribute to skill development for a larger part of the life cycle. This indicates the importance of early childhood and is in line with research that shows the difficulties of remediating skill gaps at later ages.¹

Taking into account the dynamic nature of skill acquisition is one important aspect of estimating the impacts of the school curriculum. The presence of endogeneity constitutes another. One cannot simply compare variation across countries, because the design of these systems is not arbitrarily determined. Moreover, differences within countries in how students are treated are hardly ever exogenous either. For example, schools with more low-ability students tend to spend more time on core subjects such as language and mathematics. When one would simply look at the correlation between instruction time in these subjects and performance, results will be biased downwards, since low-ability students are associated with higher levels of instruction time. Empirical estimation should correct for such confounding factors. Previous research in education has often relied on cross-sectional variation in the explanatory variable. Endogeneity is generally addressed by matching on socioeconomic status, parental education and a range of other observable characteristics. This approach does not correct for differences in unobservables. Previous research shows the importance of correcting for endogeneity when estimating effects of education and training.² For policy purposes, it is crucial that we establish causal links between changes in aspects of the educational system and relevant outcome variables. We develop specific methods to deal with endogeneity in all of the empirical chapters in this thesis. These methods rely on instrumental variable estimation or regression discontinuity designs. Instruments exploit exogenous variation induced by month of birth, policy changes over time, or variation in school characteristics across regions. Results show that correcting for endogeneity matters significantly. It leads to significantly different estimates and often very different qualitative implications as well. This dissertation also focuses on heterogeneity of treatment effects. We analyze how effects differ across demographic groups and apply a quantile regression model that estimates treatment effects across the whole distribution.

¹See e.g. Almlund et al. (2011); Heckman (2000).

²See e.g. Abadie et al. (2002); Angrist and Pischke (1999); Leuven et al. (2007); Leuven and Oosterbeek (2004).

1. INTRODUCTION

This thesis examines four specific aspects of the school curriculum. These are the starting age of formal education, the allocation of instruction time over the typical schooling week, the execution of educational tracking, and the impact of grade retention. For each of these topics, we develop a theoretical framework specifying the nature of the problem. We apply this theory to an empirical approach. All theoretical discussions build on *The Technology of Skill Formation* by Cunha and Heckman (2007). We concentrate these frameworks in Chapter 2 of this thesis. The five following chapters focus on the empirical analysis of the problem at hand.

Chapter 3 analyzes the optimal timing of school entry. We develop a framework where we compare the increase in learning productivity from postponing formal education with the economic costs of foregone time that waiting involves. How learning productivity relates to age is modeled by the Skill Technology Function. We use a within-country analysis, exploiting variation in month of birth, to estimate the slope of this function. A between-country analysis is employed to estimate its curvature. With these results, we can deduce differences in skill development from differences in starting ages and compare these with the economic costs of foregone time. We find that earlier starting ages than what is currently the norm in developed nations are more optimal for skill development, and produce substantial economic returns. This occurs while younger entrants consistently lag behind their older peers as measured by achievement tests. However, the fact that these students are younger compensates for that skill deficit, ultimately bringing them to a higher skill level at the same precise age. Additional analysis indicates that these findings are not driven by peer effects, tracking, or possible endogeneity of month of birth.

In Chapter 4, we analyze the short- and long-term consequences of changes in the quantity of language instruction in primary education for skill development. We analyze how initial effects develop over time and whether language instruction has positive spillovers towards other school subjects. We use exogenous variation in the quantity of language instruction over time, caused by policy changes in the Netherlands. Results show that more language instruction has short-term impacts on language skills. However, these gains fade away quickly, while instruction in language does have sustained effects on mathematics and environmental studies. Additional analysis shows that these results are not driven by school-level changes or general trends in scores over time.

Chapters 5 and 6 examine the implications of tracking for educational attainment and earnings. We examine two specific aspects of school tracking. In Chapter 5, we analyze what the effect is of track assignment for those students who are around the required ability level for a

specific track. Thereby, we exploit the existence of a discontinuity in treatment probability at certain scores of selection tests in the Netherlands. We conclude that students around the margin achieve higher levels of educational attainment when attending the higher track. This translates into effective gains in earnings in later life, which are around 10% per extra year of education. Chapter 6 analyzes the implications of the exact age of tracking, and how this effect depends on the ability level of the student. Thereby, we employ an instrument that exploits regional variation in the relative supply of early tracking schools. This proves to be a valid and powerful instrument for early tracking. We find that early tracking reduces the share of students that obtains a higher education degree, and this translates into lower earnings for some subgroups as well. This effect weakens the higher the student ranks in the distribution, but only reverses for a small subgroup of students of very high ability.

Chapter 7 analyzes the consequences of grade retention for school achievement. We employ an Instrumental Variable approach, using the fraction of students within a certain month of birth that have repeated a grade as an instrument for retention. We combine this with a quantile regression model that ensures estimation of specific treatment effects across the ability distribution. Our data consist of international assessment tests for a large range of countries. Retention has an overall negative impact on student achievement, but can be beneficial for small subgroups near the bottom of the distribution, provided it is executed at a very early stage. Moreover, we find that students that are being retained during formal education (i.e. after kindergarten) score relatively worst on the easiest questions, indicating they are lacking especially basic skills.

2

The Theory of the School Curriculum¹

All chapters in this thesis deal with the accumulation of skills over the life cycle. We examine how specific aspects of the educational system can influence this dynamic process, and what this implies for later-life outcomes such as earnings and educational attainment. Throughout this thesis, the Technology of Skill Formation, developed by Cunha and Heckman (2007), serves as our major framework. It specifies the development of skills over the life-cycle as a function of the skill set already present (θ), parental characteristics (h) and investments (I), in time period t :

$$\theta_{t+1} = f_t(\theta_t, h, I_t)$$

The framework recognizes that skills beget skills; a process which is labeled as self-productivity. This interdependence over time can imply that early skill deficits strongly expand over time. Such divergence can be further enhanced by the presence of static and dynamic complementarity. *Static complementarity* implies that the return on investments is higher when the skill set at that point in time is higher (complementarity between skills and investments). *Dynamic complementarity* arises because early investments make later investments more productive. This is because early investments improve skills, which increases the productivity of later investments as well (complementarity between investments over time). Strong dynamic complementarity also implies that late investment should follow early investment if the return on the initial investment is to be harvested. When complementarity is low, compensation in later periods is possible and skill levels may converge again. Hence, any estimated initial effect of any investment in skill

¹This chapter is based on Borghans and Diris (2010a,b); Borghans, Diris, Smits, and De Vries (2011, 2012); Diris (2012).

2. THE THEORY OF THE SCHOOL CURRICULUM

can develop in very different ways, which makes it crucial to estimate both short- and long-run effects. Complementarity and self-productivity are generally neglected in public policy. There is a strong belief in remedial investments as effective practices to address skill deficiencies. However, skill accumulation is not an additive process. Remediating skill gaps might not be feasible, or highly costly, because those skills developed early in life can be absolutely crucial for the whole process of skill formation.

Hence, the return on investments, such as teaching a child how to read, depends on the timing of that investment. Cunha and Heckman (2007) define critical and sensitive periods. *Critical periods* exist when a certain investment has no effect at all in any other period. A *sensitive period* occurs when the return on investment in that period is higher than in any other period. Through self-productivity and complementarity, the timing of early investments is an important component of future skill development and the productivity of any future investment. Whether we are assessing the effect of school starting ages, the quantity of language instruction in primary education, educational tracking or grade retention, we need to look at their impacts at different points in time if we want to say something meaningful regarding their ultimate effect on the total skill set and, correspondingly, various domains of social progress.

Cunha and Heckman (2007)'s theory has become the leading framework for the analysis of skill formation. It captures its dynamic nature as well as the key determinants of the productivity of investments. Moreover, it can explain several observations from empirical evidence, such as the low return on remedial investments at later ages.¹ The Technology of Skill Formation also provides a good starting point for our empirical analyses, which relies on variation in either endowments (θ) or investment (I). It specifies that, when we analyze the effect of a certain investment, we have to make sure that we compare individuals that do not differ in the skill set before investment, parental characteristics, or any other investments.

2.1 Changes in the school starting age

Chapter 3 analyzes the effect of different school starting ages on skill formation, and assesses the economic consequences of this impact. We employ the Technology of Skill Formation to specify that differences in endowments developed before school entry lead to differences in skill development.

¹See e.g. Almlund et al. (2011); Heckman (2000).

2.1.1 The Skill Technology Function

From an economic perspective, the optimal school starting age¹ depends on the tradeoff between foregone investment and foregone endowments. The former relates to human capital theory which states that postponement of investment lowers its value.² The latter refers to the argument that older children have built up more pre-school endowments and are therefore more ‘ready’ to learn a particular skill (Strøm, 2004). We split up our main analysis into two parts to estimate this particular tradeoff. First, we specify the technology of skill formation. Second, given this technology, we calculate the consequences in terms of lifetime earnings for a range of different starting ages.

The process of skill accumulation is defined using Cunha and Heckman (2007)’s Technology of Skill Formation. In this chapter we analyze the effect of differences in starting ages, which are determined by the birth month of the individual. Assuming month of birth to be exogenous, differences in parental characteristics and investment should have no effect when comparing the outcomes of individuals with different starting ages. The efficiency of the learning process depends on incumbent skills θ , which grow over the life-cycle and are hence a function of the age of the child. We introduce a Skill Technology Function (STF), which models learning productivity over age. This STF is portrayed in figure 2.1.

Learning productivity refers to how fast you learn a particular skill. Learning how to read is not effective right after birth, so the curve remains flat in the beginning. Productivity increases with age, as children become more and more ‘ready’ to start learning as they build up endowments. At some point, there is a peak³ where learning is most effective, after which productivity starts to decrease. Hence, differences in outcomes are determined by differences in skill endowments at the beginning of the period. Initially, those consist of endowments built up before anyone goes to school, which are presumably higher for older individuals, since they had more time to accumulate them (Strøm, 2004).⁴

The question remains: what does this imply for optimal school starting ages? Starting later gives higher productivity, but waiting with learning also implies lost time. Higher learning

¹Referring to the start of the formal teaching of reading and writing.

²See e.g. Becker (1964); Ben-Porath (1967).

³This is what Cunha and Heckman (2007) label as the sensitive period.

⁴Figure 2.1 is not a formal representation of any theory or belief about child development. It is our assumed shape of this particular function, based on psychological research that states that learning language before age 4 is non-productive (see e.g. Blackmore and Frith (2005); Meadows (1993)), and findings from, for example, Bedard and Dhuey (2006) that the differences in test scores by age are larger for countries that start early (which implies concavity, at least for the part of the function where currently employed starting ages are located).

2. THE THEORY OF THE SCHOOL CURRICULUM

productivity for older starters needs to compensate for what younger starters already learned when the former group was not in school yet. The optimal starting age cannot lie at the flat part in the beginning of the curve, because nothing will be learned in those early years, while learning productivity thereafter will be harmed.¹ The optimal age can also not lie after the peak, since this implies decreasing learning productivity and more lost time. In between these points, the choice is not immediately obvious and will depend on the tradeoff of the increasing learning speed for older starters, approximated by the slope of the STF, versus the increasing head start for younger starters. We estimate the slope of the STF by using month of birth as an instrument for school starting ages. Children who are born within one year after the official cutoff date all start formal education at the same point in time, but their birth month determines at which exact age they start with learning to read and to write. The effect that month of birth (and, correspondingly, the school starting age) has on skills is labeled as the Birth Month Effect (BME). The BME constitutes the slope of the STF. By executing this analysis for a multitude of countries, we can also approximate the curvature of the STF.

The above specifies how children differ in learning productivity when they enter school and learn literacy skills. It is crucial how these initial differences develop over time. The Technology of Skill Formation specifies that skills are interdependent, both over time and across different skill types. This means that reading skills in a certain period depend on reading skills developed at the very start of schooling, but also that skills in mathematics or science are dependent on that initial level of reading. The BME might be very different for performance in later grades, or in other topics than reading. It can increase, since new skills built on existing skills which reinforces any initial impact. It can decrease, since the relative difference in age can become smaller for later periods.²

We formalize the development of the Birth Month Effect over time by specifying that how fast you learn something depends on a general learning productivity (β) and the degree of complementarity with already developed skills (α).³

$$\theta_{i,t+1} = \beta_{i,t} + \alpha_{i,t}\theta_{i,t}$$

¹Assuming that someone who has not entered school yet in those early years builds up more endowments in that period.

²And other subjects like math and science could have very different Skill Technology Functions altogether.

³For simplicity, we assume a linear model here. Changing this specification will not affect the conclusions we draw.

We assume that an older individual is more proficient in learning and is better able to use these already developed skills in learning something new. Hence, the α 's and β 's of the older individual 2 are larger than those of the younger individual 1. Whether the BME diverges, converges or remains stable depends on whether the difference in the learning of new material is larger than, smaller than or equal to the original BME. Under stability,¹ we have:²

$$\begin{aligned}(\beta_{2,t} + \alpha_{2,t}\theta_{2,t}) - (\beta_{1,t} + \alpha_{1,t}\theta_{1,t}) &= \theta_{2,t} - \theta_{1,t} \\(\beta_{2,t} - \beta_{1,t}) + (\alpha_{2,t} - \alpha_{1,t})\bar{\theta}_t &= \theta_{2,t} - \theta_{1,t}\end{aligned}$$

We assume that both the α 's and the β 's converge as these individuals grow older, while the θ 's naturally increase over time. The development of the BME depends on the relative strength of the convergence of the α 's and β 's versus the strength of the increase in θ . If we assume, for simplicity, that the β 's have fully converged, the first term disappears. Under stability, we have:³

$$\begin{aligned}\frac{(\alpha_{2,t} - \alpha_{1,t})\bar{\theta}_t}{(\alpha_{2,t-1} - \alpha_{1,t-1})\bar{\theta}_{t-1}} &= 1 \\ \frac{\alpha_{2,t-1} - \alpha_{1,t-1}}{\alpha_{2,t} - \alpha_{1,t}} &= \frac{\bar{\theta}_t}{\bar{\theta}_{t-1}} \\ \frac{\alpha_{2,t-1} - \alpha_{1,t-1}}{\alpha_{2,t} - \alpha_{1,t}} &= \frac{\bar{\alpha}_t\bar{\theta}_{t-1} + \bar{\beta}_t}{\bar{\theta}_{t-1}} \\ \frac{\alpha_{2,t-1} - \alpha_{1,t-1}}{\alpha_{2,t} - \alpha_{1,t}} &= \bar{\alpha}_t + \frac{\bar{\beta}_t}{\bar{\theta}_{t-1}}\end{aligned}$$

Over time, the ratio $\bar{\beta}_t/\bar{\theta}_{t-1}$ will get smaller as $\bar{\theta}$ increases. Hence, we can essentially say that there is stability if the convergence in the α 's and the absolute level of α balance each other out. Similarly, when convergence of the α 's is strong enough to overcompensate for the growth rate $\bar{\alpha}_t$, the Birth Month Effect will be converging. Divergence occurs when the right term outweighs the left term. When there is also full convergence of α 's, the left term collapses

¹Meaning that the Birth Month Effect does not change over time. Under convergence, the left term will be smaller than the right term. Under divergence, the left term will be larger than the right one.

²For ease of interpretation, we use $\bar{\theta}$ as an approximate average of θ^1 and θ^2 . As such, we simplify $\alpha^2\theta^2 - \alpha^1\theta^1$ to $(\alpha^2 - \alpha^1)\bar{\theta}$.

³In a similar way as above, we use $\bar{\alpha}$ and $\bar{\beta}$ as approximate averages of α^1 and α^2 , and β^1 and β^2 , respectively. They are the key elements of the average rate at which $\bar{\theta}$ grows from period t-1 to period t.

2. THE THEORY OF THE SCHOOL CURRICULUM

and the BME converges to zero.¹ To sum up, whether the original Birth Month Effect remains stable, converges or diverges over time depends on the strength of the convergence of learning productivity between older and younger individuals versus the absolute levels of these learning productivity parameters. Even when learning productivity converges, initial differences can still be sustained, because skills are complementary over time and better developed early skills help in learning new skills. This illustrates the difficulty in remediating early differences in skills and the corresponding importance of (early) childhood.²

2.1.2 Costs and benefits

We can now estimate the implications of a specific school starting age (SSA) for a person's skill level after a certain years of schooling. Starting school a year later means, all other things equal, that one also finishes school a year later in life. This corresponds to a loss of foregone earnings, or additional study time, of a whole year. If we want to analyze the total effect of a change in school starting age, we also need to consider this economic component. We estimate whether the skill deficit suggested by the shape of the STF weighs up against the time gain of being younger. If the time gain exceeds the skill deficit, the individual reaches a specific skill level at an earlier (exact) age than with a later SSA.

We assume that each individual studies up to the point where the (potential) deficit from starting school earlier is made up for.³ For example, if the deficit of starting a whole year earlier (e.g. 5 instead of 6 years old) corresponds to half an academic year, one could study for half a year to compensate that difference. In that case, one can start working half a year earlier than under the alternative (later) starting age, with the same skill level. Table 2.1 gives an overview of that hypothetical situation, where I indicates investments or costs during school and Y indicates income during working life.

¹When we drop the assumption of convergence of the β 's, the left term expands, but we cannot say if it increases or decreases. Under stability, we have $\Delta\beta_t + \Delta\alpha_t\Delta\theta_t / \Delta\beta_{t-1} + \Delta\alpha_{t-1}\Delta\theta_{t-1} = 1$. Whether the left term increases or decreases depends on the ratio of the β 's relative to the ratio in the original left term. Still, it holds that the stronger the convergence between the β 's, the more likely it is that the BME also converges. In general, we may assume that the convergence of the β 's plays a smaller role over time than that of the α 's, since the weight of the β 's decreases as θ naturally grows larger.

²The importance of early childhood has been well-documented in recent work by Cunha and Heckman (2008); Cunha et al. (2010); Heckman and Masterov (2006).

³One could also study longer and enter the labor market with a better skill set, or start working (a lot) earlier with a lower skill set. The results from the economic analysis should be seen as a conservative estimate; the gains are higher if any of the other choices is more optimal.

Table 2.1: Difference in Cost-benefit structure

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Start at age 6						I ₁	I ₂	I ₃	I ₄	I ₅	I ₆	I ₇	I ₈	I ₉	I ₁₀	I ₁₁	I ₁₂	Y ₁	Y ₂
Start at age 5					I ₁	I ₂	I ₃	I ₄	I ₅	I ₆	I ₇	I ₈	I ₉	I ₁₀	I ₁₁	I ₁₂	I _x	Y ₁	Y ₂

This involves gains from starting earlier. These are based on the gain in time of finishing earlier, with the same skill level. This *time surplus* is equal to the change in school starting age minus the additional study time needed to compensate for a potential skill deficit. The largest benefit consists of the fact that all income streams are incurred earlier. These benefits (labeled R_1) are equivalent to the effective interest rate over lifetime income. The effective interest rate depends on how much earlier the income streams are incurred, which is equal to the change in school starting age (ΔSSA) minus the time to compensate for the skill deficit (δ). Lifetime income exists of all income streams from the start of working life (T_w) until retirement (T_r). Benefits are discounted to the start of formal education. The parameter r indicates the nominal interest rate.

$$R_1 = ((1 + r)^{\Delta SSA - \delta} - 1) \sum_{t=T_w}^{T_r} \frac{Y_t}{(1 + r)^t}$$

The change in starting age, ΔSSA , always refers to a *decrease* in the starting age. We take a specific SSA as a benchmark and then estimate the returns for a range of earlier starting ages. Thus, a large ΔSSA refers to a very young school starting age. This method allows a specific country to set its own starting age as a benchmark and estimate the effects of any changes in that ‘benchmark starting age’.¹

In addition to the earlier incurred income streams, younger starters incur an extra income stream as well. Because one can start working earlier, one also has a longer working life if we assume the retirement age to be fixed. The length of this extra income stream (labeled R_2) again

¹This ‘benchmark starting age’ is also the exact age to which the income streams are being discounted to. This is irrelevant for the optimal age, but matters for the exact magnitude of our estimated profits.

2. THE THEORY OF THE SCHOOL CURRICULUM

depends on the time surplus; $\Delta SSA - d$:¹

$$R_2 = \frac{(\Delta SSA - \delta)Y_{T_r+1}}{(1+r)^{T_r}}$$

Moving to an earlier starting age also involves costs. The nature of these costs are analogous to that of the benefits. There are costs resulting from the fact that educational costs are also incurred earlier. These costs (labeled C_1) are equal to the effective interest rate over all educational costs. The effective interest rate depends on how much earlier we start with formal education; ΔSSA . Educational costs are all cost streams from the start of formal education (T_{ss}) until the very end of the educational career (T_{se})

$$C_1 = ((1+r)^{\Delta SSA} - 1) \sum_{t=T_{ss}}^{T_{se}} \frac{I_t}{(1+r)^t}$$

Extra educational costs are incurred to compensate for the academic deficit that starting earlier with formal education (presumably) causes (I_x in table 2.1). The size of this extra cost stream (labeled C_2) depends on the size of the academic deficit (δ):

$$C_2 = \frac{\delta I_{T_{se}+1}}{(1+r)^{T_{se}}}$$

Whether moving to an earlier starting age results in a positive economic profit depends on how the costs weigh up to the benefits. We label the exponent $\Delta SSA - \delta$ as γ . This γ corresponds to the time that is left after making up for the academic deficit. It refers to how much one can start earlier with earning an income; the *time surplus*. Profits become:

$$\Pi = R_1 + R_2 - C_1 - C_2$$

$$\Pi = ((1+r)^\gamma - 1) \sum_{t=T_w}^{T_r} \frac{Y_t}{(1+r)^t} + \frac{\gamma Y_{T_r+1}}{(1+r)^{T_r}} - ((1+r)^{\Delta SSA} - 1) \sum_{t=T_{ss}}^{T_{se}} \frac{I_t}{(1+r)^t} - \frac{\delta I_{T_{se}+1}}{(1+r)^{T_{se}}}$$

One can draw a couple of preliminary conclusions from this system of equations:

¹It might be counterintuitive that we are discounting the extra income stream, Y_{T_r+1} , by T_r instead of T_{r+1} . This extra income stream is incurred *before* retirement (otherwise it would not constitute income). If T_r is equal to 65 and the deficit is half a year, income stream Y_r is incurred half a year earlier. At age 65, half a year of interest on this income stream is incurred. During that same half a year, additional income is earned. At age 65, we have Y_r including half a year of incurred interest and additional income of half a year in the form of Y_{T_r+1} . Both have to be discounted by the same amount of years.

Property 1: The optimal school starting age lies before the maximum of the Skill Technology Function

At the maximum of the STF, the academic deficit (δ) is 0. In this case, the effective interest rates in R_1 and C_1 are exactly equal. We can also assume that lifetime income outweighs all educational costs, since the net present value of optimal investment in education is non-negative. Because R_2 is positive and C_2 is 0 (since δ is 0 and there is no deficit to compensate), Revenues must outweigh Costs. Thus, an incremental decrease in SSA brings positive profit, thereby outperforming the starting age corresponding to the maximum of the STF.

Property 2: The starting age cannot be optimal if the academic deficit is equal to or larger than the change in the starting age ($\delta \geq \Delta SSA$)

If the change in starting age and the deficit are exactly equal, $\Delta SSA - \delta$ is equal to zero. This implies that both R_1 and R_2 are zero. Since educational costs are positive, net gains are negative. In practice, this would mean that we would start working at the same time as before, but incur extra educational costs from starting earlier with formal education. The (negative) effect of starting one month earlier with formal education (the Birth Month Effect) should never outweigh the returns of one month of education.

Property 3: Revenues are maximized where the change in deficit is equal to the change in starting age ($d\delta/dSSA=1$)

This represents the point where skill accumulation is most optimal. It is in between the two points previously mentioned. When we assume that individuals will enter the labor market once they reach a specific skill level, all parameters aside from ΔSSA and δ are fixed. Hence, Revenues will be maximized when γ is maximized. The slope of γ will be zero, when the change in δ equates the change in ΔSSA . Beyond this point, the deficit is increasing faster than the change in starting age. It will never be efficient to reduce the starting age beyond this point, because Revenues would decrease while Costs are constantly increasing. However, it might still be efficient to *increase* the starting age from this point. Although Revenues would decrease, Costs might decrease at a faster rate. The optimal starting age has to be somewhere between the maximum of the STF and this point. This is where the slope of the Skill Technology

2. THE THEORY OF THE SCHOOL CURRICULUM

Function ($dd/dSSA$) is somewhere in between 0 and 1. The exact location of this point depends on estimation of the STF and on the relative magnitude of lifetime income and educational costs. The three properties above are all irrespective of cost and income streams.

The point where profits are maximized is the economic optimum. Additionally, the point where Revenues are maximized ($dd/dSSA=1$) represents a secondary optimum: the skill optimum. It is that specific school starting age at which we reach a specific skill level at the earliest age, or it is that SSA where our γ is largest. This optimum is different from the economic optimum. At some point, Costs will rise stronger than Revenues, when we keep increasing ΔSSA . Since costs are constantly increasing, the skill optimum lies at an earlier starting age than the economic optimum. This property discards any possible criticism that taking economic factors into account leads to a bias towards earlier school starting ages. If we only look at skills, we reach an earlier optimal school starting age than if we take all relevant economic factors into account.

Whether starting school earlier is beneficial from an economic point of view depends on two tradeoffs. First of all, the time gain needs to be larger than the skill deficit, because otherwise benefits are negative (or zero at most). Secondly, if benefits are positive, they need to be large enough to offset the extra educational costs incurred when starting earlier. We can use part 1 of our analysis to estimate the skill deficit (or surplus) for every earlier (or later) School starting Age. Then, we can take a specific ‘benchmark’ school starting age, and use the specified analysis to see what the economic payoffs are of any deviating starting age. The method can easily be transformed to measure the effects of later school starting ages. In that case, the costs would be revenues (educational costs are incurred later and less education is needed) and the revenues would become costs (we can only start working at a later age, assuming the BME is small enough).¹

One could argue that costs are already made before C_1 , in the form of child care and kindergarten. If we would incorporate these costs as well (which we will not do specifically in this study), the economic optimum will move earlier and closer to the skill optimum (which would not change), since this involves larger costs for later starting ages. There are still costs from reducing the starting age even in that case since cost streams generally increase as students advance in their educational career. The smaller these differences between cost streams, the closer

¹Or, the ΔSSA and d would change sign.

to each other the two optima are.¹

2.2 Allocating Instruction Time

The process of skill formation depends, among other things, to a strong extent on the amount of time spent on specific topics and the order in which these topics are taught. Schools and teachers generally have a lot of freedom in deciding how they divide the range of school subjects over the school week. There is no consensus whatsoever on what the optimal allocation of instruction time over these subjects is. Economic theory tells us that optimal allocation depends on the initial productive value of (instruction in) each subject, how this value develops over time and the extent to which the acquired skills in a certain topic have spillovers towards other topics. The last two issues depend on the degree of complementarity between skills, both within a skill over time as well as across different type of skills.

2.2.1 Long-run effects

We approach the concept of instruction time as an investment in skills, analogous to the ‘Technology of Skill Formation’ (Cunha and Heckman, 2007). We specify that more instruction time (IT) means a higher investment in skills and therefore leads to higher outcomes. It is of great importance to investigate how an initial effect of investment develops over time. Direct positive effects of increased investment could either fade away or increase over time. This depends on two concepts: the degree of self-productivity and the degree of complementarity between skills (and thereby investments) over time. The initial value of additional instruction time can increase when it enhances the learning of new topics and thereby makes future investments also more productive. When early acquired skills have little to no value in later grades and complementarity is low, the initial impact of more instruction can fade away over time. Complementarity could be as low that it does not matter if investments take place early or late, as long as they take place at some point in time. In that case, the investments are perfect substitutes, and the learning process is purely additive. Since we conduct our analysis at three different age levels, we can get an indication of the degree of self-productivity and complementarity.

We assume that instruction time is the only productive investment. The skill set in the current period depends on skills in the previous period by a coefficient α and on investments in

¹If all cost streams (including the ones before C_1) would be exactly equal, costs are negative in our example in table 2.1. The total length of the cost streams is shorter in the case with the earlier starting age.

2. THE THEORY OF THE SCHOOL CURRICULUM

instruction time by a coefficient β . We ignore parental characteristics here.¹ We examine three different time periods or grades, being grades 2, 4 and 6. We have the following relationship between skill formation and the levels of instruction time in these three different periods:

$$\theta_1 = \alpha_1\theta_0 + \beta_1IT_1$$

$$\theta_2 = \alpha_2\theta_1 + \beta_2IT_2$$

$$\theta_3 = \alpha_3\theta_2 + \beta_3IT_3$$

We define γ as the effect of instruction time on skill formation. For second grades this is equal to:

$$\gamma_{i,t} = \beta_1$$

For 4th grades, we can estimate the effect of instruction time in that period, but also of instruction time in the previous period (being grade 2):

$$\gamma_{i,t} = \beta_2$$

$$\gamma_{i,t-1} = \beta_1\alpha_2$$

For 6th grades, we can retrieve the effect of instruction time on skill formation in two previous periods:

$$\gamma_{i,t} = \beta_3$$

$$\gamma_{i,t-1} = \beta_2\alpha_3$$

$$\gamma_{i,t-2} = \beta_1\alpha_2\alpha_3$$

In the model specified above, the parameter β captures both the degree of self-productivity and the degree of skill complementarity. These aspects determine how an initial effect of instruction time develops in later periods. We can specify these effects separately in the model, by introducing an interaction term between instruction time and incumbent skills. In this model, δ measures

¹In our empirical analysis, we ensure that other investments and parental characteristics remain constant.

self-productivity and τ measures the degree of complementarity:

$$\begin{aligned}\theta_1 &= \delta_1\theta_0 + \beta_1IT_1 + \tau_1\theta_0IT_1 \\ \theta_1 &= (\delta_1 + \tau_1IT_1)\theta_0 + \beta_1IT_1 \\ \theta_1 &= \alpha_1\theta_0 + \beta_1IT_1\end{aligned}$$

Although we can specify this theoretically, it is not possible to separately identify both these effects in the actual analysis, given the data at hand.¹ Therefore, both effects are picked up by the α 's in the previous model. If the parameter β is smaller than 1, the effect of IT in a specific time period (i.e. grade) on skills decreases in the long run. If this parameter is larger than 1, the implications of early investment will only become bigger in the long run (either through high self-productivity or high complementarity, or a combination of the two).

2.2.2 Spillover effects

The empirical analysis of Chapter 4 does not assess an extra-curricular program, but looks at changes in language instruction within the typical school week. This means that increases in instruction time for one specific subject (in this case language) take time away from other subjects. Instruction time in these other subjects can also have productive value for the skill we measure. Effects will be underestimated if the increase in IT comes at the expense of another subject that is productive for that skill. It is apparent that we need to take into account how instruction time for math has changed when estimating the effects of language instruction on mathematics performance. But spillover effects could be present in any area, requiring that we incorporate all school subjects.

The concept of spillover effects refers to a potential value of instruction time for one topic for performance in any other topic. Can language instruction only affect specific language skills, such as spelling or reading, or could it have more general effects? There are different schools of thought on this relation. Learning could be seen as comprehending specific skills. One is taught a specific concept and by learning, executing and repeating, one can master such a skill. In order to improve the performance on other skills, separate and specific instruction on these skills is necessary. Alternatively, it could be the case that instruction in a specific skill has beneficial effects for other skills as well. This can work directly when, for example, reading skills are

¹Since all these variables are endogenous, this would require a multitude of instruments, and hence extremely rich data.

2. THE THEORY OF THE SCHOOL CURRICULUM

productive for studying mathematical topics. It can also work by indirectly enhancing general intelligence. There is a big and ongoing debate about the malleability of general intelligence. The controversial book ‘The Bell Curve’ (Herrnstein and Murray, 1994) argues against it, but there are also strong voices in favor of it.¹ Almlund et al. (2011) summarizes evidence on the development of cognitive and non-cognitive skills and concludes that IQ stabilizes at age 8. Whether through general intelligence or through ‘intermediate’ skills, spillover effects are present in both these cases.

The implications of these different schools of thought are different. In the former case, increases in language instruction should only affect performances on language related tests. Moreover, it would imply that increases in instruction in one specific subject would always come at the expense of other subjects, which naturally would have less instruction time available. In the latter two cases, increases in language instruction would have positive implications for performance in multiple school subjects and might even enhance a general intelligence level (and could further enhance performance in other subjects, through IQ).

Hence, the actual estimates of increased language instruction do not only depend on the value of language but also on where this time is retrieved from and on the productive value of this particular activity. The underestimation of the estimate is equal to the product of the two, summed up for all school subjects. We define the degree with which instruction in any other subject decreases as σ . This decrease is measured relative to the increase in instruction of the former subject (i.e. language), so that all σ ’s add up to one. We define the productive value of each subject as ω . Changes in language instruction (ΔIT^L) affect the levels of instruction in other subjects:

$$\Delta IT^L = -\Delta IT^M - \Delta IT^X - \dots - \Delta IT^Z$$

Any increase in language instruction comes at the expense of other subjects. The superscript M refers to mathematics and the superscripts from X until Z refer to all other school subjects (besides language and mathematics). The total decrease is divided over all these subjects. We specify how much time each subject loses, by applying σ ’s as weights:

$$\Delta IT^M + \Delta IT^X + \dots + \Delta IT^Z = -\sigma_1 \Delta IT^L - \sigma_2 \Delta IT^L - \dots - \sigma_n \Delta IT^L$$

with $\sigma_1 + \sigma_2 + \dots + \sigma_n = 1$.

¹See e.g. Devlin et al. (1997).

Combining this with the productive value of each topic (ω), the payoff on language skills can be represented as:

$$\begin{aligned}\Delta\theta^L &= \beta^L \Delta IT^L + \omega_1^L \Delta IT^M + \omega_2^L \Delta IT^X + \dots + \omega_n^L \Delta IT^Z \\ \Delta\theta^L &= \beta^L \Delta IT^L - \sigma_1 \omega_1^L \Delta IT^L - \sigma_2 \omega_2^L \Delta IT^L - \dots - \sigma_n \omega_n^L \Delta IT^L \\ \Delta\theta^L &= (\beta^L - \sigma_1 \omega_1^L - \sigma_2 \omega_2^L - \dots - \sigma_n \omega_n^L) \Delta IT^L\end{aligned}$$

Similarly, the effect of changes in language instruction on mathematical skills is equal to:

$$\Delta\theta^M = (\beta^M - \sigma_1 \omega_1^M - \sigma_2 \omega_2^M - \dots - \sigma_n \omega_n^M) \Delta IT^L$$

To simplify, we specify the whole negative component of increases in language instruction, resulting from changes in other IT variables, as λ . λ is a weighted average of all the added values of all other school subjects, with the σ 's applied as weights:

$$\lambda = \sigma_1 \omega_1 + \sigma_2 \omega_2 + \dots + \sigma_n \omega_n$$

Taking all of this together, the actual effects of instruction time that we are estimating are equal to:¹

$$\begin{aligned}\gamma_{i,t}^L &= \beta_3^L - \lambda_3^L \\ \gamma_{i,t-1}^L &= (\beta_2^L - \lambda_2^L) \alpha_3 \\ \gamma_{i,t-2}^L &= (\beta_1^L - \lambda_1^L) \alpha_2 \alpha_3\end{aligned}\tag{2.1}$$

The estimated effect of instruction time for language incorporates a potentially negative effect of receiving less instruction in other subjects. The presence of this negative effect depends on whether these other subjects are productive for that specific skill. The value of the specific instruction that we examine (in our case language) has to outweigh the weighted average of the values of instruction in those subjects that suffered decreases in instruction, if one wants to find a positive effect of language instruction. The fact that we do not measure the pure added value of language instruction (β) does not limit the relevance of our findings. Any increase in instruction time for a specific subject comes at the expense of instruction in a different subject, if we hold the length of the schooling week fixed. From a policy perspective, ' $\beta-\lambda$ ' is the

¹This case refers to 6th grade language, but one can put any skill in here.

2. THE THEORY OF THE SCHOOL CURRICULUM

coefficient of interest. Only, it is not a fixed coefficient since it depends on changes in other subjects. If one is estimating the effect of increases in a specific subject within the typical schooling week, one needs to know where this extra time has come from (and, ideally, also the added value of instruction in these topics), since this has large implications for the interpretation of the estimates.

2.3 Tracking

It is common practice for the majority of countries to select students into specific tracks at some point in their educational careers, based on the ability of these students. Virtually all developed nations employ some sort of differentiating mechanism to deal with heterogeneity in ability. Countries such as the U.S. or the U.K. make use of the more informal practice of ability-grouping, which groups students within a class according to their capabilities in a certain school subject. We are specifically looking at formal tracking, which means separating students into distinct educational programmes. Countries that employ tracking differ in the number of tracks employed and the specific age at which tracking takes place, as well as the exact division of students to tracks. Although the assignment mechanism and its strictness differs across countries, a student generally needs to have a certain ability level in order to be eligible for a certain track. Two important questions with respect to the efficiency of tracking systems is whether currently employed thresholds send students to their appropriate tracks and what the impact is of changes in the age of tracking. These questions are analyzed in Chapters 5 and 6.

In Chapter 5, we examine whether tracking systems send all students to their most optimal track. If this is the case, the threshold lies before the first student in the distribution for which the higher track is best suited and after the last student in the ranking distribution for which the lowest track is most suitable. In other words, the threshold will lie at the student that is indifferent between the two tracks. Figure 2.2 illustrates this reasoning. In the example, there are three different tracks: a low, a medium and a high track. The low track is best suited for students low in the ranking, but is not a good fit if we move up along the distribution. The higher tracks are being outperformed by the lower track low in the distribution. As students improve in ability, this gap in performance decreases until, at some point, students are better off with a higher track. When thresholds lie at these cutoff points, each student goes to the one track of the three that is best-suited for him. The effective outcome line follows the highest points

of all three lines (the thick line) and individual outcomes are maximized.¹ When thresholds are located at different points, some students will be better off in other tracks. The effective outcome line will not follow the highest points anymore and there will be discontinuities in outcomes along the distribution. Figure 2.3 describes the case when the lower threshold is more strict or more lenient, respectively. In the former case, the better students in the low track would have been better off with the medium track and hence there is a relatively large jump between the highest low-track students and the lowest medium-track students. In the latter case, the lowest medium-track students would be better off with a lower track and there is a downward jump in the effective outcome line. When such discontinuities (whether positive or negative) are found in actual outcomes, the assignment system in place does not send each student to his best option.

This is not necessarily the same selection that maximizes societal output. The placement of thresholds will affect more than just those individuals around the threshold. Hence, we are making no claims with respect to the optimal track assignment from the point of view of society as a whole.

From an economic perspective, a difference in track assignment can be seen as a difference in investment. This investment consists of a specific curriculum and a set of peers. We again interpret this from the perspective of The Technology of Skill Formation from Cunha and Heckman (2007). When we compare two identical individuals who are sent to different tracks, the parameters θ_t and h will be equal, but I will differ. Presumably, this will lead to a different θ_{t+1} in the future period. This difference represents the effect of the Investment on skill formation. Ideally, estimation of the effect of track assignment holds both θ_t and h constant (and any unrelated investments).² Depending on the degree of self-productivity and complementarity, any estimated initial effect of any investment in skill can develop in very different ways, which makes it crucial to estimate both short- and long-run effects.

The difference in investment we are analyzing consists of a different curriculum, different peers and possibly different teachers as well. The data to disentangle all three of these effects are not available, which is why our estimates pick up on all of them. We explicitly take an individual point of view in this analysis. ‘Efficient’ selection as in figure 2.2 means that individuals cannot switch and be better off. Estimates should not be interpreted from a public policy perspective,

¹This refers to individual outcomes given the assignment of all other students. No one can switch and make himself better off, which essentially represents a Nash equilibrium. This is not necessarily the division that maximizes societal output.

²We will do so by employing a regression discontinuity design.

2. THE THEORY OF THE SCHOOL CURRICULUM

because changing effective thresholds also changes the whole system of peer effects. By changing thresholds, the slopes and intercepts of the lines in figures 2.2 and 2.3 would change as well. When we implement thresholds that maximize individual outcomes given the assignment of the rest of the distribution, this will result in a too lenient assignment mechanism if we assume that peer influences are more beneficial in higher tracks. A stricter selection (such as in the left panel of figure 2.3) can be more beneficial since it will increase peer quality in tracks on both side of the threshold. Our estimates are valuable for the individuals who are around the required ability level for a specific track. They answer what would be the best choice for the person at the border of the two tracks in figures 2.2 and 2.3. This represents valuable information for children and their parents who face this choice.

The framework can be applied in the same way to estimate the impact of changes in the age of tracking. This essentially introduces one additional track, which is a comprehensive track. This again represents a different investment, including a different curriculum, different peers and different teachers. Being in a comprehensive track for one additional year represents a shorter treatment than when comparing two specific tracks which are different throughout secondary education. Still, because of self-productivity and complementarity, the long-term impacts of such treatment can still be large and can continue to expand when the treatment has finished. This is emphasized by the fact that tracking age can also affect the track you end up in.

2.4 Grade Retention

The effect of grade retention is often presented as a tradeoff between receiving an extra year of repeated instruction and the (presumably negative) psychological effects that come with the ‘stigma’ of being a repeat student.¹ This does not capture the total impact of being retained. The total effect is a sum of psychological effects and the effect of a difference in curriculum (receiving the same instruction again versus receiving new instruction that builds upon the previous). These effects do not need to have opposite signs, hence we cannot ex ante speak of a tradeoff here.

The choice whether to retain a student or not is a choice between two different pathways of skills formation. As such, we approach it in the same manner as we do in our discussion of track assignment in the previous section. We again use The Technology of Skill Formation, where

¹See e.g. Alexander et al. (2002); Jackson (1975); Pagani et al. (2001a).

Investments differ between the person that is being retained versus the identical individual that is being promoted.

When we are comparing a retained student with his promoted counterfactual, parental characteristics and incumbent skills should not differ.¹ Retention influences skill development because it affects the investment in the next period.² Investment takes the form of a more basic, repeated, curriculum (I_0) or a more advanced, new, curriculum (I_1). Investment I_0 provides a stronger basis by repeating core concepts, while I_1 develops a broader set of skills. The notion that grade retention can be beneficial is driven by the concepts of self-productivity and complementarity. Earlier skills beget later skills. Repeating material from an earlier grade can develop such fundamental skills and enhance the productivity of future investments in skills. There are two scenario's possible. In scenario 1, the retained student has learned more one year after retention than when progressed. He will have both a better overall skill set and a stronger basis. Retention should provide a positive impact, both in short and long run. In scenario 2, the skill set is higher from the promoted curriculum, after year one. There can still be a positive effect of retention because fundamental skills should be better developed. Repeated investment I_0 should at least have some positive impact, implying that the student enters the next grade with a higher skill level than when promoted. Because skills are self-productive, this has positive impacts on future skill development. When there is strong complementarity between skills, it is possible that the retained student will catch up to and even overtake his promoted equivalent. The above also implies that the retained student can never be more than one schooling year behind in skills, since he should always learn something in the repeated year and will enter any grade with a higher skill level than when promoted.

We now add psychological effects to the model. There can be positive psychological effects, because retention puts students on an easier pace and can lower self-confidence problems from being behind the rest of the class. Psychological effects can be negative through the social stigma that retention can cause and the fact that it removes students from their peers. Previous discussion has emphasized these negative consequences. If the balance of all psychological effects (labeled ψ) is indeed negative, this can have a negative impact on further skill development throughout the life cycle. The student who already had a higher skill set one year after repeating (scenario 1) could then even be overtaken by his promoted counterfactual. Moreover, the

¹By employing month of birth as an exogenous instrument for grade retention, we ensure this holds true in our empirical analysis as well.

²And also in subsequent periods, since the repeated student is always one grade behind.

2. THE THEORY OF THE SCHOOL CURRICULUM

negative impact of retention can now also exceed one schooling year. The Technology of Skill Formation becomes:

$$\theta_{t+1} = f_t(\theta_t, h, I_t, \psi_t)$$

The model also recognizes that the impact of grade retention differs by ability level and by age of the student. Better students benefit relatively more from the advanced investment I_1 than from I_0 and are less likely to gain from grade retention. The model suggests that retention has more favorable effects when executed at younger ages. First, a student who is retained at a younger age has more time to catch up to his promoted counterfactual, through the presumed effects of self-productivity and complementarity. Secondly, complementarity effects are higher at younger ages (Cunha and Heckman, 2007; Cunha et al., 2010). This is primarily because the skills developed early contribute to learning for a larger part of the life cycle. Moreover, skills developed at earlier ages generally have stronger spillovers to other topics than those at later ages, which are generally more specific. Developing better early reading skills will be more beneficial in the learning process than a better understanding of differentiating equations.¹ Additionally, it is likely that negative psychological effects are lower for early retention. It does not remove children from long-established peers and the social stigma from being held back a year in kindergarten is generally smaller. We also expect this to hold if we compare retention in primary school versus secondary school. The former is generally decided between the pupil, the teacher and the parents, while the latter is usually based on a formal requirement with respect to the students' grades. The sense of failure is expected to be stronger in the second case.²

Our approach focuses on the effect of retention on those who are effectively retained. Retention can also have effects on the untreated. Primarily, the threat of potential retention can work as an effective motivator to work harder. Additionally, separating students that lag behind can put the pace of instruction at a higher and more appropriate level for the rest of the class. Hence, retention also leads to a change in peer quality and to psychological effects on the untreated. These effects are an interesting topic for future research, but beyond the scope of the study we present in Chapter 7.

¹For example, Cunha et al. (2010) find that non-cognitive skills are cross-productive for cognitive skills, but only at early ages. The existence of stronger complementarity early in life is confirmed in Chapter 4, where early language instruction has spillovers to other skills. These spillovers are absent for language instruction in later grades.

²The proposition that the stigma from retention is higher when retained at later ages is made in, for example, Mantzicopoulos and Morrison (1984); Shepard and Smith (1986).

2.5 Figures

Figure 2.1: The Skill Technology Function

Notes: The graph specifies the development of learning productivity over age, for the formation of literacy skills. Learning productivity refers to the speed at which you learn. It determines how the skill set develops from period t to period $t+1$.

Figure 2.2: Effect of track on outcomes: efficient selection

Notes: The graph shows the hypothetical effect of track assignment on outcomes for the full ability distribution, for three different tracks. Higher tracks are better suited for students higher in the ranking and hence have a stronger slope but a lower intercept. The effective outcome line traces out the maximum points of all three tracks, since track assignment is efficient.

Figure 2.3: Effect of track on outcomes: inefficient selection

Notes: The graphs show the hypothetical effect of track assignment on outcomes for the full ability distribution, for three different tracks. Higher tracks are better suited for students higher in the ranking and hence have a stronger slope but a lower intercept. The effective outcome line shows the realized outcomes, given the specified thresholds between two certain tracks. In the left panel, assignment is stricter than under the smooth system of figure 2.2. This results in a positive discontinuity in outcomes at the threshold between low and medium track. In the right panel, assignment becomes more lenient, resulting in a negative discontinuity in outcomes at the threshold between low and medium track.

3

An Economic Analysis of the Optimal School Starting Age¹

3.1 Introduction

Acquiring literacy skills is crucial for the development of a child's skill set. An important question is at what exact age schools should start with teaching these skills. Psychologists and school teachers often argue that children need to reach a certain maturity level before they can effectively learn reading and writing. They therefore advocate relatively late school starting ages (Shepard and Smith, 1986). However, from a human capital point of view, delay of investment lowers its value (Becker, 1964).²

The aim of this study is to empirically investigate the timing of learning literacy skills from an economic point of view. We model this tradeoff between starting either earlier or later in the child's life by introducing a skill technology function where learning becomes more effective as children grow older and where the effective returns are larger when learning is scheduled earlier in life. Empirically, we identify the slope of the skill technology function by comparing students by their month of birth and the curvature of this function by comparing different countries. We use data from TIMSS, PIRLS and PISA for 25 different countries. We find that students that start school at an earlier age perform worse than their peers that started at a later age. However, this negative effect is offset by the (economic) advantages that these younger students have because they are younger when they finish school.

¹This chapter is based on Borghans and Diris (2010b).

²Assuming that the costs of investment are the same. Chapter 2 specifies that they differ in this case, which should also be incorporated.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

We compare test results of students with respect to their month of birth. Birth months determine at what exact age someone starts reading and writing. Thus, students from different birth months start at different (exact) ages with learning literacy skills. Learning productivity is likely to differ across these different ages. More specifically, learning productivity over age, as modeled by the Skill Technology Function, is an S-shaped curve. By comparing students with different birth months, we can determine the slope of this Skill Technology Function at a specific interval. By conducting the same analysis for countries with different school starting ages, we can also determine the curvature of the function, since we are looking at different intervals. We measure inter-student differences at the 4th grade, 8th grade and 15-year-old level. In this way, one can see how differences in performance develop over time. This allows us to judge whether differences in reading proficiency have strong complementarity effects for the remainder of the educational career. We combine these estimates with human capital theory to reach our optimal school starting age.

The organization of this chapter is as follows. Section 3.2 discusses prior literature on school starting ages, both in the field of economics and psychology. We describe the data that we use in the empirical analysis in section 3.3. Section 3.4 discusses methodological issues concerning the data and the estimation method. Section 3.5 shows our main empirical results and the economic implications of these results. Robustness checks and additional estimation are discussed in section 3.6. Section 3.7 discusses the implications of our results and presents directions for future research.

3.2 Literature

3.2.1 Economic literature

This study is related to economic literature estimating the optimal starting age for school. Strøm (2004) presents a theory that initial (meaning pre-school) endowments play an important role. Children already learn and develop before they start formal schooling, which comprises these endowments. It is intuitive that those that are older have built up more endowments. These children are also expected to perform better in school, at least in the early grades. This endowment theory can be interpreted in different ways. On the one hand, older children built up more endowments. On the other hand, younger children already go to school at an age at which the older ones are still at home. In his paper, Strøm (2004) focuses on this accumulation of endowments or human capital. He does not consider the effects of the timing of investments on their

subsequent value, which is the main focus of this study. The endowment theory of Strøm relates to the concept of sensitive periods in Cunha and Heckman (2007), which refers to periods that are more effective in producing certain skills. Such differences in endowment and sensitivity ultimately determine which students are more productive in learning skills.

How initial differences in performance develop over time depends on several issues. Peer effects is one of these factors. The concept of peer effects refers to the fact that individuals can be affected by interactions with other classmates (Winston and Zimmermann, 2003). If a child is lagging behind, which could be caused by relative age or maturity differences, this can hamper his self-esteem. This can further increase initial differences in performance. On the other hand, a child can also be stimulated to catch up with the rest of the class. Secondly, the development of initial performance differences can depend on the degree of complementarity between skills (Cunha and Heckman, 2007). Complementarity implies that investments are more productive when a high level of skills is present. As such, early investments in skills make later investments more productive. When complementarity is low, skill deficits levels in early grades can be compensated in later periods. Thus, with strong complementarity initial performance differences between older and younger children within a grade will only increase, while they can decrease or even dissipate in absence of such effects (see section 2.1.1).

There have been several studies in recent years using the exogenous variation in month of birth when assessing student performances or labor market outcomes. Bedard and Dhuey (2006) look at the effect of quarter of birth on student's test scores, for a wide range of countries, using the predicted age of the student based on his quarter of birth and possible retention and skipping. They find that quarter of birth has a significant effect on the performance of students and that the younger students in class have a lower chance of attending college. There are several country-specific studies on differences in starting ages.¹ These studies generally focus on the relative disadvantage that younger entrants have in class and do not specifically examine the implications for absolute starting ages.

The above studies examine the effect of starting ages or month of birth on student performance. Other papers look also at long-term effects in terms of educational attainment or earnings. Fredriksson and Ockert (2005) find that those that start school later have better GPA's, a higher chance of getting a college degree and earn a higher wage in the labor market as well (although the latter effect is insignificant). Black et al. (2008) find a small positive effect of

¹See Bell and Daniels (1990) for the U.K., Puhani and Weber (2005) for Germany, and McEwan and Shapiro (2008) for Chile.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

starting younger on IQ at age 18, and a short-term positive effect on earnings, which disappears by age 30. Although such long-run effects are very informative, they run an increasing danger of picking up an relative age effects, which are generally emphasized over time by institutional aspects. The fact that early starters have a lower chance of getting a college degree (through the fact that their test results are lower because of lower maturity at the time of the test) can translate into lower wages. This is a relative, not an absolute, age effect. In the current educational setup, relatively younger students are always at a disadvantage (as long as we are on the upward part of the STF), regardless of whether the starting age is 5, 6 or 7 years. This is a consequence of the fact that issues such as grade retention, track selection and university attendance are based on test scores which are taken at different exact ages. Although this is certainly an important issue, it is independent of (absolute) starting ages.

The added value of our study is that we specifically measure the tradeoff between the endowment and the human capital argument. Additionally, we estimate birth month effects for an extensive range of countries, at different age levels, using several data sources. This contrasts with other studies, which either focused their analysis on one country or did not (fully) address the relation to school starting ages, nor the implications in terms of economic effects.

3.2.2 Psychological literature

This study is also related to work in cognitive psychology. Psychologists and neurologists have been conducting a vast amount of research concerning optimal school starting ages. Cognitive theory suggests that the effect of starting school early is twofold. On the one hand, younger children could be more receptive for learning than older ones and, therefore, learning should start as early as possible. On the other hand, it could also be the case that first a certain age needs to be reached before they can understand the (more complex) material (Shepard and Smith, 1986). In the 1990s, there was a preference for the former argument, especially in the US. This perception was based on three major findings: there are dramatic increases in the number of connections in the brain during infancy; there are critical periods for learning in this same time period; and enriched environments (such as schooling) can further increase the forming of connections in the brain. However, Blackmore and Frith (2005) argue that despite the bulk of neurological research, we still know very little about the developing brain. Most research is based on monkey studies and we cannot simply generalize these findings to humans (Bruer, 1999). Additionally, most studies concerning the development of the brain have focused on visual development. Much less is known about the frontal cortex, where planning and decision-making takes place. There

are indications that the development of the frontal cortex takes place over a much longer period than the development of the visual cortex (Huttenlocher and Dabholkar, 1997). This questions whether there truly are critical periods for learning. It is more common to speak about sensitive periods; stages where learning is most effective, but not necessarily vital. However, whether (strong) sensitive periods exist for reading, writing and arithmetic is not known, nor where they are located. These recent arguments have led to a movement towards advocating later starting ages by the majority of psychologists in this century.

Although psychological and neurological research do not give clear answers on optimal school starting ages, there is consensus on the fact that only very few children have the necessary skills and competencies to start reading and writing before the age of four. In general, psychologists are more focused on the individual child and do not draw conclusions on what would be ideal from a general point of view. Whether it is better to start reading at five or seven and for whom, is a question that has not yet been answered (Eden et al., 2003). Thus, in relation to the Skill Technology function, the curve is expected to be rather flat until at least the age of 4. However, where the function is at its steepest and at what age the maximum lies are questions still left unanswered. Moreover, whenever psychologists are interested in finding the optimal point to start reading, they want to find the top of the STF. Section 2.1.1 indicates that both the skill optimum and the economic optimum lie somewhere before that top.

3.3 Data

The data we use in this chapter stem from three international assessment tests: TIMSS, PIRLS and PISA. TIMSS assesses the knowledge of 4th and 8th graders in mathematics and science, on a regular four-year cycle. Overall scores on math and science are reported for each student. These overall scores are the focus of our study. The study started in 1995 and has also been conducted in 1999 and 2003. The '99 study is an exception in the sense that it only tested 8th graders, while the other two included both 4th and 8th graders in their study. TIMSS is presented as a study of 4th and 8th graders, but this is not entirely accurate. In practice, it is a survey for those grades that contain the largest proportion of 9-year-olds and 13-year-olds. This means that, for countries that start school one year earlier, 5th or 9th graders are examined. For countries that start school one year later, the dataset includes 3rd or 7th graders. The exception to this rule is the TIMSS 2003 study, which only contains 4th and 8th graders. One has to be aware of this when interpreting the results from these studies. The '95 dataset has significantly

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

fewer observations per country than the other two waves. Additionally, this wave reports data for students in two different grades. The official population is composed of "Students enrolled in the two adjacent grades that contained the largest proportion of 9-year-old students at the time of testing" and "Students enrolled in the two adjacent grades that contained the largest proportion of 13-year-old students at the time of testing" respectively. In most countries, this corresponds to 3rd and 4th grade in the former case and 7th and 8th grade in the latter. The more extensive setup of this dataset gives us some important possibilities on which we will elaborate later on.

PIRLS (Progress in International Reading Literacy Study) focuses on trends in the reading achievement of 4th graders, on a five-year cycle. PIRLS studies were conducted in the years 2001 and 2006. All countries contain only 4th graders, with the exception of New Zealand (5th grade) in PIRLS 2006 and England (5th grade) in PIRLS 2001. This is somewhat inconsistent, since 4th graders are included again for New Zealand in 2001 and England in 2006. Again, we take this into account when evaluating our results. All countries in both datasets exhibit at least 3,500 observations.

PISA (Program for International Students Assessment) is an international test of school performance of 15-year old students (largely 10th-graders). It contains a multidisciplinary test for students at the end of compulsory education. PISA studies were executed in the years 2000, 2003 and 2006. The PISA 2006 study is the most sizeable wave, with 9,000 observations per country on average. All countries exhibit at least 3,500 observations. PISA 2000 contains the fewest observations per country. The assessment is being described as a study among 15-year olds, which in practice means that students who are born within one year of each other in a particular country are examined. For all countries that are included in our sample, this comprises students who are born in the twelve months after the formal cut-off date. They test student in reading, mathematics and science, thereby combining all of the disciplines that are measured in the TIMSS and PIRLS studies.

We use scores on either math, science or reading as the dependent variable in our within-country analysis. All three studies do not report final scores, but plausible values (PV's). On every topic, five different plausible values are reported. Our ultimate score variable is the average of these five PV's. All datasets also contain a large range of background questions.¹ To sum up, data are available for 4th graders for reading (PIRLS '01, PIRLS '06) and mathematics and science (TIMSS '95, TIMSS '03), for 8th graders on mathematics and science (TIMSS '95,

¹E.g. gender, country of origin, language spoken at home, parental education, number of people in the household etc.

TIMSS '99, TIMSS '03) and for 15-year-olds on mathematics, science and reading (PISA '00, PISA '03, PISA '06). The fact that we measure the effect of month of birth on score for different ages allows us to estimate skill growth levels along the Skill Technology Function.

The analysis includes a total of 25 countries for which the birth month effect is estimated. Not all countries are included in every dataset. All countries under examination are OECD countries, in order to ensure that they all have sufficiently developed educational systems. Some countries are excluded from the analysis. The U.S. and Australia employ region-specific cutoff dates and the data did not properly identify which students belonged to which specific area. Additionally, Slovenia and Norway have recently changed their school starting ages. Since these changes were done gradually and the individual data for these samples are hard to identify, these two countries are discarded from the analysis as well. Some countries are excluded for specific data sets, because of unsuitable samples. For example, several countries in the PISA data include students from within one calendar year, instead of one year after the cutoff date. Although some countries have exactly similar school starting ages, there is still substantial variation among (exact) starting ages. In total, that variation presents us with 12 unique ranges for which to estimate the Birth Month Effect. Starting ages (for the average-aged student) range from 60 months for Scotland to 86 months for Finland, Poland and Sweden.

Not all countries strictly enforce cutoff dates, allowing some flexibility to parents on when to enter their child. By introducing this choice, the starting age becomes endogenous for these countries. However, this should be corrected with the use of our particular model (we will elaborate on this in section 3.4.1.1). Hence, these countries are included.¹

Additionally, we use data on the costs of education and average income statistics. This information is used for the cost-benefit analysis specified in section 2.1.2. For both income and educational costs, we used Euro-Area averages reported via Eurostat. We specify five different age groups, which each have a distinctive mean income. Costs are estimated separately for primary, secondary and tertiary education.

3.3.1 Descriptive Statistics

We first display graphs to get a first impression of the nature of the Birth Month Effect. Based on section 2.1.1, we would expect mean scores steadily declining as birth month increases (children become younger). Figure 3.3 shows the pattern of mean reading scores in grade 4 per birth

¹There are only three countries in our study that can be classified as non-strict: Germany, Hungary and the Netherlands.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

month, as an example. The first month on the x-axis corresponds to the first month after the official cutoff date. We observe the expected pattern of decreasing performance when birth month increases. Students who are born later and therefore are relatively younger score worse than their relatively older classmates. We also observe that this effect is fairly linear along a one-year range. The predicted difference between the very youngest and very oldest student is around 15 points, on a standard deviation of approximately 70.¹

3.4 Methodology

The analysis of this study focuses on the effect that month of birth has on the test scores of students, for different age groups and in different countries. Since a person's birth month decides his or her exact school starting age (SSA), we can use these results to draw conclusions concerning optimal school starting ages. In the first part of our analysis, we focus on estimating the Skill Technology Function. We can thereby analyze the implications of differences in productivity when starting reading and writing, throughout schooling life. The estimation of the STF can be split up in two further parts: the within-country analysis and the between country-analysis.

3.4.1 Within-country analysis

We conduct several within-country analyses to estimate the *slope* of the STF on different points. We examine children born within one year after the school cutoff-date for every country. These children started formal education at the same point in time, but differed in (exact) age at that point. We estimate what the implications of these differences in SSA are on skill levels, which corresponds to the slope of the STF for that one-year range. We label that slope the Birth Month Effect (BME). By executing this for different countries with different school starting ages, we have different ranges along the curve, and, hence, different slopes. We use test scores as a proxy for skill level in our analysis. Additionally, we employ a range of controls. The list of potential controls is not equally extensive in every dataset, which means that not all regressions employ exactly the same controls. Gender, language spoken at home and country of birth of child, mother and father are included in every data set. Parental education level data are available for

¹These scores are already corrected for the presence of older repeaters, as discussed in section 3.4.1.1.

the 8th grade and 15-year-old samples, while number of people living at home is available for all TIMSS and PIRLS datasets. Hence, our model becomes:

$$Score_i = \beta_0 + \beta_1 Month_i + \beta_2 X'_i + \epsilon_i \quad (3.1)$$

Where X' refers to a vector of controls. Month is defined from 1 to 12, where 1 is the first month after the cutoff date for each country (so not necessarily January) and 12 the month just before the cutoff date. This equation assumes that the estimated effect is linear on this one-year range. Figure 3.3 already showed that this is a fair approximation.

3.4.1.1 Correction method

There is an important difference in the setup of TIMSS and PIRLS on the one hand and PISA on the other hand. The former have a population sample that includes students that are in a particular grade, while the latter examines students born within one year. The PISA data are more suitable for measuring the Birth Month Effect, because it is the date of birth and not the grade that is exogenous. The other two datasets contain repeaters that are older than their classroom peers (and also older than the repeaters in PISA). For TIMSS and PIRLS, the test scores for repeaters contain a surplus, resulting from the extra schooling that repeaters have had.¹ This problem is aggravated by the fact that repeating is more frequent for later birth months. The overestimation in score is therefore strongest in later months, leading to an underestimation of the BME. For those who have skipped a grade, a similar but opposite reasoning applies. Test scores increase with age and years of education, which are both higher for repeaters in the grade-samples. On the other hand, repeaters are of lower ability. We need to disentangle these two effects to estimate what the ‘bonus’ from being older (and having had more schooling) is. This bonus needs to be subtracted from their score, so that we are left with the right measure of their skill level.

We construct a correction method, which allows us to estimate this score bonus for every dataset. We instrument retention by the mean level of repeating for that month in that particular country, net of the own value for retention status. We label this instrument as *meanrep*. It is a strong predictor for repeating, but will not be correlated with ability levels, which are assumed to be evenly distributed over the year. If we estimate the effect of repeating using OLS, the coefficient will both measure the (positive) age and years of schooling effects and the (negative)

¹In relation to the Technology of Skill Formation (Cunha and Heckman, 2007), the difference in endowment is now also related to a difference in Investment, in the form of Years of Schooling.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

ability effect of repeating on score. By instrumenting, we take away the latter component and are left with the appropriate ‘score bonus’. Month and country dummies are included as well. We then get the following IV-model:

$$Repeat_i = \beta_0 + \beta_1 meanrep + \sum_j \gamma_j Country_j + \sum_k \delta_k Month_k + \epsilon_i$$
$$Score_i = \beta_0 + \beta_1 Month_i + \beta_2 Repeat_i + \sum_j \gamma_j Country_j + \sum_k \delta_k Month_k + \epsilon_i \quad (3.2)$$

If we estimate this model, the coefficient on repeat gives us the bonus in score that ‘older’ repeaters have because they have had more years of schooling. This is the number we have to subtract from the test score of repeaters. Their new score is what they would have scored if the test had been taken at the same age as their non-repeating peers. We call this adjustment in test score the *correction factor*.

We can test this correction method with the help of the TIMSS ’95 study, which presents data for two adjacent grades. With this dataset, we can discard all observations outside the relevant birth cohort to create our ‘pure’ sample of one birth cohort (since we can now locate repeaters from the same birth cohort in the earlier grade) and compare it to the ‘rough’ sample of one single grade. This dataset does not include skippers and those that have repeated more than once. However, the double repeater group is only very small (5 % of total repeaters) and the essence of the created sample is that it gives the opportunity to create a correction for repeaters. If this is established, a similar method can be applied to skipping. We apply the correction factor estimated in model 3.2 on the rough grade-sample and see whether that gives us a similar coefficient as the one we find in the pure cohort-sample.¹

The coefficient we find for repeat, applying model 3.2 for the 8th-grade dataset, is 29. This is very close to the actual difference between the two different groups of repeaters (30). When we apply this 29 as a downward correction on our repeaters in the grade-sample and estimate model 3.1, we get a coefficient of -1.32. The normal model for the pure sample, gives us a coefficient of -1.27. When we do not apply the correction, the coefficient is -0.918, hence correcting matters significantly. Figures 3.1 and 3.2 further show that the estimates with the correction factor are much closer to those from the ‘pure’ birth cohort sample than when we

¹Skippers and double repeaters are excluded in the former, to make the samples better comparable.

do not apply any correction. The correction factor is estimated for the whole sample, since the number of observations is too low to do this on a country level.

The correction method is also applied to correct for late starters, thereby making it possible to include countries that do not strictly enforce cutoff dates. We treat these late starters as if they are repeaters. Although they did not have extra schooling as did the repeaters, they did spend an extra year at home, and are one year older than their classroom peers. Hence, we assume that the gain of this extra year at home is equal to the gain of the additional schooling year.¹ Overall estimates are accurate irrespective of the validity of this assumption, since our estimated correction method treats both actual repeaters and late starters as repeaters. The estimated correction factor is a weighted average of the ‘score bonus’ of both these groups and is therefore accurate if we look at the sample as a whole, even if the two gains differ. On a country-level, estimates are only affected when both the difference in ‘score bonus’ between the two groups and the difference in distribution of late starters versus actual repeaters compared to the sample as a whole are large. Therefore, it appears to be warranted to include the non-strict countries into our analysis.²

In order to receive the most accurate measure for the correction factor, we grouped together those datasets from the same study and grade level³ when we applied the correction method. This also gives us enough observations to estimate a correction factor for skipping, which is less frequent than repeating.⁴ There are 10 different correction factors; 2 for PIRLS (repeating and skipping for reading in 4th grades) and 8 for TIMSS (repeating and skipping for both math and science in both 4th and 8th grades).

3.4.2 Between-country analysis

The within-country analysis allows us to estimate the slope of the STF at different ranges along the function. By conducting a between-country analysis we can use these different ranges to estimate the *curvature* of the function. This allows us to estimate whether the Birth Month Effect (which is the slope of the STF) increases if we move further back along the curve. We

¹This assumption is also implicitly made in the ‘pure’ PISA sample. Late starters have had one year less schooling than the rest of the cohort, but also one more year at home. These two effects are assumed to cancel each other out. Therefore, no correction is needed for PISA.

²The correction method is just as accurate for these countries as for the rest of the sample.

³I.e. the two PIRLS studies, the two TIMSS 4th grade studies and the three TIMSS 8th grade studies.

⁴In this case, scores need to be adjusted *upward*, to compensate for the lack of schooling that the younger skippers have had.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

regress the estimated BME of a specific country on the starting age for that particular country. We have the following model, for k different countries:

$$BME_k = \beta_0 + \beta_1 SSA_k + \epsilon_k \quad (3.3)$$

In other words, we estimate if there is a quadratic component of the effect that the starting age has on skill levels. The model is estimated on a country-level. The number of observations is therefore limited (9-19 observations per dataset). Nevertheless, we have 10 different datasets for which to estimate this model, providing 21 different coefficients in total. This allows us to estimate the quadratic component with a fair degree of precision.

3.4.3 Cost-benefit analysis

The first two models estimate the skill deficit for every school starting age, to the exact month. In order to draw economic conclusions, we need to compare this skill deficit with the time gain of starting earlier. As long as the academic deficit, which is expressed in terms of what is learned in a schooling year, is exceeded by the change in school starting age, there is a gain of being able to work earlier. We estimate whether this weighs up against the extra educational costs that are incurred when starting earlier. In the example in table 2.1, we assumed that an individual has 12 years of schooling. We expand this with 4 years of tertiary education in the actual analysis. We then have 16 years of education and thus 16 cost streams (without the cost of additional study time, I_x). Let us repeat the total profit function that specifies the net economic outcome for a specific individual:

$$\Pi = ((1+r)^\gamma - 1) \sum_{t=T_w}^{T_r} \frac{Y_t}{(1+r)^t} + \frac{\gamma Y_{T_r+1}}{(1+r)^{T_r}} - ((1+r)^{\Delta SSA} - 1) \sum_{t=T_{ss}}^{T_{se}} \frac{I_t}{(1+r)^t} - d \frac{I_{T_{se}+1}}{(1+r)^{T_{se}}}$$

The first part of our analysis allows us to estimate the skill deficit d for every ΔSSA and, correspondingly, every γ for every ΔSSA . The deficit δ is equal to the estimated Birth Month Effect multiplied by the change in school starting age; i.e. how much you lose from starting a month earlier, times how many months you actually start earlier. It includes both a linear (a) and a quadratic (b) component:

$$\delta = \Delta SSA(a + b\Delta SSA)$$

The net time gain γ is simply equal to the change in starting age minus this deficit:

$$\gamma = \Delta SSA - \Delta SSA(a + b\Delta SSA)$$

The above allows us to estimate the net effect of any change (in months) in SSA, since model 3.1 estimates our a and model 3.3 estimates our b . The above analysis also allows us to find the ‘ideal’ school starting age to the exact month.

By comparing the Birth Month Effect to the gain in time from starting earlier, we effectively correct for the fact that estimates of the Birth Month Effect pick up on both starting age effects and age at testing effects. Expressing the BME in terms of what is learned in a schooling year allows us to deduce the skill difference between young and old starters at the same exact age.¹

3.5 Results

3.5.1 Within-country analysis

This section presents and discusses the results of the analysis that has been specified above. We focus on overall results (which are summarized in table 3.2). An example of country-specific results (for the two PIRLS studies, which contain the largest set of countries) is given in table 3.1.² One can observe that the Birth Month Effect (BME) is very significant for all datasets.³ We also observe that adding controls hardly has any effect, both in terms of magnitude as well as significance. Therefore, we only explicitly discuss results without controls here. All three studies use the same standardization method of using a mean of 500 and a standard deviation of 100, thus scores can be viewed as comparable.⁴ The BME decreases somewhat as students get older. However, the coefficients only give us the effects on a specified test score. It would make more sense to look at what this means in relation to what is learned in one schooling year. The gain from one year of schooling differs by grade. It is approximately 60 points for 4th graders and 38 points for 8th graders.⁵ A difference in school starting age (SSA) of one whole year (multiply coefficients by 12) results in a performance difference between 0.45 and

¹There is still a difference between both individuals in excess of starting ages, namely that the younger starters have a longer educational life. However, we correct for this by taking this into account in the cost function.

²Country-specific results for all other datasets are available on request.

³To give an idea about the sensitivity to the correction method: if we add one standard error to the correction factor for both repeating and skipping, the overall coefficient increases, in absolute terms, in the range of 0.10-0.13 for all datasets.

⁴The only difference is that this standardization is based on a slightly different set of countries.

⁵Estimated from the TIMSS '95 datasets, which includes two adjacent grades.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

0.5 of a schooling year for both grades (about 28/60 vs. 18/38, see top panel of table 3.3). The BME thus stabilizes early, in terms of relative skill differences.¹ Moreover, this implies that the very youngest students in a birth cohort are about half a schooling year behind the very oldest students at the end of their academic career (assuming that stabilization continues). When we express magnitude in terms of a standard deviation in test scores, relative effects decrease from 4th to 8th grade (bottom panel of table 3.3). This only tells us that the effect is decreasing with respect to the variation in the total sample. It is more meaningful to express magnitude in terms of schooling years, from which we conclude that the BME is indeed stable.² Moreover, coefficients are close to stable in terms of standard deviations as well, when we compare 8th graders to 15-year-olds.

3.5.2 Between-country analysis

The BME constitutes the slope of the Skill Technology Function at different points. We now estimate the curvature of the function. Results from that analysis (model 3.3) are presented in table 3.4. There is a positive effect of school starting age (SSA) on the Birth Month Effect (BME). The effect becomes more negative when starting earlier. Coefficients are largely significant, even with only 9-19 observations per dataset. Thus, our estimated BME's follows closely the concave pattern of (the relevant range of) the STF. Coefficients in table 3.4 are per month. For a difference in SSA of a whole year, we multiply coefficients by 12, which gives an average coefficient of around -0.50. This means that, if we decrease SSA with a whole year, the BME would change from -1.50 to -2.00. This implies an increase from 1/2 of a schooling year to 2/3 of a schooling year, when starting a whole year earlier.

3.5.3 Cost-benefit analysis

The results from the previous section can be used to evaluate the costs and benefits of changes in school starting age. We choose a certain benchmark school starting age and measure whether gains are positive or negative for any other school starting age. In our study, we choose an SSA of 74 as a benchmark. This is the age of the average student in a large group of countries.³ We

¹We do not have a measure of this gain on the 15-year-old level, but assume that it is equal to the gain on the 8th grade level (largely 13-year-olds).

²We also estimated overall BME coefficients using a meta-analysis, where we assigned weights to countries according to within-country standard error. These results are very similar to the ones in table 3.2 (which takes each dataset as a whole), with a maximal deviation in coefficient of 0.12 and an average deviation of less than 0.04.

³The mode SSA is the range of 68-80 months, which is employed in Belgium, Canada, France, Iceland, Italy, Portugal and Spain.

continue the analysis specified in section 3.4.3. The Birth Month Effect stabilizes around a skill deficit (δ) of approximately half a schooling year. Since the youngest students gain a full year in time, they could compensate this deficit and still start to earn an income at an earlier age. The cost-benefit analysis estimates whether this weighs up to the additional educational costs. We have seen that the BME is dependent on the absolute starting age employed.¹ The gain in time is the change in SSA minus the deficit. We specified the following formula for γ in section 3.4.3 (with SSA measured in years):

$$\gamma = \Delta SSA - \Delta SSA(a + b\Delta SSA)$$

We mentioned that the linear component of the BME (a) lies around 0.5, where the quadratic component (b) is approximately equal to 1/6th. A precise calculation gives coefficients of 0.475 for the linear component and 0.179 for the quadratic component, expressed in terms of a whole schooling year. We can estimate γ for every ΔSSA with the following function:

$$\gamma = \Delta SSA - \Delta SSA(0.475 + 0.179\Delta SSA)$$

At the benchmark of 74 months, we assume the BME to be 0.475, which is the mean BME over all the datasets. An SSA of 74 months is approximately the median of all starting ages and, moreover, the average BME of all countries with a range of 68-80 months also centers around a coefficient of -1.50 (which corresponds to 0.475 of a schooling year). Therefore it is plausible to assume the BME has this value at exactly this starting age. We assume an interest rate (r) of 5%. Cost and income streams are Euro-Area averages. Since these data are from a particular year in time (2002), while the income streams are incurred in different time periods, we also incorporated a real GDP growth rate for the cost and income streams. This assumed growth rate is 2.5%.

Figure 3.4 shows the outcome for the cost-benefit analysis. Any positive profit corresponds to an improvement compared to the generally accepted SSA of 74 months. Reducing the school starting age by a whole year (SSA=62) is a definitive improvement. Profits are only negative for an SSA of 55 and earlier. The gain in time (which is maximized at an SSA of 56 months) does not weigh up to the increased educational costs for these starting ages. We can see that profits

¹We employ the SSA of the average student as our value for SSA. An SSA of 74 thus corresponds to a range for the whole cohort of 68-80 months. The overall result corresponds to the economic payoff of the average student for that specific range.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

are maximized at an SSA of 65 months. In this case, the whole cohort ranges from 59 to 71 months. The large majority of children will be 5 years old when they start to learn how to read and to write.

We address the sensitivity of these results, in order to establish realistic lower and upper bounds for our estimates. We can do this by adding or subtracting a standard error to either the BME or to the quadratic component. Results of these alterations are portrayed in table 3.5. The sensitivity is small if we alter the BME but somewhat larger for the quadratic component. This is intuitive since the latter is estimated for a country sample of 9-19 observations and therefore has a relatively larger standard error. Especially in the case where we subtract a standard error, the change in optimum is relatively large. In that case, the quadratic component is close to zero and therefore too low to keep the time surplus from rising. Hence, this is a very low bound. Estimation below the 60 months mark tends to be less robust, since we have fewer estimated slopes in that area. Estimates on the range above 60 months are very precise, also because the very robust linear component of the BME is the main determinant of our optimum for ages close to the benchmark.

We can also add sensitivity to our income and educational cost data. Table 3.5 shows that changing the length of the educational career, the size of income streams or the size of the real GDP growth rate does not change our optimum by much. The *height* of the profit function can be affected more strongly by these parameters (as can be seen in the last column), but the starting age at which this optimum takes place depends largely on the BME itself. The skill optimum is not affected by changes in income and cost data, since it fully depends on γ .

3.6 Robustness

3.6.1 Peer effects

Peer effects could affect estimates of the Birth Month Effect (BME). Differences in relative maturity within a class can have psychological effects on individuals. Younger individuals can be induced to work harder to catch up with their older peers, or they could become demotivated because they are lagging behind. We perform a robustness test, by comparing self-confidence values across individuals. The existence of differences in confidence levels between older and younger individuals could be an indication of the psychological influence of other-aged peers. We conducted a factor analysis, combining the answers to several questions with respect to how comfortable children feel at school into an aggregate measure for self-confidence. Our measures

show that confidence is somewhat higher for older students, but this effect is only very small and largely insignificant on a country-level (table 3.6), and is especially low at the 8th grade level. If confidence effects are a big factor in our measurement of the BME, we would expect a lower relative BME for 8th graders, which is not the case. There is no difference between early and late starters in the degree to which they like being in school, although early entrants report more bullying and have less friends. Additionally, Fredriksson and Ockert (2005) consider the role of peer effects in their study as well, and they conclude that relative starting age effects (which are a measure for peer effects in this case) are strongly dominated by absolute starting age effects. This is estimated by instrumenting the rank in the age distribution in class with the rank in the starting age distribution. Relative age effects are only significant for sports, but are neglectable for all academic subjects. Fredriksson and Ockert (2005) therefore conclude that the relative rank in the age distribution has no noticeable influence on performance, which implies that peer effects are weak or absent.¹ In addition, Elder and Lubotsky (2009) find slightly higher relative age effects, but they are still only a small fraction of absolute starting age effects, and the estimates are largely insignificant.

We also examine whether early and late starters differ in measures of non-cognitive skills. The data contain several measures that are closely related to the Big 5 Trait of Extraversion. Younger entrants are *more* likely to be members of a club, although the effect is just short of being significant. They are significantly less likely to be active in sports. We constructed a latent factor for Agreeableness which measures enjoyment of working with others, working in groups and helping other people. No significant effect of month of birth is found.² Similarly, we constructed a factor variable for Neuroticism, which combines self-reported questions on whether students feel nervous in school, worry a lot, feel helpless or feel tense.³ Again, we find a favorable but insignificant impact of starting school earlier.

The above does not present any conclusive picture for the effect of starting ages on psychological or non-cognitive factors. Disentangling absolute from relative age effects is cumbersome. We cannot completely rule out any impact from relative age in the class, but the evidence from Fredriksson and Ockert (2005), Elder and Lubotsky (2009) and our own analysis suggests that such effects are modest at best.

¹Or the possible negative effect of being the youngest in the class cancels out with the positive effect of wanting to catch up, on average.

²Younger entrants have higher values but the effect is insignificant.

³All these questions are stated with respect to the topic of math, which was the core topic of the PISA 2003 wave.

3.6.2 Tracking

Another aspect that might influence the Birth Month Effect is tracking. The fact that relatively younger students obtain lower test scores than their older peers means that these younger pupils have a higher chance of being selected into a lower track. Differences in investment between those tracks can further increase differences in performance between these two groups of students, thereby inflating our estimates of the Birth Month Effects. We have seen that the BME is stable from an early age on. Thus, it does not seem to be the case that there is an increase in the BME after tracking has taken place. It could still be that tracking prevents a decrease of the effect over the educational career. There are several reasons why we believe this is not the case. There are multiple countries in our sample that only start formal selection after the age of 15 (including Canada, Iceland, Sweden and New Zealand). There is no decrease over age for the Birth Month Effect for these countries. Furthermore, at the 8th grade level, only one third of all countries had selected their students, of which only three for more than one year. This can hardly explain the stability that we find in the BME from 4th to 8th grade. Stability between these two age groups is also strong for those countries that only started tracking after the 8th grade level. The only evidence we find of an effect of tracking on the BME is for those countries that tracked very early. For these countries, the PISA estimates are slightly larger than the estimates at the two lower age levels. Thus, it could be the case that, several years after tracking has taken place, the BME is slightly higher because of this. This does not invalidate our assumption of stability. The evidence even suggests that stability would be stronger in the absence of tracking.

3.6.3 Inequality

We examine the effect that school starting ages have on the variance of the score variables. We do this by first estimating the variance within each month, within each country and then take the by-country-average of these variances. In this way, one excludes any variance that occurs from the BME itself. We then estimate the effect of the official SSA on this specified variance. Results are listed in table 3.7. There is a negative and significant effect between starting ages and inequality. Earlier school starting ages increase the variance in test scores, thereby leading to more inequality among students. Any economic advantages that earlier school starting ages may bring, can come at the expense of increased inequality among students. A decrease of a whole year in SSA increases inequality by around 0.05 to 0.15 of a standard deviation in test score.

3.6.4 Month of birth

Recent papers argue that month of birth is not independent of social class.¹ We checked the possibility of such a correlation, using number of books at home as an indicator for social class. This estimation is executed for our most extensive dataset, the PIRLS 2006 study. We find weak evidence for a relation between month of birth and social class. Children from lower social classes are slightly more likely to be born later in the year. On a country level, this effect only passed the 5% significance mark for 5 out of 20 countries. Moreover, the effect is very small. Moving from the lowest social class to the highest,² decreases the predicted month of birth by 0.2 (which effectively is 6 days earlier). It is therefore plausible to say that this will not have strong effects on our estimates. Additionally, there does not seem to be a consistent pattern of birth months to socioeconomic status (SES). Where Buckles and Hungerman (2008) find lower SES for children born in winter compared to summer-borns, we find SES to be average for both groups in the U.S. sample. Moreover, any such correlation between month of birth and social class would lead to an *overestimation* of the Birth Month Effect (since scores of younger children are underestimated), which would point even stronger to earlier starting ages (be it very weakly).

3.7 Conclusion

This study uses variation in month of birth to examine the effect of different school starting ages on school performance. We show that these results can be applied to estimate the costs and benefits of changes in school starting ages. Students that start school at an earlier age perform worse than their older peers, but this negative effect can be compensated by the positive effect of finishing school earlier. These positive effects come in the form of earlier incurred income streams and a longer working life. We use data from three different studies, covering 25 countries, three age groups and three school disciplines to measure our Birth Month Effect (BME). The BME is significant across ages and across disciplines and stabilizes early. By conducting a between-country analysis, we also find that the effect increases when starting ages are lower. We also take into account that grade-samples, as opposed to birth-cohort-samples, lead to biased estimates. We develop an IV-method to correct for this.

¹See e.g. Buckles and Hungerman (2008).

²Based on number of books, meaning 0-10 books versus more than 200 books.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

Taking all this into account, we find that the benefits of starting school earlier (meaning earlier than 74 months, which is currently the most frequently used starting age for the average student) weigh up against the increases in educational costs, up until a school starting age (SSA) of 55 months. Profits are at their maximum at an SSA of 65 months. Thus, our findings suggest an earlier starting age than the current norm, opposing the current trend of advocating later school starting ages. These results are, to an extent, rough findings, since they are sensitive to changes in variable inputs. However, further analysis shows that the degree of sensitivity to changes in any parameter is low, and results consistently point to earlier starting ages than the current norm. Moreover, the analysis shows that it is possible to weigh costs and benefits of changes in SSA against each other using variation in month of birth. Thereby, it is crucial that we incorporate both the human capital theory argument as the endowment theory argument in a proper way. The benefits of earlier school starting ages in the form of gains in time have been largely neglected in previous research and policy analysis.

One could argue that the concepts of tracking, peer effects and possible endogeneity of month of birth bias our estimates. We show that these effects are either very small or absent. Moreover, even if any such bias is present, it would point in the direction of even earlier starting ages than the ones we find.

Nonetheless, more elaborate analysis on the non-linear component of the BME should be performed, in order to get a more complete picture of the slope of the STF at every possible point. Still, slight imprecision of this non-linear component would only have noticeable implications for the estimation of very large changes in SSA. Changes up to at least one year from the current norm are very precise. The developed method is therefore very suited for country-specific analysis and allows each country, by linking its estimated Birth Month Effect to its economic parameters, to estimate the economic implications of changes in its currently employed school starting age.

3.8 Figures and tables

Figure 3.1: Birth cohort vs. grade: no correction

Notes: The graph shows the correlation between the estimated Birth Month Effect from the grade dataset without correction and the birth cohort dataset, for 8th grade math (TIMSS '95). The birth cohort dataset contains students who are born within one year after the cutoff date. The grade dataset contains all students who are in the same grade.

Figure 3.2: Birth cohort vs. grade: with correction

Notes: The graph shows the correlation between the estimated Birth Month Effect from the grade dataset with correction and the birth cohort dataset, for 8th grade math (TIMSS '95). The birth cohort dataset contains students who are born within one year after the cutoff date. The grade dataset contains all students who are in the same grade. For the grade dataset, scores for students who are from a different birth cohort (repeaters, skippers and early and late starters) are corrected using model 3.2.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

Figure 3.3: Mean reading scores

Notes: Graph shows mean Reading scores over month of birth (PIRLS: 4th grade). Month of birth is defined as the 12 months after the cutoff date. Hence, a higher month of birth means a lower starting age.

Figure 3.4: Cost-benefit analysis

Notes: Graph shows the expected payoff for the average-aged student, compared to the mode starting age of 74 months. Benefits relate to reaching a certain skill level at an earlier age (and thereby entering the labor market earlier). Costs exist of extra educational expenses. We assume an interest rate of 5%, years of schooling of 16 years and a real GDP growth rate of 2.5%. Income and educational cost data are Euro-area averages (2002, in Euro PPP)

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

Table 3.1: The Birth Month Effect (BME) for 4th grade reading

	2006	2006	2001	2001		2006	2006	2001	2001
Scotland	-3.69*** (0.365)	-3.64*** (0.358)	-2.37*** (0.447)	-2.31*** (0.441)	Netherlands	-2.16*** (0.237)	-2.18*** (0.230)	-1.72*** (0.254)	-1.76*** (0.244)
New Zealand	-2.09*** (0.324)	-2.04*** (0.314)	-2.61*** (0.529)	-2.47*** (0.501)	Austria	-1.91*** (0.258)	-1.90*** (0.246)	-	-
England	-3.09*** (0.372)	-2.89*** (0.361)	-2.79*** (0.422)	-2.58*** (0.398)	Czech Republic	-	-	-2.10*** (0.334)	-2.19*** (0.322)
Greece	-	-	-	-	Luxembourg	-1.52*** (0.277)	-1.46*** (0.257)	-	-
Cyprus	-	-	-	-	Slovakia	-1.93*** (0.290)	-1.91*** (0.280)	-2.61*** (0.320)	-2.61*** (0.307)
Belgium:	-1.98*** (0.244)	-1.94*** (0.234)	-	-	Germany	-2.80*** (0.215)	-2.84*** (0.210)	-2.56*** (0.230)	-2.54*** (0.215)
Flanders					Hungary	-2.98*** (0.309)	-3.02*** (0.303)	-2.20*** (0.283)	-2.10*** (0.271)
Belgium:	-2.62*** (0.302)	-2.58*** (0.297)	-	-	Japan	-	-	-	-
Wallonia					Denmark	-2.13*** (0.314)	-2.14*** (0.307)	-	-
Canada	-2.43*** (0.179)	-2.44*** (0.178)	-2.73*** (0.348)	-2.59*** (0.320)	Finland	-	-	-	-
France	-2.37*** (0.307)	-2.29*** (0.301)	-1.87*** (0.355)	-1.85*** (0.346)	Poland	-1.95*** (0.310)	-1.92*** (0.307)	-	-
Iceland	-2.55*** (0.310)	-2.55*** (0.304)	-2.51*** (0.346)	-2.19*** (0.337)	Sweden	-1.91*** (0.271)	-1.83*** (0.262)	-1.84*** (0.237)	-1.69*** (0.223)
Italy	-2.54*** (0.316)	-2.58*** (0.311)	-1.73*** (0.339)	-1.78*** (0.335)	All	-2.39*** (0.067)	-2.37*** (0.066)	-2.24*** (0.090)	-2.17*** (0.086)
Portugal	-	-	-	-					
Spain	-2.28*** (0.304)	-2.26*** (0.300)	-	-					
Controls	Yes		Yes			Yes		Yes	

Notes:*Significant at 10% level **Significant at 5% level ***Significant at 1% level

Coefficients estimate the effect of month of birth on score, both without and with including controls (wc), according to model 3.1. Scores are standardized with mean 500 and standard deviation 100, based on all countries included in the study (this is somewhat lower for our datasets, ranging from 70 to 100, since our country group is more equal). Birth months are numbered as the 12 months after the cutoff date. Scores for early and late starters, repeaters and skippers are adjusted. The applied correction factors for 4th grade reading are 30.1 for repeating and 50.0 for skipping. Source: PIRLS.

Table 3.2: Overall coefficients of birth month on score

	4th grade Reading	4th grade Math	4th grade Science	8th grade Math	8th grade Science	Age 15 Reading	Age 15 Math	Age 15 Science
PIRLS 2006	-2.39*** (0.067)	-	-	-	-	-	-	-
PIRLS 2001	-2.24*** (0.090)	-	-	-	-	-	-	-
TIMSS 2003	-	-2.33*** (0.120)	-2.43*** (0.115)	-1.54*** (0.109)	-1.53*** (0.103)	-	-	-
TIMSS 1999	-	-	-	-1.42*** (0.110)	-1.54*** (0.107)	-	-	-
TIMSS 1995	-	-2.56*** (0.136)	-2.55*** (0.131)	-1.52*** (0.118)	-1.60*** (0.110)	-	-	-
PISA 2006	-	-	-	-	-	-1.38*** (0.078)	-1.50*** (0.072)	-1.63*** (0.076)
PISA 2003	-	-	-	-	-	-1.14*** (0.082)	-1.35*** (0.083)	-1.40*** (0.088)
PISA 2000	-	-	-	-	-	-1.67*** (0.142)	-1.25*** (0.130)	-1.83*** (0.136)

Notes: *Significant at 10% level **Significant at 5% level ***Significant at 1% level

Coefficients estimate the effect of month of birth on score for each dataset as a whole. Scores are standardized with a mean of 500 and a standard deviation of 100.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

Table 3.3: Relative magnitude of Birth Month Effect

	4th grade Reading	4th grade Math	4th grade Science	8th grade Math	8th grade Science	Age 15 Reading	Age 15 Math	Age 15 Science
<i>In Schooling Years</i>								
PIRLS 2006	0.478	-	-	-	-	-	-	-
PIRLS 2001	0.448	-	-	-	-	-	-	-
TIMSS 2003	-	0.466	0.486	0.486	0.483	-	-	-
TIMSS 1999	-	-	-	0.448	0.486	-	-	-
TIMSS 1995	-	0.512	0.510	0.480	0.505	-	-	-
PISA 2006	-	-	-	-	-	0.436	0.474	0.515
PISA 2003	-	-	-	-	-	0.360	0.426	0.442
PISA 2000	-	-	-	-	-	0.527	0.395	0.578
<i>In Standard Deviations</i>								
PIRLS 2006	0.406	-	-	-	-	-	-	-
PIRLS 2001	0.366	-	-	-	-	-	-	-
TIMSS 2003	-	0.346	0.361	0.240	0.242	-	-	-
TIMSS 1999	-	-	-	0.209	0.229	-	-	-
TIMSS 1995	-	0.350	0.333	0.215	0.223	-	-	-
PISA 2006	-	-	-	-	-	0.183	0.187	0.218
PISA 2003	-	-	-	-	-	0.154	0.172	0.174
PISA 2000	-	-	-	-	-	0.229	0.162	0.230

Notes: The table shows the predicted difference in skills between the very oldest and the very youngest individual in class. It is the (average) Birth Month Effect multiplied by 12; the predicted difference in score for a change in starting age of a whole year. In the top panel, this is divided by the gain of a whole schooling year, which is 60 for the 4th grade level and 38 for the 8th grade level. We thereby assume that the gain on the 15-year-old level (largely 2 years later than 8th grade) is also 38 for all school topics and the gain for 4th-grade reading (PIRLS) is also 60. In the bottom panel, the Birth Month Effect is multiplied by 12 and divided by the standard deviation of the corresponding test score.

Table 3.4: Effect of school starting age on Birth Month Effect

	4th grade Reading	4th grade Math	4th grade Science	8th grade Math	8th grade Science	Age 15 Reading	Age 15 Math	Age 15 Science
PIRLS 2006	0.038** (0.016)	-	-	-	-	-	-	-
PIRLS 2001	0.022 (0.015)	-	-	-	-	-	-	-
TIMSS 2003	-	0.057** (0.024)	0.050* (0.023)	0.031* (0.017)	0.024 (0.015)	-	-	-
TIMSS 1999	-	-	-	0.036 (0.026)	0.048* (0.023)	-	-	-
TIMSS 1995	-	0.069*** (0.020)	0.081*** (0.017)	0.025* (0.014)	0.037** (0.011)	-	-	-
PISA 2006	-	-	-	-		0.089*** (0.018)	0.055** (0.022)	0.047** (0.019)
PISA 2003	-	-	-	-		0.007 (0.024)	-0.021 (0.020)	-0.005 (0.017)
PISA 2000	-	-	-	-		0.054** (0.020)	0.040 (0.024)	0.039 (0.025)

Notes: *Significant at 10% level **Significant at 5% level ***Significant at 1% level

Executed on a country level (model 3.2). The table shows the effect of the formal school starting age (for the median student in the country) on the country-wide Birth Month Effect, according to model 3.3. Starting ages are measured in months. In order to get the implications for a change in starting age of a whole year in terms of a schooling year, one has to multiply by 12 and divide by the gain in points from a whole schooling year (60 for 4th grade and 38 for 8th grade and 15-year-olds). This gives, on average, a measure for the non-linear component of 0.179.

3. AN ECONOMIC ANALYSIS OF THE OPTIMAL SCHOOL STARTING AGE

Table 3.5: Sensitivity of cost-benefit analysis

	Skill Optimum (in months)	Profit Optimum (in months)	Economic payoff (in euro PPP)
Estimated values	56	65	4467,04
BME + 1 s.e.	58	66	3435,77
BME - 1 s.e.	55	63	5649,63
Quadratic component + 1 s.e.	61	67	3204,61
Quadratic component - 1 s.e.	<50	58	7417,58
Upper extreme	62	68	2464,64
Lower extreme	<50	56	9373,03
Add 1 year of Schooling	56	65	4055,55
Add 2 years of Schooling	56	65	3653,86
Increase income streams by 10%	56	64	5536,86
Decrease income streams by 10%	56	65	3465,14
Real GDP growth rate of 4%	56	62	13034,76
Real GDP growth rate of 1%	56	68	1095,33

Notes: The table shows the starting age in months for which either skills or profits are maximized for the median-aged student. The skill optimum only looks at revenues. It is the age at which we reach a specific skill level at the earliest age. The profit optimum also takes increased educational costs into account. The economic payoff is the expected gain for the median-aged student under this new starting age. S.e. stands for standard error. The first estimate refers to the case where the BME is 0.475, the quadratic component 0.179, we have 16 years of schooling and a real GDP growth rate of 2.5%. The upper extreme refers to the case where a standard error is added to both the Quadratic component and the BME. The lower extreme is where a standard error is subtracted from both.

Table 3.6: Effect of month of birth on confidence

	4th grade level	Significance	8th grade level	Significance
PIRLS 2006	-0.0053*** (0.00095)	5 out of 19	-	-
PIRLS 2001	-0.0035** (0.0014)	1 out of 11	-	-
TIMSS 2003	-0.0083*** (0.0016)	4 out of 9	-0.0026* (0.0014)	0 out of 12
TIMSS 1999	-	-	-0.0033** (0.0015)	1 out of 10
TIMSS 1995	-0.0086*** (0.0018)	3 out of 7	-0.0061 (0.0015)	2 out of 11

Notes: *Significant at 10% level **Significant at 5% level ***Significant at 1% level

Confidence is a latent factor constructed from multiple self-report variables on how comfortable students are with school, teachers, peers etc. Values are standardized with a mean of 0 and a standard deviation of 1. Significance measures for how many countries in the particular dataset the country-specific estimate passes the 5% significance mark.

Table 3.7: Effect of school starting age on inequality

	4th grade Reading	4th grade Math	4th grade Science	8th grade Math	8th grade Science	Age 15 Reading	Age 15 Math	Age 15 Science
PIRLS 06	-0.128*** (0.040)	-	-	-	-	-	-	-
PIRLS 01	-0.166*** (0.037)	-	-	-	-	-	-	-
TIMSS 03	-	-0.070 (0.052)	-0.066 (0.071)	0.010 (0.040)	-0.016 (0.036)	-	-	-
TIMSS 99	-	-	-	-0.083 (0.055)	-0.107* (0.052)	-	-	-
TIMSS 95	-	-0.052 (0.031)	-0.131*** (0.042)	0.011 (0.039)	-0.067* (0.036)	-	-	-
PISA 2006	-	-	-	-	-	-0.085** (0.032)	-0.032 (0.023)	-0.072** (0.027)
PISA 2003	-	-	-	-	-	-0.084** (0.032)	-0.021 (0.032)	-0.012 (0.029)
PISA 2000	-	-	-	-	-	-0.068 (0.044)	-0.061 (0.039)	-0.031 (0.041)

Notes: *Significant at 10% level **Significant at 5% level ***Significant at 1% level

The estimates measure the effect of a change in starting age of a whole schooling year on the standard deviation in test scores. A coefficient of -0.166 means that if we reduce the starting age by a whole year, the standard deviation in scores (which measures inequality) will go up by 0.166 of a standard deviation. Our method excludes variation that is caused by the Birth Month Effect itself.

Allocating Instruction Time: How Language Instruction Can Affect Multiple Skills¹

4.1 Introduction

What is the value of time spent on education in a certain school subject? The immediate knowledge acquired in that specific subject is an obvious gain, but the long run value is highly dependent on the degree of complementarity between skills. Moreover, what is learned in one subject could benefit subsequent learning in that same subject and in other subjects. For example, well-developed listening and reading skills can make one more adept at learning and understanding a subject such as Geography or History. If one wants to assess the value of instruction in a specific subject, all these effects need to be taken into account.

The aim of this study is to analyze the direct and indirect effects of changes in the instruction time for language on student achievement. We use a skill technology function, where the acquisition of skills depends on incumbent skills and the allocation of time to all school subjects. Three dimensions of the effect of instruction are crucial: direct effects on the specific skill that is being trained, the long-run development of initial effects of instruction and effects on seemingly unrelated skills. Both short- and long-term effects on language performance are estimated, as well as effects on mathematics, environmental studies and non-verbal IQ. In our analysis, we exploit variation in instruction time (IT) over time, caused by policy interventions in the Dutch

¹This chapter is based on Borghans and Diris (2010a).

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

educational system. We find that there is a strong positive relationship between instruction time and performance. However, this effect dissipates in the long run. We find strong spillovers towards other areas, including non-verbal IQ measures and performance in mathematics. For the latter, we find long-term effects to be dominating. The true value of (increased) language instruction therefore lies in spillover effects into other type of skills rather than pure language development.

We use data from the Dutch PRIMA study, which is a bi-annual longitudinal study of Dutch primary schools. PRIMA was executed from the academic year 1994/'95 until the year 2004/'05 and contains six waves in total. This specific dataset provides us with the means to estimate direct and indirect effects of changes in language instruction, because it gives us data for multiple years, multiple grades and performance in multiple subjects. The study reports time spent in class on language instruction per week, which is employed as our main independent variable. Results are measured at the 2nd, 4th and 6th grade level.

Previous research in economics on the effect of time spent in school has largely focused on the duration of educational life. The first study that tried to identify the effect of education on earnings was by Mincer (1974). There have been many studies since then that have measured the returns to schooling.¹ Among those studies, Angrist and Krueger (1991) and Card (1993) were the first to tackle the apparent problem of endogeneity when measuring the return to education. There are only a few papers in the field of economics² that deal with the actual content of education. Such studies are more prominent among educationalists.³ However, these studies focus on short-term direct effects instead of taking the perspective of long-term skill acquisition. Moreover, they tend to make use of cross-sectional analyses, which questions the causality of their estimates.

Although this topic has largely been the domain of educationalists, decisions on the content of education are economic decisions. Schools have to make tradeoffs with regard to which type of instruction they deem most valuable. Each subject has a specific payoff on one or several skills. From an economic perspective, increases in instruction in a specific subject are justified if the marginal effect on a student's skill set outweighs that of instruction in all other subjects. It is crucial that the measure of this marginal effect also incorporates long-run development of initial effects and effects on other skills. Since skills acquired in a certain period are dependent

¹See Psacharopoulos (1994), Card (1999) and Harmon et al. (2003) for overviews.

²E.g. Angrist and Lavy (2002); Machin and McNally (2008).

³See reviews by Scheerens (2000) and Sammons (1999).

on skills already present before that period, relatively small initial effects on skill acquisition can have strong implications in the long run, depending on the degree of complementarity within that specific subject. Moreover, if complementarity between subjects is strong, skill acquisition in a certain subject can benefit from proficiency in seemingly unrelated skills. Gaining better insight into the added values of different school subjects is vital in achieving an efficient allocation of instruction time.

The organization of this chapter is as follows. Section 4.2 gives a brief overview of relevant characteristics of the Dutch educational (primary school) system. Section 4.3 presents a literature overview, in which we discuss previous research in this field. Section 4.4 discusses the PRIMA data sample that we use for our analysis. Section 4.5 presents the methodology, where section 4.6 discusses our main results. Further robustness analysis is discussed in section 4.7. Section 4.8 concludes.

4.2 Dutch educational system

We use variation in instruction time (IT) over time to estimate its effect on school performance. Changes in educational guidelines in the middle of the 1990's in the Netherlands caused a jump in the allocated instruction time for language. This section discusses the background and implications of this policy change.

Schools in Dutch primary education have a lot of freedom in designing the weekly curriculum for their students. There are no strict demands with respect to the time spent on certain subjects, but there are demands on what pupils should be able to master at the end of the year. These demands are specified in the so-called Core Targets (Dutch: *Kerndoelen*), which specify the skills students should have for every grade and every subject. Schools are evaluated on this. The Targets have a guiding function towards the setup of the weekly curriculum, albeit indirect. In the beginning of the 1990's, these Targets expanded horizontally, meaning that a lot of new topics and fields were being brought into the list of demands. These new topics were largely related to the building of citizenship.¹ In this same period, the Dutch government assigned an independent commission of educational experts to make an elaborate investigation of the Dutch primary school system. The final report from this *Commissie Evaluatie Basisonderwijs* (CEB) was very critical of Dutch primary schools, arguing that the expansion of Targets has taken too

¹How children should behave in social life or in traffic, that they should care about the environment etc.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

much attention away from core activities, leading to a decrease in especially language performance (Commissie Evaluatie Basisonderwijs, 1994). A separate evaluation of the 1992 Targets in the academic year '97/'98 reached similar conclusions after which both these advices were translated into an effective change in policy, namely the formation of new Core Targets in 1998. These new Targets were more demanding towards language performance, while the chapter relating to citizenship was cut back and made more voluntary.

Schools were not given strict orders to increase the quantity of instruction time, but these new guidelines have removed schools and teachers from obligations with respect to 'secondary activities', which has freed up instruction time that should have been largely directed towards language, given the emphasis in the new Targets on especially this subject.¹ Not all schools might have picked up on this policy change to the same degree, but it is expected that this has raised the time spent on language for the average school,² which is what we need for our identification (provided the increase is sufficiently strong).

4.3 Literature

4.3.1 Instruction time

There is a large body of research measuring the effects of instruction on student performance. Studies have focused both on the effects of general (effective) time spent in class as well as more qualitative aspects of instruction time, such as specific methods of instruction. Research in this area is rare in the field of economics, and has almost exclusively been conducted by educationalists.

The earliest of these papers often focused on the length of the school day or school year. They find that there is no significant relationship with academic achievement (Harnischfeger and Wiley (1976), Stallings et al. (1979)). These findings have led to a general belief that pure quantity changes in instruction time are not that worthwhile and that it is not the quantity of instruction time that is crucial, but rather how that time is being used (Stallings et al., 1979). Other studies distinguish between different types of instruction (individual or group, written or oral, involved or engaged, etc.) and analyze which types are most effective (Powell, 1978; Stallings, 1980). Some papers focus on the effects of more or less time in specific subjects on

¹The fact that the Commission also criticized, although to a lesser extent, math performance, was hardly noticeable in the effective new targets. The political focus really has been on language instruction.

²In other words, it shifts the distribution of time spent on language to the right.

academic achievement. Results are very inconsistent both across and within studies. Many find a positive and significant relationship, although there are inconsistencies across subjects or grades (Daniels and Haller, 1981; Jacobson, 1980). Analysis on some specific samples even provide negative correlations (Bell and Davidson, 1976). Multiple studies show that opportunity to learn or content exposure is the biggest predictor for test scores of all ‘classroom information’ (Cooley and Leinhardt, 1980; Wang, 1998).

Kiesling (1984) finds that (individual) instruction on vocabulary has a negative marginal return beyond a certain threshold. A similar finding is made by Soar (1978), who states that for all students there is a point at which more instructional time is not productive anymore. Stallings et al. (1975) also find that time spent in more advanced activities is positively associated with scores on non-verbal problem-solving tests. This suggests that there might be a direct relationship between instruction in certain subjects and IQ.

Although the effect of instruction time on student achievement has been widely investigated, there has been large variation in dealing with selection bias. Selection is expected to put a downward bias on estimated coefficients, since schools with poorly performing students tend to spend more time on core subjects, i.e. language and mathematics. Most papers recognize this, but there has not been a clear and consistent solution for this problem. For example, Kiesling (1984) use pre-test scores or interactions of pre-test scores with instruction time as proxies for ability. Although this has led to positive and significant findings, these estimates are expected to still underestimate the effects of IT on performance, since one cannot control for unobservables. There are more recent studies that use an experimental setup. Machin and McNally (2008) evaluate the use of the National Literacy Project (NLP), which involved more focused literacy instruction and effective classroom management. They compare schools that used NLP with schools that did not, in geographically adjacent areas. The program increases scores on reading and English tests, which is sustained until the end of primary education. The magnitude of the increase is around 0.1 of a standard deviation and is larger for boys. NLP presumably replaced existing literacy instruction and, as such, did not take time away from any other subjects.

Machin and McNally (2008)’s study is quasi-experimental. We take a different statistical approach by employing IV. Moreover, the above mentioned studies focus on one-to-one correlations between IT in a specific grade, either scheduled or effective, and test scores. We take a broader approach by also examining whether long-term effects are present, whether the effect differs for different groups and whether there are spillovers of instruction in one specific subject into performances in other subjects.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

4.3.2 IQ literature

We also examine if language instruction can have a positive influence on non-verbal IQ measures. This would require that IQ can still change for the age group under examination (being 7-12 year olds) and, more importantly, that this change can be influenced. There is still no consensus in the literature on either of these two puzzles. We examine current literature concerning this possible malleability of fluid intelligence.

Many studies have examined the stability of (fluid) intelligence over a person's lifetime. This has been done mainly by looking at correlations of IQ tests at different ages. Rank order correlations are high after the age of 6, in the range of 0.75-0.90 (Almlund et al., 2011). This suggests that genetics and early environment are strong determinants of fluid intelligence, but that there still can be (relatively small) changes after early childhood. For our study, the relevant question is whether (non-verbal) IQ scores can be influenced after the age of 7 or 8. Pearce et al. (2005) show that (physical) growth had an effect on cognition beyond the age of 9 years. They find that height at 13 remained significant after adjustment for earlier height. McArdle et al. (2002) execute an elaborate analysis on growth patterns of different types of intelligence. They conclude that fluid intelligence continues to increase into early adulthood and reaches a peak at 23 years of age.

Even if IQ can still change at later ages, this does not immediately imply that it can be influenced by policy. Changes could be caused by biological programming (called ontogenic change) rather than intentional change. McCrae and colleagues argue¹ that changes in IQ reflect genetically programmed scripts. In that case, investment aimed at improving traits would be futile.

A study by Jaeggi et al. (2008) shows that a task designed for the training of working memory improved performance on a seemingly unrelated task. The training task involved matching letters with images, which were portrayed to the subjects before. The subjects that took part in this memory task performed significantly better on visual analogy tests. These results seem to indicate that it is possible to train our ability to solve cognitive problems, at least in the short run.

It appears that schooling is one of those factors that can influence IQ. Research has shown that schooling changes certain mental abilities, thereby enhancing children's fluid intelligence

¹(see Costa and McCrae, 1994; Costa et al., 1999; McCrae and Costa, 1990, 1994, 1996, 2003).

(Ceci, 2001). Brinch and Galloway (2012) find that an educational reform in Norway that increased the compulsory schooling age led to higher measures of IQ at age 19. It is still largely uncertain what aspects of schooling are especially critical in this respect.

There is a very large range of preschool interventions that had short- and also long-run impacts on IQ. The Perry Preschool Program led to strong short-run impacts in IQ, of up to 12 points, but these disappeared by age 8.¹ Similar patterns are visible for many other preschool and school-age interventions.² The majority provides short-term boosts in IQ, but they do not sustain. The only exceptions appear to be interventions that already start in infancy, such as the Abecedarian Project (Heckman et al., 2010), or the Nurse-Family Partnership (Olds et al., 2004), who still present positive impacts on IQ at later ages. To summarize, policies and interventions that change IQ in the short run are present over a large range of ages, but sustained improvements have only been made by those that start before the age of four.

4.4 Data

4.4.1 The PRIMA-dataset

We use data from the Dutch PRIMA study, which assessed primary school students in the Netherlands. PRIMA consists of tests and background questions and was executed every two years, from 1994/1995 to 2004/2005. Information is collected for students in kindergarten and 2nd grade, 4th grade and 6th grade of primary education. Tests are conducted on language, mathematics, reading (from 1999 on, only for 4th and 6th grade) and non-verbal IQ. The IQ test is a strictly non-verbal test, consisting of two parts: combining (putting two figures together to form a pre-specified shape) and exclusion (selecting which figure is different from the other three).

Questions for language and IQ tests are identical in every wave. This is not the case for the math test, which contains different questions and a different number of questions per wave, which makes scores incomparable.³ Since the policy change takes place in between waves 2 and 3, it is crucial that these waves are comparable. The test also reports results for the Dutch Cito test. This test is executed at the end of elementary schools and determines to a strong extent into which track students are selected at the secondary school level. The data report an overall score

¹See e.g. Anderson (2007), Heckman et al. (2010).

²See overviews by Barnett (1995); Blau and Currie (2006); Currie (2001); Karoly et al. (2005); Yoshikawa (1995).

³Both raw and calibrated scores are incomparable, since a different calibration method is used after wave 2.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

and scores per subject,¹ in the form of number of correctly answered questions. Although the questions are not identical for every year, they are selected to have the same difficulty level on average. We use these test results to estimate the effects of language instruction on performance in mathematics and environmental studies.² Estimation is only possible at the 6th grade level, since the test is only conducted there.³ To sum up, we have comparable test scores for language and IQ in grades 2, 4 and 6 and for mathematics and environmental studies in grade 6.

Since there is no formal language instruction in kindergarten, we only look at data from 2nd, 4th and 6th grade. PRIMA also contains background questionnaires, answered by pupils, their parents, teachers and schools. Teachers are asked how much time they spend per week on each school subject.⁴ For the first two studies, this only applies to language, reading and math, but it is expanded with other subjects in later waves. Teachers can answer the question to the exact minute. For interpretative purposes, we rescaled these values to hours. Thus, the values of IT in the remainder of this chapter refer to hours of instruction time, but these are not (necessarily) integer numbers. We take reading and language instruction together for our measure of instruction time, since both these subjects aim at improving the understanding of language.⁵

The first PRIMA wave was executed in '94/'95. This wave differs from the others because tests are conducted at the beginning of the academic year⁶ and questions concerning instruction time are aimed at the previous year. This means that scores are lower (pupils have had less schooling) and IT-values are from another grade (3rd, 5th and 7th). The first PRIMA study is therefore discarded. Although the '96/'97 is the earliest study included in the analysis, we refer to it as the second wave in the remainder of this chapter. As an abbreviation, we label this wave as T2. The other waves are similarly indicated as T3, T4, T5 and T6.

¹Math, language and environmental studies, which is a combination of biology, geography and history.

²The results for the Cito test for language is not comparable, since the number of questions in the test was changed between waves 4 and 5. Therefore, we cannot check whether the results would be similar as those for the 'conventional' language test.

³In addition, we cannot employ reading scores. The reading test was only conducted from the third wave on, which means we do not have any observations from before the policy change.

⁴The exact question reads: "How much school time did the pupils receive on average per week on the following subjects:"

⁵The coefficients are highly similar when we exclude time spent for reading.

⁶The other waves were executed around March.

4.4.2 Descriptive statistics

We exploit variation in instruction time caused by policy interventions to estimate the effect of IT. The new guidelines were enacted right before the new academic year '98/'99. One would therefore expect any changes to show up from the third wave on (i.e. '98/'99). Figure 4.1 shows how IT fluctuates over time. We only include schools that recur in every wave, to make the sample comparable over time. There is a clear increase of roughly have an hour of average instruction time in language between waves 2 and 3.¹ This means half an hour extra language instruction every week, which adds up to a substantial sum over a whole schooling year. Table 4.1 shows that half an hour corresponds to almost half a standard deviation. There are a few outliers in grade 6,² which we discarded, since these appear to be misreported values. Summary statistics on instruction time per subject and test scores are listed in table 4.1.

We only show mean levels until wave 5 because this gives us a larger number of observations. This involves dropping wave 6, but we also retrieve schools that are in all waves but the 6th. In the end, the tradeoff in terms of total observations is favorable.³ In order to estimate long-run effects, we need to include at least the fifth wave, since this wave includes 6th graders that experienced the jump in IT in the 2nd grade as well. Thus, reducing to even fewer time periods is not possible, but the 6th wave is not a necessary time period. Moreover, we also believed that a larger number of observations per time period is more valuable than an extra (not strictly necessary) time period.

Since we are also looking at the effects of language instruction on mathematics achievement, we need to know how IT for especially math has developed over the same period. If mathematics instruction has risen along with language instruction, σ_1 is positive, and a large part of our estimated effect might originate from the added value of math instruction. If mathematics has fallen strongly, our λ (the negative component, originating from the fall in instruction in other subjects) could be very strong. Figure 4.2 shows that volatility in math-IT is very low, especially for the periods before and after the policy change (T2 and T3). The only noticeable difference occurs between the 3rd and the 4th wave in the 2nd grade. Even then, the difference is only 0.32 hours, or 19 minutes of math instruction more per week. Moreover, we only estimate effects

¹The trend for grades 4 and 6 is less apparent if one would look at the sample as a whole. This is because there are a lot of schools entering the sample by wave 3 with low levels of IT, lowering the average for wave 3. To illustrate, the schools that first entered in wave 3 had mean-IT's of 7.34 and 6.59 for 4th and 6th grade respectively, in the third wave. For schools that are present in all waves, this is 7.87 and 7.05.

²Most notably one school that gave 8 hours on language instruction less in wave 3 than in wave 2.

³With 4 waves, we have about 800 observations more per grade.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

for math at the 6th grade level and students who are in 2nd grade in the 4th and 5th wave never appear in grade 6 in our data. Therefore, even this slight increase does not affect math estimates.¹ All other variation is lower than 10 minutes of instruction per week. Thus, the σ for mathematics lies close to zero.

Where has this extra language instruction come from? With respect to the jump from wave 2 to wave 3, we cannot tell exactly (aside from the fact that it is not math). For later waves, the data on instruction time include more subjects. Increases in language tend to come at the expense of, foremost, ‘all other time’.² This is defined as the difference between the sum of all reported subjects and the length of the schooling week. This consists of activities that could not be grouped in any of the other eight categories (language, math, environmental studies, Physical Education, Expressive Subjects, Music, English and ‘Dealing with computers’) on which the quantity of IT was reported. The policy change led to Targets with a stronger focus on language and reduced demands on the ‘building of citizenship’. It is likely that the bulk of this other time went into those activities, also given the rise in importance of this topic in the period before the policy change. These are activities which are presumably less productive in the subjects that we have test scores on. Therefore, it is expected that our λ is low and our estimates lie close to the ‘pure’ effects of increased language instruction.³

4.5 Methodology

If we want to estimate the effect of instruction time (IT) on test results, we have to correct for endogeneity. The allocation of instruction is decided by schools and teachers. This can be depend on the average performance level of the school, which would mean that there is an ability bias when employing OLS. We correct for endogeneity by using (exogenous) variation over time for language instruction.

We would expect that schools with proportionally more low-ability students have higher levels of instruction time for language, since they would need to spend more time on the in-

¹Only effects for IQ in 2nd grade and the first lag in 4th grade could be slightly affected.

²The fact that this was the case for later periods does not automatically imply that this worked in the same way for the period after the policy change. However, the decision-process is believed to be the same; schools feel that they need to spend more time on language and the new Targets have an overall guiding effect in where this time will be taken from.

³The negative correlation between ‘other time’ and language IT is strongly apparent in all three grades. It is very strong in 2nd and 4th grade (-0.50 and -0.60 respectively) and lower, but still high and significant, in grade 6 (-0.22). The only other subjects that have significant negative correlations with language IT are Religion (all grades) and English (grade 6). The strength of these correlations are much lower than that of ‘other time’ and language IT.

struction of basic skills, while schools with more high-ability students can spare more time for ‘secondary’ subjects, such as English or expressive subjects. We check whether this is what we find in practice, by using several measures for ability or socioeconomic status. These are education level of the mother and the father and the average Cito score of the school over all waves. Table 2 shows that there is a (negative) correlation between instruction time for language and these indicators. This relation is strongest for 2nd grades and weakest (but still significantly negative) for 4th grades. This implies that the effect of IT on scores, estimated cross-sectionally, will be underestimated or even negative, since higher values for IT are associated with low-performing schools and therefore lower scores. This could explain that many studies find very low or insignificant effects of instruction time on performance.¹

We need an instrument that is correlated with instruction time, but not with the average level of the school. The educational policy changes in the middle of the 1990’s presents such an instrument. The PRIMA-study was repeated each two years, given a total of 6 different studies, of which 5 are comparable. Between wave 2 (’96/’97) and wave 3 (’98/’99), there are large changes in the allocation of instruction time for language. This variation can be used as an instrument for IT. More specifically, we employ a dummy (D) which takes the value 0 for the ’96/’97 study and the value 1 thereafter, as an instrument. This presents us with the following model, where IT stands for instruction time:

$$IT_i = \beta_0 + \beta_1 D + \epsilon_i$$

$$Score_i = \beta_0 + \beta_1 IT_i + \epsilon_i$$

The jump in instruction time from wave 2 to wave 3 should ensure a strong correlation between this dummy variable and instruction time. The dummy variable will not be correlated with the error term, as long as there are no other changes over time that affect scores. We estimate effects for each specific grade separately. Thus, we are not following the same individuals over time (as these naturally advance in grade) but are comparing the same grades on the same schools in different periods. This sample can be seen as a school panel, but still containing observations on an individual level. Our dependent variable *Score* refers to four different test scores: language, IQ, mathematics and environmental studies. The effect of IT is estimated separately for all four test scores. The β_1 in the model above corresponds to our $\beta\text{-}\lambda$ from model 2.1. It incorporates

¹The Cito scores themselves can be affected by IT as well. Still, correlations between language IT and overall Cito scores are positive (although insignificant), including the negative ability bias, so this would only underestimate the extent of the negative selection bias.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

both the added value of having more instruction time and the negative component of having less instruction time for other subjects.¹

Additionally, we are interested if IT in earlier grades still has an effect on performances in later grades and how this effect develops exactly. We employ lagged IT variables in regressions for 4th and 6th grades. Since these lagged variables are endogenous as well, they need to be instrumented. We again employ dummies as instruments. If the lagged IT-value is incurred before the ‘policy change’ the value of the dummy is 0. If it is incurred after, it takes the value 1. This means that for the first lag, the first two time periods correspond to a dummy value of 0 and the last two to a value of 1.² For the second lag, the first three periods have a dummy value of 0 and the last period has value 1 (see table 4.2).

$$IT_i^6 = \beta_0 + \beta_1 D6_i + \epsilon_i$$

$$IT_i^4 = \beta_0 + \beta_1 D4_i + \epsilon_i$$

$$IT_i^2 = \beta_0 + \beta_1 D2_i + \epsilon_i$$

$$Score_i = \beta_0 + \beta_1 IT_i^6 + \beta_1 IT_i^4 + \beta_1 IT_i^2 + \epsilon_i \quad (4.1)$$

The above pertains to estimation in 6th grades. D6, D4 and D2 are the (dummy-) instruments for instruction time in language in grades 6, 4 and 2, respectively. For estimation in 4th grades, the second lag is absent. For 2nd grades, no lags are present, since it is the earliest observation we have. The lags combine the effect of language instruction with our measure for complementarity (the α ’s from model 2.1). This model effectively uses changes in the degree of treatment. Some students are treated in no grade; some in all grades; one group of students is treated in grade 6, but not in 4 and 2; another in both grades 6 and 4 but not in 2. This variation is used to estimate the effect of instruction time in all grades.³

¹Which is presumably very small for the test domains we are looking at, as explained in section 4.4.2.

²Again, we use 4 waves: the 2nd to the 5th.

³We do not have lagged values for the earliest observations, since these are incurred before the first time period. We assume that these values are equal to the value of IT for that particular school in the second wave (which is the earliest observation we have). This refers to the first lag of 4th graders in T2, the first lag of 6th graders in T2 and the second lag of 6th graders in T2 and T3.

4.6 Results

We present results of both OLS and IV estimation for the effects of language instruction time on language, IQ, math and environmental studies scores. IV estimation uses treatment dummies as instruments. Table 4.4 reports the F-values of the first stages for all instrumented variables in all three grades. All values easily exceed the commonly accepted threshold of 10,¹ ranging from 35 to 80. Hence, there is a strong enough correlation between our dummy variables and our instruction time variables. The fact that the F-values for instruction time in a specific grade are slightly different is mainly due to some degree of collinearity among the instruments, when more than one instrument is included.

4.6.1 Language

The presence of selection bias reported in table 4.3 suggests that OLS results will be underestimated. OLS results for language scores are listed in the first column of table 4.5. They are positive for all grades, but only the effect for 4th grades is significant. The fact that the 4th grade estimate is highest is likely related to the fact that the negative selection bias is weakest for that sample. The effects are low in magnitude. The highest coefficient (4th grade) would only correspond to an increase in score of 0.02 of a standard deviation for an increase in instruction of one hour each week.

Column II of table 4.5 shows that IV effects are clearly larger for all grades. This suggests that OLS estimates are indeed biased downward, because of the (negative) correlation between school-level and instruction time in language. Measurement error in the IT-variable can also have a downward effect on OLS-estimates. We only observe significant effects of Instruction Time on language for 2nd grade. However, we need to add lags to get a complete picture of IT-effects.

Estimates when lags of instruction time are added are listed in columns III and IV of table 4.5. The 2nd grade estimates are the same as the Instrumental Variable coefficients in column II, since there are no lags available for 2nd grade estimation. For 4th grades, we can add the instruction time experienced in the second grade, as a second explanatory variable. We observe positive and weakly significant effects on score for current IT (meaning IT received in that particular grade), but insignificant (positive) results for lagged IT (meaning IT received in an earlier grade). The positive effect of more instruction fades out quickly. When we add lags to

¹First proposed by Staiger and Stock (1997).

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

the 6th grade model, a similar picture occurs. Current IT becomes significantly positive. Both lags are negative, but insignificant. Again the effect of instruction in earlier grades fades away. This suggests that the effect of increased instruction time for language has only a temporary ‘boosting’ effect but no long-lasting positive payoffs. In relation to model 2.1, the α ’s are clearly lower than 1.

The lack of long-term effects could be due to the specific nature of the 6th grade language test, which might not be discriminative enough. However, table 4.1 shows that the standard deviation is not lower for grade 6 than for the other two grades. Moreover, we can compare the PRIMA language score with the Cito language score. The Cito test is specifically designed to test all the aspects of what students have learned in their six years of primary school. Its discriminatory power should therefore be very high. We observe that ‘our’ PRIMA language test correlates strongly with the Cito language test. Correlations are in the range of 0.67 to 0.76 for all waves. The correlation between PRIMA and Cito test is slightly higher for math (0.84) but this difference is not surprising given the more homogeneous nature of math as a test subject (especially in primary school). This suggests that a lack of discriminatory power of the 6th grade language test is not the reason for the absence of any long-run effects of language instruction on language scores in that grade.

4.6.2 IQ

We find that instruction time for language has a strong positive effect on scores on non-verbal IQ tests. These effects are more significant and larger in magnitude than those of IT on language scores. We can observe these results in table 4.6. OLS estimates are again low, although they pass the 5% significant mark for 2nd and 4th grades. IV estimates are strongly significant with and without lags. For 4th grades, we observe both a positive and significant effect of IT in the current grade and of the first lag of IT. Thus, in contrast to language scores, IQ scores are also positively affected in the longer run. For 6th grades, the first lag is significantly positive as well. We do not observe significant effects of 2nd grade IT on 6th grade test outcomes, although the effect is positive. Hence, although the effect of language instruction is lasting longer than for language skills, there is no evidence of a significant long-run effect. Both the initially strong impact on IQ as well as the later fade-out are very common findings in the literature of childhood interventions, including programs such as Perry and Head Start.¹

¹See e.g. Heckman et al. (2010); Westat (2010).

The positive effect of language IT suggests that children's brains are trained when they experience instruction in language, which positively influences their non-verbal intelligence as well. The fact that this is not (significantly) present for the long run could mean that effects of IT have a temporarily boosting effect that fades away after a few years. More language instruction could enhance the proficiency to solve these IQ tests (similar to Jaeggi et al. (2008)), without making long-term improvements in general intelligence. Still, the degree of fading out (our α) is limited, compared to the results for language.

4.6.3 Effects on other subjects

We also estimate effects of increases in language instruction on the performance in other school subjects than language. This allows the detection of possible spillovers of language instruction towards other skills. Test scores for math and environmental studies (ES) are only available at the 6th grade level. When we estimate the same model as before on scores for mathematics, we find that long term effects dominate (table 4.7). Instruction time in 2nd grade is the only of the three IT-variables that is significant at the 5% level. Additionally, we find significant long term effects for IT for language on ES. Again, only the IT in the 2nd grade has a significant effect on scores at the 6th grade level. Thus, early investments in language have strong and long-lasting effects into a range of other skills. This is intuitive, since students make use of reading and language (including listening) skills when they are studying other subjects as well. Especially in 6th grade, where math questions tend to be framed in the context of small stories, these skills certainly can be valuable. In other words, there is a strong complementarity between language, as taught in the early grades, and other skills, such as mathematics. It is also logical that this effect is only present for IT in early grades, since it is there where the fundamental reading and language skills are shaped. In later grades, language instruction becomes more specific and payoffs to other skills are less likely to occur. Machin and McNally (2008)'s study of the National Literacy Project also finds some evidence of changes¹ in literacy instruction on mathematics scores.

Spillovers can be present through direct effects of language instruction on skills other than language, or through an intermediate effect on general intelligence. Based on our results, we cannot say which of these effects is stronger. General intelligence is enhanced, but long term

¹These are qualitative changes, providing a different method of instruction.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

effects for mathematics and Environmental Studies seem to be too strong to fully contribute this to IQ effects.

In addition, since we do not have scores on these subjects in multiple grades, we cannot say whether the stronger impact of early instruction time is due to high complementarity over time or a high initial impact. As Cunha et al. (2010) point out, the optimal ratio of early to late investment is positively related to both the degree of complementarity over time and the relative investment productivities in different periods. We know the combination of these two concepts is high in the case of language investment on math skills, but we cannot say which one dominates.

4.6.4 Magnitudes

So far, we have only focused on the direction and the significance of IT-effects. In terms of magnitude, the effect of language instruction is both stronger and longer-lasting for IQ than for language itself. As stated before, such strong short-run IQ effects are not that uncommon in the literature. The magnitudes are generally high. However, this refers to one hour instruction time more every week. Over the line of a whole academic year, these are significant increases. On an average of around seven hours, an increase of one hour comprises a big share of total hours of language instruction. Such an increase is close to a full standard deviation for the instruction time variable, as can be seen in table 4.1.

Some coefficients can pick up on variation in intermediate grades. For grade 2, the variation between T2 and T3 exists only of more language in grade 2. For T4 and T5, this also includes more language in grade 1. This also applies to the first lag in grade 4. The other coefficients are not affected, since they are purely based on the variation between the periods just before and just after the policy change, since there is no variation in the value of the other dummies for the other time periods. We can identify the effect of only more language in grade 2 by dropping periods T4 and T5 for grade 2 results. This indeed leads to a decrease of the effect to 0.342 of a standard deviation in the test score. If we only look at the effect of more language in both grades (dropping T3), this increases to 0.410. Hence, although we indeed pick up an effect of more language in grade 1, this extra effect is not large. Even if we assume that the increase in language in grade 1 was of the same magnitude, this result is not surprising. The test in grade 2 is designed to measure topics studied in that grade and we have seen that effects from earlier grades fade away quickly.

4.6.5 Effect of week of testing

We also developed an alternative measure for magnitude, by expressing the effect of an extra hour of language instruction each week in number of weeks of instruction. The 4th wave of our dataset reports the week at which the test was taken. All tests are taken between January and March, but there is variation within this time period. This allows us to estimate how a whole week of schooling affects language scores. We thereby assume that this week of testing is randomly chosen; it is not affected by the average level of the school.¹ Results show that testing one week later has a significant positive effect on test scores.²

The effect of week of testing on scores centers around 0.15 of a standard deviation for 2nd and 4th grades. For 6th grades it is clearly lower: 0.06 of a standard deviation.³ This means that an extra hour of language instruction each week substitutes for around 2.6 weeks of schooling for 2nd grades, 1.8 weeks for 4th grades and 3.1 weeks for 6th grades, in terms of language performance. Test are made after 25 weeks of schooling, on average, while the average level of language instruction per week is 7,5 hours. This means that one extra hour of language every week adds up to the same sum as the amount of language taught in 3.5 weeks of schooling. However, the effect of week of testing already achieves the same impact after 2 to 3 weeks. This can have two main reasons. First of all, other subjects are taught in an extra week of schooling as well. These subjects can also have a productive value with regard to language skills, either directly (through ω) or through general intelligence. Secondly, it could mean that putting too much of the same subject in one fixed time period can reduce the return on that instruction. Children might pick up less from yet another hour of language instruction in that week than if that instruction would be spread out more over the academic year.

4.6.6 Heterogeneity

To address heterogeneity in the effect of instruction time along the distribution, we estimate effects for different parts of the distribution. Our method is a variation on the estimation of quantile regressions, first developed by Koenker and Bassett (1978):

¹There is no significant relationship between week of testing and average Cito score (which is taken on the same date by every school) or educational level of the parents.

²No significant correlation is found between week of testing in wave 4 and test scores in other waves. This suggests that the correlation between weeks of testing in different periods is low, since otherwise a significant positive correlation with other waves should have been found as well.

³The direct effect of language instruction is only slightly lower in grade 6 than in the other two grades. The larger difference here can be due to lower complementarity between skills. Subjects become more specific, implying that everything that is taught outside of the language lessons has less value than in other grades.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

$$y_i = x_i' \beta_\tau + u_{\tau i} \quad \text{with} \quad Quant_\tau(y_i | x_i) = x_i' \beta_\tau$$

where τ indicates the specific quantile of the distribution. Our approach is similar, but we recode the outcome variable to the rank in the distribution as well. Rank is based on the percentage of people in the whole dataset (thus taking all waves) which had a specific number of questions correct. If 65 percent of all pupils had 45 or more questions correct, then someone with the exact score of 45 has a rank of 0.35. We then estimate to what extent increases in IT move you up in the rank and for which part of the distribution this effect is most prominent. We split up the ranking in three parts and conduct separate analyses:

$$\begin{aligned} Score1_i &= \min(1/3, score_i) \\ Score2_i &= \max(1/3, \min(2/3, score_i)) \\ Score3_i &= \max(2/3, score_i) \end{aligned}$$

We estimate model 4.1 with this alternative outcome variable. The effect for the whole distribution is the sum of the three separate effects. This method is an alternative to the more straightforward approach of conducting split regressions for different groups. One benefit of quantile regression is that it uses all observations in each separate regression. Additionally, it is less vulnerable to outliers.

Results are listed in tables 4.8 (for language score), 4.9 (for IQ) and 4.10 (for math and environmental studies). The overall picture from table 4.8 is that the two lower groups gain the most from receiving more instruction in a particular grade. The only exception are the results for grade 6, where effects are stronger for the highest group.¹ Short-term effects fade away for every subgroup.

Table 4.9 shows a similar picture for 2nd grades, for IQ. The effect is consistently strong and strongest for the middle group. The effects of instruction time for the ‘top group’ are relatively low for grades 4 and 6, but the lags are strong for this same group. Hence, the highest

¹This could be due to the nature of the test, which is more diverse and introduces ‘new’ topics (there are more questions on understanding and interpreting sentences). Better students might be more receptive for (the presumably limited) instruction in these new sub-topics.

quantile experiences stronger complementarity and/or self-productivity. The effect of 4th grade instruction lasts longer than that of 2nd grade instruction, holding for all groups.

The results for math scores show that the long-run effect originates mainly from the low end of the distribution. Still, the effects for the other two groups are only slightly lower. The lower group has a significantly negative effect of IT in the current grade. For Environmental Studies, the middle group gains the most from language instruction, holding for all three IT-variables. The long-run effect is lowest for the highest group, which is consistent with earlier results and suggests that the added value of instruction is in general lower for these students. This can be because they are already at a high level and are able to further develop their skills also with a relatively low (quantitative) level of instruction.

We also estimate differences in effects between boys and girls. These results (not portrayed here) show that those results that are significant are stronger for boys. The only exception is the 4th grade direct effect, which is slightly higher for girls. However, direct effects for language are not sustained in the long run and fade away more strongly for boys. The first lag for math achievement is close to a 10% significance level and strong in magnitude for boys (standard errors are naturally high with split samples). Hence, there is suggestive evidence for an effect of 4th grade language instruction on 6th grade math skills for boys.

4.6.7 Panel data

The results we find for language achievement are confirmed when we look at the development in scores of individual panels. We can construct a panel of individuals on which we have observations in all three groups. We can distinguish three panels; one running from the second to the fourth wave, one from third to fifth and one from fourth to sixth wave. The ‘oldest’ group (T2-T4) is characterized by the fact that IT levels are still low in the 2nd grade. The other two groups (T3-T5 and T4-T6) have all their observations after the ‘change’. Mean scores for each panel in each grade are portrayed in figure 4.3. Mean scores are substantially lower for the oldest group in 2nd grade, but they already catch up this deficit by grade 4. Thus, the increase in IT for the two later cohorts does nothing but accelerate skill development in language that otherwise would have taken place at a later time. In the end, there is no long-term payoff from investments in instruction time in 2nd grade. The data contain a fourth panel with students that are already in grade 4 in wave 2 (there is no observation in grade 2), where they receive a relatively low level of language IT. Once they receive additional instruction time in grade 6, they catch up again. We can check if an absence of a jump in all grades would have prevented such a catchup, by

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

checking 6th grade scores from T2. This is the group of students that did not receive more IT in any period. The mean score for this group is 52.56, which is very close to the other 4 estimates in the graph (they range from 52.61 to 52.76). Hence, the policy intervention that led to jumps in instruction time for all these grades has had no payoff in terms of *language* skills at the end of primary school.

This illustrates that there is low self-productivity and complementarity over time. However, the absence of these two effects cannot explain the complete fade-out. The level of language skills in the last period is the same between the very oldest and the very youngest cohort. Even when investments are perfect substitutes, the younger cohort should have a higher skill level, because the *absolute* amount of investments differs.¹ Hence, the marginal value of these extra investments decreases, and essentially becomes zero. They have an effect in the short run, but since they lower the value of later investments, their net long-term effect becomes zero. The fact that marginal effects decrease is in line with results from Kiesling (1984) and Soar (1978). This means that either there is an early saturation point for learning language skills,² or there is fast depreciation when learning these skills. There is not enough exogenous variation in investment over time to say which one applies here.³ The above does not mean that the quantity of language instruction is irrelevant for the long-term level of language skills. It means that the old (relatively lower) levels of language instruction were already high enough to bring the pupils to their ‘maximum’ levels of language performance. These results depend on the original level of language IT. If this level would have been much lower, long-term positive effects of language instruction could have been identified.

The pattern in score in figure 4.3 closely follows the pattern in instruction time in figures 4.1 and 4.2. Besides the more obvious trend we explained above, any other fluctuations between the three different panels could be explained by trends in IT. The fact that scores in 2nd grade are also higher for the T4-T6 panel compared to the T3-T5 panel can be linked to the fact that IT also increased slightly from T3 to T4 in that grade. Additionally, the relatively low average 4th grade score for the T3-T5 panel can be explained by the fact that IT in that grade is somewhat lower in T4. This is further evidence that fluctuations in score are strongly related to fluctuations in instruction time, at least in the short run.

¹And since investment is higher for the youngest cohort in every grade, this cannot be explained by differences in the relative productivity of investments in different periods, as presented in Cunha et al. (2010).

²The fact that we find this for multiple grades suggests that this would be a saturation point for the individual, and not of the particular test.

³This would require one comparison group where the investment takes place in an early period, but not in a later one. We do not have such a group since investments will be made consistently after the policy change.

We also provide the same figure for IQ scores in figure 4.4. Scores should not be compared for different grades, since both the content of the questions and the exact number of questions differ per grade. A very similar picture as in figure 4.3 occurs. The two later cohorts start at a higher level again, since they had extra investments in grade 2. The other two cohorts catch up when they also receive higher levels of language instruction time in that particular grade. One difference with the pattern for language scores is that the 6th graders from wave two, who had lower IT levels in all grades, have lower scores than all other waves. The mean IQ score for this group is 25.8, which is somewhat lower than all other four estimates. Hence, the policy change has had some effect for IQ in the longer run, which is also reflected in our estimation results for IQ.

4.7 Robustness

4.7.1 Other changes over time

4.7.1.1 School-level changes

When exploiting variation in the explanatory variable over time, there is a danger of also incorporate other (qualitative) changes within the same time window, which also influence test scores. The nature of the change in policy guidelines does not suggest strong accompanying changes in the qualitative dimension. The underlying reason for the jump in IT is that schools were being freed of rather strict guidelines with respect to secondary school subjects. The more elaborate and mandatory nature of the guidelines on language instruction ensured that the bulk of that freed up time went into language instruction. There seems to be no reason to believe that this would involve significant changes in quality or other relevant factors as well.

We can formally analyze this, by looking at information reported by schools. We test whether there are statistically significant differences in the means of the two time periods. Results are displayed in table 4.11. The vast majority of school-related measures has not changed from wave 2 to wave 3. The only two exceptions (at a 10% significant level) are the degree of extra financial support and the fraction of schools that work with end-of-year-targets for the core topics. The latter only emphasizes our notion that the more demanding Targets for language, which were set at the macro-level, also effectively reach the average school. We observe no increase in the setting of intermediary targets, suggesting that the setting of targets did not effectively led to stronger tracking of performance. It seems unlikely that this is behind the large

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

effects on test performance that we find.

The increase in spending is not only present between these two time periods, but over the last twenty years in the Netherlands (and worldwide). Government reports show that the majority of this increase goes to increased wage costs and technical adjustments (largely IT-related costs) (Centraal Bureau voor de Statistiek, 2010). Table 4.11 shows that these extra finances are not directly aimed at the enhancement of language instruction. Moreover, if the extra resources would have led to strong increases in performance, one would expect long-term increases in test scores in all subjects, especially taking into account the strong and continuous increase over the last two decades. This is not the case. Additionally, the gradual increase in spending cannot explain that students in all grades get a boost in performance exactly between time periods 2 and 3, while there are no strong shifts in other time periods (as figure 4.3 illustrates). This confirms that the overall (mean) increase in language instruction has been indirect, because obligations on secondary topics were being relieved. There are no changes in school philosophy or execution of language instruction.

4.7.1.2 Class-level changes

Teachers report several measures at the class-level as well, which give an indication of the content of language classes.¹ The large majority of this wide range of variables is not significantly different between the time periods. There are no strong differences in the learning material used for language instruction, which are still strongly dominated by two specific methods.² Also, there are no differences on the amount of homework for language and the registration of progress in language. The data also report information on how frequently eight subtopics in language are discussed. Naturally, these values are somewhat higher after the change since overall language instruction has increased, but there are no strong shifts. There is a relatively strong reduction in teaching students how to use dictionaries and a relatively strong increase in practicing phone conversations and interpreting tables and graphs. The latter might have contributed to the positive effect of language on math, although this difference is only significant at the 10% level. There is also a difference in the frequency of testing for language, but that is a likely consequence of overall time spent on language going up. The only qualitative change in language

¹These results are not reported here, but available on request.

²The share of the most popular method goes up from 49% to 57% and that of the second most popular method goes down from 20 to 14%. The share of all other methods remains stable at around 30%.

instruction that we find strong evidence for is a reduction in the degree of letting students develop language skills on their own pace, which is significant at the 1% level for both grade 4 and 6. We can only speculate on the origins of that change, but it might be a consequence of having more group instruction, and less time for independent work.

The latter result raises one other issue. We have seen that increases in language IT were the result of decreases in ‘other time’. Although the nature of the policy change suggests that this has largely been time on the building of citizenship, there might have been reductions in time in which students can work independently. Some of that time could have been devoted to independent learning in language. If this is the case, our explanatory variable might effectively measure the effect of somewhat less than 1 hour of extra language instruction, and the coefficients would be larger in absence of such a change. However, since our long-run effects for language instruction on language performance are far from significant, this is unlikely to be behind the lack of any long-term effect in that area.

Hence, we find no strong qualitative differences over time in the way that language was taught, across a wide range of measures. The occurrence of significant differences is very low, especially taken into account that these are vulnerable to multiple hypothesis testing.

4.7.1.3 Cross-sectional changes

We also add controls to check whether the coefficients pick up on changes in the composition of the cohorts over time. A mean comparison of observable characteristics (not shown) indicates that the cohorts are not fully identical in background characteristics. The difference for ethnicity, month of birth, gender and whether the child lives in a two-parent household are insignificant, but parental education is significantly different across time periods. This is part of a general trend that education levels, and hence parental education levels, are increasing over the last decades. When we add all these variables as controls, this induces some changes (see column V of tables 4.5 to 4.7). The coefficients are hardly affected for 2nd and 4th grades and for IQ in all grades. The changes are somewhat stronger for language scores at the 6th grade level and for the effects on other subjects. For the latter, weaker instrumentation naturally leads to stronger variation. The significant effect for current IT on language score in grade 6 disappears, while it becomes significantly negative for math scores. The second lag for ES is marginally affected, but loses its significance ($p=0.106$). All these changes stay within one standard error. One could question whether adding controls provides a good indication in this setting. As mentioned, the rise in parental education is part of a general trend. The pattern in test scores over the last

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

decades does not follow that same pattern, as we see in the next subsection. The cross-sectional effect of parental education on test scores is strong, also because it picks up on other unobserved variables. The question is whether that relation will be as strong for variation over time and whether this does not lead to ‘overcontrolling’ for this variation. Nonetheless, adding controls still leaves our conclusions in tact and the induced variation is moderate.

4.7.2 Trends

There is a danger of picking up on a general trend in test scores. Extracting this general trend is not feasible in this setup, since our instrumentation is already based upon time indicators. Additionally, we can also not simply look at development of test scores in all waves after the policy change since pupils in these waves still differ in the received instruction in earlier grades. An alternative way of approximating a general trend is by looking at trends in alternative data sources, such as PIRLS test results (4th grade reading tests) for the Netherlands. Comparison shows that pupils scored 0.1 of a standard deviation lower in 2006 than in 2001.¹ Mean scores for all countries together are extremely similar (for those countries appearing in both studies), so the difficulty level of the test is comparable. We repeat our analysis of language IT on language scores, thereby extracting a general trend whereby test scores decrease with 0.04 of a standard deviation per two years. Results are listed in table 4.12 (for 6th grade). Our estimates increase somewhat in magnitude, but there are no strong changes in significance levels.² Both lags become positive, where the second lag is virtually zero. Hence, the effect of instruction in 2nd grade might be slightly more long-lasting if there has indeed been a slight negative trend, but initial effects still fade away at higher grade levels, leaving our initial conclusions intact.

Similarly, we check the development of 4th grade math scores by comparing TIMSS '03 with TIMSS '07. They show a similar decline of slightly more than 0.1 of a standard deviation. Extracting this trend in a similar way leaves our main conclusions intact. The only change is that the first lag has also become significant now, meaning that 4th grade language instruction also has some added value for math performance in the long(er) run.

¹The 2001 group entered first grade in the academic year '97/'98 and the 2006 group in '02/'03. Hence, only the first grade for the 2001 group occurred before the policy change. Since we saw that early IT effects on language skills fade away at higher grade levels, this is not likely to affect our main results.

²The only change in significance in all grade levels is that the first lag at the 4th grade level is now significant at a 5% level. The coefficient has increased to 1.83, through both a higher initial effect (β_1) and the fact that extracting the trend leads to lower fade-out as well (α_2).

The trends in PIRLS or TIMSS deal with different time periods than the one we are looking at, but they nevertheless serve as a good approximation of how scores would have developed in absence of changes in instruction time. Additionally, the analysis shows that our conclusions are not sensitive to trend effects, if even a assumed negative trend of 0.1 of a standard deviation per five years (which is severe in this context) does not affect these conclusions, or even strengthens them. On the other hand, the results would be more sensitive to incorporating a positive trend. A positive trend of 0.04 of a standard deviation per two years would affect our 6th grade language results, by making both lags significantly negative (effects with opposite trends can also be seen in table 4.12). This is not a sensible result. Similarly, a positive trend of 0.04 for math would lead to very strong negative effects for current IT and the first lag of IT (and would make the second lag insignificant, but still clearly positive). This is also hard to explain, since IT in math stays constant over time.

Our two main findings, namely that early language instruction has no payoff in the long run, and that it does enhance math achievement, are only changed when we use very extreme trends. The former is robust up until a negative trend of 0.1 per two years (which means 0.5 of a standard deviation per decade). In that case, the second lag becomes just significant at a 10% level, while the current and first lag effects become unrealistically large. The effect for math becomes insignificant with a positive trend of 0.04, but is still large (standard errors are higher for math since we have fewer observations there). The coefficient is positive up until a negative trend of 0.09, and leads to unrealistically large negative effects for the other two coefficients. Hence, our main conclusions are only sensitive to extreme linear trends, whereby this would both lead to non-sensible results for the other coefficients and be in contrast to observed patterns in score from PIRLS and TIMSS data. The fact that trends are decreasing over time, while parental education has been rising, also confirms that adding controls can lead to an ‘overcorrecting’ of scores for later cohorts.¹

We also check sensitivity to trends for IQ estimates. We take the controversial ‘Flynn effect’ as a guideline. This effect states that there is a general trend of IQ scores increasing by 3 points every ten years (Flynn (1987)). This pertains to 0.2 of a standard deviation per 10 years. This again leads to a trend of 0.04, now in a positive direction. Incorporating this trend still leads to significant and positive effects for current IT and the first lag of IT. The second lag is reduced to

¹Naturally, non-linear trends can change our results more drastically. However, this would have to be an extremely irregular non-linear trend.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

virtually zero, hence it could be the case that the positive effects of language instruction on IQ fade away more strongly in the long run than based on our original estimates.

4.7.3 Volatility after change

Our method assumes that there are no changes in instruction time after the dummy changes sign. However, there is some variation of IT also after the policy change. Figure 4.3 suggests that other fluctuations in language IT only strengthen our points with respect to the effect of language instruction on language skills. Moreover, this variation is both relatively small (always smaller than the initial jump) and inconsistent. It does not move consistently in one direction. On average, the levels of instruction time for the periods after wave 3 are close to the levels of wave 3 itself. In other words, the upward and downward volatility cancel each other out. Our assumption that IT is linear after wave 3 therefore seems appropriate in this case.

Additional analysis suggests that results are not sensitive to these fluctuations. When we drop time periods for our estimation in grades 2 and 4,¹ where volatility is at its highest, we obtain coefficients that are very similar. Standard errors are affected stronger, since we have fewer observations, but significance is still very high. We also add wave 6 and see how that affects our estimates. Again, results are hardly affected. Results for language in grade 2 go slightly up (from 0.320 to 0.373) which can partly be explained by the fact that we add another period that picks up on possible changes in language instruction time in grade 1. Effects for grade 6 are highly similar. A similar picture occurs for IQ and mathematics. Additionally, we use the average instruction time for language per time period as an alternative instrument. Results with this alternative instrumentation are highly similar to our original results.²

4.7.4 Non-cognitive skills

We have seen that the increase in language instruction does not come at the expense of instruction in any core topics, which makes it safe to assume that there are no reductions in other topics that could impact our outcome variables. Section 4.2 indicated that the extra time came largely at the expense of topics related to the building of citizenship. Such instruction could still have benefits for certain types of skills, presumably those which are non-cognitive. There are some non-cognitive indicators that are consistently reported and measured in the exact same way across waves. These are factor variables for self-confidence, feeling well in school and social

¹We cannot do this for 6th grades, since we need a minimum of 4 time periods.

²These results are available on request.

integration in class.¹ All these measures are teacher-reported. Results show that the increase in language instruction (and the corresponding decrease in instruction focused on building citizenship) has different impacts per grade. More language in grade 2 has positive impacts on measures of non-cognitive skills in grade 6,² while the effect is negative for instruction in grade 4,³ and insignificant (and negative) for instruction in grade 6. Thus, the fact that early language instruction provides spillovers towards other academic topics even carries over to non-cognitive domains. More language instruction in later grades, which already did not have clear payoffs for academic subjects, can lead to harmful effects for non-cognitive measures.

4.8 Conclusion

In this chapter, we specify that changes in the allocation of instruction time (IT) for a specific topic should be analyzed from a broad and multi-dimensional point of view. Changes in IT for one particular topic will have implications for a diverse set of skills. Besides the direct effect on achievement in the same subject, there can also be spillover effects towards skills in other subjects. Moreover, if we want to reach an efficient allocation of instruction time, it is crucial to estimate how any positive effect of IT on performance develops over time. Effects can fade away, or accelerate over time. Moreover, one has to take into account that increases in IT for one subject take time away from other subjects. Such an increase is only advisable when the productive value of the subject that experiences the increase exceeds that of the subject that suffers a decrease in IT.

We apply this theoretical concept to a case where we have exogenous variation in the given quantity of language instruction over time. This variation occurs through a policy change in the Netherlands. We find that changes in the quantity of language instruction have both direct and indirect effects on skill acquisition. We find that direct effects on language skills dissipate quickly, while indirect effects on mathematics and environmental studies and non-verbal IQ are long-lasting. Therefore, the real value of instruction time (IT) in language lies in spillover effects into other subjects. Extensive and early instruction in language is a foundation for skill development in other subjects. The absence of long-run direct effects suggests that there is

¹All outcome variables are measured in grade 6, where we can present a complete overview of the effects in language instruction in grades 2, 4 and 6.

²This effect is significant for feeling well in class and social integration.

³Statistically significant at the 5% level for self-confidence and feeling well in school; insignificantly positive for social integration.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

either a saturation level of language skills for each specific individual, or strong depreciation of skills learned in language. High (early) levels of IT can accelerate skill development, but have no added value in the long run as long as an adequate level of IT is being given. Additional analyses on individual panel data and analysis by subgroup confirm this.

By using an IV-approach, we can eliminate the endogeneity that normally plagues estimates of instruction time on performance variables. One could be concerned that our way of instrumenting incorporates qualitative changes in the same time period in which the policy change took place. However, given the background of the policy intervention that caused the increase in instruction time, it is not likely that there are strong non-quantitative changes within this same time frame, as additional analysis on school-level characteristics confirms. Results are also robust to incorporating time trends and variation in the number of included time periods. Still, the ideal way to get a complete picture of the effect of instruction time on skill development in various areas is to conduct a field experiment, where one would randomly variate the instruction time that students are subject to. Moreover, our analysis only solves one piece of the big puzzle of how to allocate time to subjects. We need additional research on the value of other subjects than language to know more precisely what we should teach children during which stages of life.

4.9 Figures and tables

Figure 4.1: Mean levels of language instruction time

Notes: Graph shows the mean level of language instruction time in a particular wave, for all schools that are present in all four waves of the PRIMA dataset. Waves 1 and 6 are discarded.

Figure 4.2: Mean levels of math instruction time

Notes: Graph shows the mean level of mathematics instruction time in a particular wave for all schools that are present in all four waves of the PRIMA dataset. Waves 1 and 6 are discarded.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

Figure 4.3: Mean language scores by panel

Notes: Graph shows Language scores in all three grades for four separate individual panels of students. The indicators T2 to T6 refer to time periods. The earliest time period refers to the period when the student was in 2nd grade, the next one to 4th grade and the last one to 6th grade. The exception is the T2-T3 panel, which only has observations for 4th grade (T2) and 6th grade (T3). The scores correspond to mean scores for that specific panel in that time period and grade.

Figure 4.4: Mean IQ scores by panel

Notes: Graph shows IQ in all three grades for four separate individual panels of students. The indicators T2 to T6 refer to time periods. The earliest time period refers to the period when the student was in 2nd grade, the next one to 4th grade and the last one to 6th grade. The exception is the T2-T3 panel, which only has observations for 4th grade (T2) and 6th grade (T3). The scores correspond to mean scores for that specific panel in that time period and grade.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

Table 4.1: Summary statistics

	2nd Wave		3rd Wave		4th Wave		5th Wave	
	Mean	Stand. dev.	Mean	Stand. dev.	Mean	Stand. dev.	Mean	Stand. dev.
2nd grade	Language Score	49.55	5.92	49.82	6.68	50.48	50.88	5.70
	IQ Score	26.98	5.33	28.35	5.24	28.11	28.21	5.26
	Pure Language Instruction	4.53	1.31	4.94	1.69	5.28	4.96	1.29
	Pure Reading Instruction	2.81	0.724	2.97	0.964	2.87	2.91	0.890
4th grade	Combined Language Instruction	7.33	1.60	7.91	1.72	8.15	7.88	1.58
	Math Instruction	4.40	0.901	4.40	0.674	4.72	4.57	0.775
	Language Score	50.85	7.82	51.35	7.17	50.59	7.18	7.39
	IQ Score	25.56	4.81	26.43	4.23	26.02	4.51	4.19
6th grade	Pure Language Instruction	4.82	1.14	5.36	1.61	5.05	1.06	5.07
	Pure Reading Instruction	2.62	0.987	2.74	0.932	2.54	0.849	1.33
	Combined Language Instruction	7.44	1.35	8.10	1.78	7.62	1.18	0.836
	Math Instruction	4.74	0.762	4.72	0.913	4.85	0.794	1.55
	Raw Language Score	52.56	6.41	52.81	6.29	52.15	6.40	6.46
	IQ Score	25.73	4.46	26.41	4.17	26.52	4.25	4.23
	Pure Language Instruction	4.80	0.922	5.01	1.52	4.88	1.19	5.08
	Pure Reading Instruction	1.92	0.727	2.09	0.749	2.03	0.800	0.752
	Combined Language Instruction	6.71	1.14	7.10	1.61	6.93	1.31	1.40
	Math Instruction	4.97	1.02	4.84	0.829	4.88	0.945	0.906
	Cito Score	534.83	9.65	534.36	9.93	534.37	9.85	9.50
	Cito Math Score	43.42	10.64	42.11	11.48	42.50	10.98	10.45
	Cito Environmental Studies Score	41.76	8.37	41.26	9.01	42.16	8.81	8.69

Notes: Summary statistics, referring to the dataset on which the empirical analysis is executed. This refers to all schools which are included in all four waves. The first wave is disregarded because it is not comparable; the last (6th) wave in order to increase the relevant number of schools. The Cito test determines selection into secondary education and is therefore only conducted in 6th grade (last grade of elementary school in the Netherlands). Combined language instruction is the sum of pure language and pure reading instruction. This sum serves as the main explanatory variable in this chapter.

Table 4.2: Value of dummy instruments

	'96/'97	'98/'99	'00/'01	'02/'03
D6	0	1	1	1
D4	0	0	1	1
D2	0	0	0	1

Notes: The table shows the values of the dummy instruments in every period, for all three grades. The above refers to estimation in 6th grade. D6 is the instrument for language instruction in grade 6, D4 for that in Grade 4 (two years earlier) and D2 for that in grade 2 (four years earlier). For estimation in grade 4, the first two rows are the same. Only, they belong to instruction in grades 4 and 2. For estimation in grade 2, only the first row applies.

Table 4.3: Effect of ability indicators on instruction time for language

Grade level	Average Cito score	Educational level of father	Educational level of mother
2nd grade	-0.029** (0.012)	-0.574*** (0.112)	-0.604*** (0.108)
4th grade	-0.022** (0.011)	-0.426*** (0.094)	-0.437*** (0.096)
6th grade	-0.038*** (0.010)	-0.545*** (0.086)	-0.497*** (0.088)

Notes: Standard errors are in parentheses

*Significant at 10% level **Significant at 5% level ***Significant at 1% level

In the first column, number of hours of language instruction is regressed on the average final score on the Dutch Cito test for that school on all waves. In the last two columns, number of hours of language instruction is regressed on the mean education level of either the mother or father on a school level, separately per wave. Educational level is categorized into 4 categories, from low to high.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

Table 4.4: First stage results (F-values)

Instrumented Variable	2nd grade	4th grade	6th grade
Instruction time	68.99	38.61	34.77
First lag of IT		81.47	51.08
Second lag of IT			44.19

Notes: Table reports F-values of the first stage of model 4.1. The instrument corresponds to a dummy that takes a value of 0 if the corresponding instruction time is received before the policy change and 1 thereafter. Hence, the table shows whether the policy change has a strong effect on instruction time in language.

Table 4.5: Effect of instruction time (IT) for language on test scores in language

Grade level	Expl. variable	I	II	III	IV	V	n
2nd grade	Instruction Time	0.019* (0.011)	0.320*** (0.075)	0.019 (0.011)	0.320*** (0.075)	0.339*** (0.085)	3636
	First lag of IT						
4th grade	Instruction Time	0.053*** (0.012)	0.178 (0.143)	0.053*** (0.012)	0.118* (0.063)	0.116* (0.061)	3366
	First lag of IT			-0.014 (0.012)	0.067 (0.050)	0.025 (0.049)	
	Second lag of IT						
6th grade	Instruction Time	-0.015 (0.012)	0.072 (0.098)	-0.0017 (0.014)	0.187** (0.089)	0.117 (0.095)	3101
	First lag of IT			0.023 (0.012)	-0.016 (0.061)	-0.091 (0.063)	
	Second lag of IT			-0.022** (0.010)	-0.056 (0.055)	-0.063 (0.056)	
	Second lag of IT						
Estimation Method		OLS	IV	OLS	IV	IV	
Lags included				Yes	Yes	Yes	
Controls Included						Yes	

Notes: Standard errors are in parentheses
 *Significant at 10% level **Significant at 5% level ***Significant at 1% level
 Coefficients report the effect of an extra hour of language instruction per week on language test scores. Scores are standardized with a mean of 0 and standard deviation of 1. IV uses a dummy variable as an instrument, which takes value 0 if the specific observation originates from before the policy change and 1 thereafter. The lags (which are equivalently instrumented) refer to the effect of the instruction time in language that the students of that group in that particular school experienced two years earlier. For a child in grade 4 in wave 3 this refers to the value of IT in grade 2 in wave 2 etc. The last column controls for parental education level, ethnicity, gender, month of birth, whether the child live in a two-parent household, and the number of children in class.

Table 4.6: Effect of instruction time (IT) for language on IQ test scores

Grade level	Expl. variable	I	II	III	IV	V	n
2nd grade	Instruction Time	0.022** (0.011)	0.457*** (0.109)	0.022** (0.011)	0.457*** (0.109)	0.496*** (0.127)	3634
4th grade	Instruction Time	0.050*** (0.012)	0.534*** (0.170)	0.050*** (0.012)	0.289*** (0.071)	0.293*** (0.070)	3365
	First lag of IT			-0.0069 (0.012)	0.225*** (0.057)	0.187*** (0.056)	
6th grade	Instruction Time	-0.020 (0.012)	0.446*** (0.114)	-0.0045 (0.014)	0.242** (0.101)	0.256*** (0.102)	3100
	First lag of IT			0.035*** (0.012)	0.219*** (0.078)	0.201*** (0.071)	
	Second lag of IT			-0.024** (0.010)	0.096 (0.111)	0.059 (0.107)	
Estimation Method		OLS	IV	OLS	IV	IV	
Lags included				Yes	Yes	Yes	
Controls Included						Yes	

Notes: Standard errors are in parentheses

*Significant at 10% level **Significant at 5% level ***Significant at 1% level

Coefficients report the effect of an extra hour of language instruction per week on IQ test scores. Scores are standardized with a mean of 0 and standard deviation of 1. For an explanation on the instrumentation and the control variables included, see table 4.5.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

Table 4.7: Effect of instruction time (IT) for language on math and environmental studies

Test score	Expl. variable	I	II	III	IV	V	n
Mathematics	Instruction Time	0.013 (0.017)	-0.058 (0.114)	0.00090 (0.020)	-0.107 (0.089)	-0.189* (0.099)	1675
	First Lag of IT			0.020 (0.017)	0.0067 (0.137)	-0.129 (0.179)	
	Second Lag of IT			0.025* (0.015)	0.171** (0.085)	0.224** (0.107)	
Environmental Studies	Instruction Time	-0.040** (0.020)	-0.013 (0.139)	-0.062*** (0.024)	-0.068 (0.080)	-0.082 (0.078)	1328
	First Lag of IT			0.030 (0.022)	0.0055 (0.149)	0.0012 (0.183)	
	Second Lag of IT			0.025 (0.016)	0.159* (0.088)	0.134 (0.083)	
Estimation Method		OLS	IV	OLS	IV	IV	
Lags included				Yes	Yes	Yes	
Controls Included						Yes	

Notes: Standard errors are in parentheses
 *Significant at 10% level **Significant at 5% level ***Significant at 1% level
 Coefficients report the effect of an extra hour of language instruction per week on the relevant test score. Both tests are only taken in 6th grade. Scores are standardized with a mean of 0 and standard deviation of 1. Environmental studies is a combination of biology, geography and history. For an explanation on the instrumentation and the control variables included, see table 4.5.

Table 4.8: Effect of instruction time on ranking: language scores (IV)

		Language score Lowest third	Language score Middle third	Language score Highest third	Language score All
2nd grade	Instruction Time	0.099*** (0.027)	0.168*** (0.039)	0.087*** (0.023)	0.118*** (0.026)
4th grade	Instruction Time	0.084*** (0.036)	0.051 (0.048)	0.0060 (0.029)	0.047* (0.028)
	First lag of IT	0.045 (0.028)	-0.021 (0.039)	-0.020 (0.024)	0.00091 (0.025)
6th grade	Instruction Time	0.033 (0.026)	0.045 (0.039)	0.060** (0.024)	0.046* (0.024)
	First lag of IT	-0.0039 (0.018)	-0.030 (0.026)	0.0069 (0.016)	-0.0094 (0.017)
	Second lag of IT	-0.011 (0.016)	-0.017 (0.024)	-0.017 (0.015)	-0.015 (0.016)

Notes: Standard errors are in parentheses

*Significant at 10% level **Significant at 5% level ***Significant at 1% level

Coefficients report the effect of an extra hour of language instruction per week on the ranking within the whole sample, using IV in combination with quantile regression on the rank in the distribution. Ranking is based on the fraction of the sample population that has a score that is equal to or lower than that particular individual. The exact ranking is 1 minus this fraction. The quantile regression model sets the score of those individuals outside of the relevant score range at either the lower or the higher threshold. The method initially divides the overall coefficients over all subgroups, hence we multiply coefficients by three to receive interpretable coefficients. For an explanation on the instrumentation, see table 4.5. No control variables are included.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

Table 4.9: Effect of instruction time on ranking: IQ (IV)

		IQ score Lowest third	IQ score Middle third	IQ score Highest third	IQ score All
2nd grade	Instruction Time	0.066*** (0.025)	0.177*** (0.039)	0.135*** (0.0026)	0.126*** (0.027)
4th grade	Instruction Time	0.096*** (0.033)	0.126*** (0.048)	0.030 (0.027)	0.084*** (0.031)
	First lag of IT	0.048 (0.029)	0.087 (0.042)	0.060** (0.024)	0.065** (0.028)
6th grade	Instruction Time	0.075** (0.030)	0.090** (0.042)	0.054** (0.024)	0.073*** (0.028)
	First lag of IT	0.081** (0.020)	0.096*** (0.028)	0.057*** (0.017)	0.078*** (0.019)
	Second lag of IT	0.019 (0.019)	0.0060 (0.0026)	-0.012 (0.015)	0.0035 (0.017)

Notes: Standard errors are in parentheses

*Significant at 10% level **Significant at 5% level ***Significant at 1% level

Coefficients report the effect of an extra hour of language instruction per week on the ranking within the whole sample, using IV in combination with quantile regression on the rank in the distribution. Ranking is based on the fraction of the sample population that has a score that is equal to or lower than that particular individual. The exact ranking is 1 minus this fraction. The quantile regression model sets the score of those individuals outside of the relevant score range at either the lower or the higher threshold. The method initially divides the overall coefficients over all subgroups, hence we multiply coefficients by three to receive interpretable coefficients. For an explanation on the instrumentation, see table 4.5. No control variables are included.

Table 4.10: Effect of instruction time on ranking: math and environmental studies (IV)

		Lowest third	Middle third	Highest third	All
Mathematics	Instruction Time	-0.045* (0.026)	-0.012 (0.039)	-0.0057 (0.024)	-0.021 (0.0025)
	First lag of IT	0.042 (0.042)	-0.048 (0.060)	-0.033 (0.036)	-0.013 (0.038)
	Second lag of IT	0.048* (0.026)	0.045 (0.036)	0.030 (0.023)	0.041* (0.0024)
Env. studies	Instruction Time	-0.033 (0.023)	-0.029 (0.036)	0.0084 (0.022)	-0.017 (0.023)
	First lag of IT	-0.014 (0.066)	0.014 (0.022)	0.016 (0.014)	0.0055 (0.043)
	Second lag of IT	0.039 (0.026)	0.072* (0.042)	0.029 (0.024)	0.047* (0.025)

Notes: Standard errors are in parentheses

*Significant at 10% level **Significant at 5% level ***Significant at 1% level

Coefficients report the effect of an extra hour of language instruction per week on the ranking within the whole sample, using IV in combination with quantile regression on the rank in the distribution. All results are for grade 6. Ranking is based on the fraction of the sample population that has a score that is equal to or lower than that particular individual. The exact ranking is 1 minus this fraction. ES stands for Environmental Studies. The quantile regression model sets the score of those individuals outside of the relevant score range at either the lower or the higher threshold. The method initially divides the overall coefficients over all subgroups, hence we multiply coefficients by three to receive interpretable coefficients. For an explanation on the instrumentation, see table 4.5. No control variables are included.

4. ALLOCATING INSTRUCTION TIME: HOW LANGUAGE INSTRUCTION CAN AFFECT MULTIPLE SKILLS

Table 4.11: Mean differences of school characteristics

Variable	Mean Wave 2	Mean Wave 3	Difference	P-value
Extra Teachers (in working hours)	3.597	3.638	0.041	0.832
Extra Resources	2.292	2.667	0.375	0.082
Extra Resources: devoted to Language stimulation of low-performing pupils	1.441	1.416	-0.025	0.749
Extra Resources: devoted to development of Language policies	1.358	1.302	-0.056	0.408
Nr of students per classroom	25.032	25.096	0.064	0.915
Use of end-of-year Targets	1.783	1.856	0.073	0.051
Use of intermediary Targets	1.771	1.753	-0.018	0.754
Homework in Language	2.356	2.405	0.048	0.703
Homework in Math	2.210	2.118	-0.092	0.423
Homework in Env. Studies	2.498	2.718	0.220	0.121
Individual tracking of vocabulary skills	2.032	2.095	0.063	0.767
Individual tracking of spelling skills	3.215	3.241	0.026	0.767
Frequency of testing	3.505	3.562	0.057	0.504
Ability-grouping in Language: At top	1.326	1.265	-0.061	0.164
Ability-grouping in Language: Overall	1.161	1.146	-0.014	0.676

Notes: Table shows mean differences in school-reported variables between waves 2 and 3. Answers to questions are reported on a scale: no, moderate, extensive etc.; or simply yes or no. 'Extra resources' question is measured on a scale of 1-5 (1 means nothing, 2 means under 1000 euro, 3 means [1000, 5000] etc). Hence, the values have no absolute meaning. The last column lists p-values of a two-sided Wald test conducted on the difference between means.

Table 4.12: Sensitivity to extracting trends (grade 6)

Test score	Expl. variable	Original Effect	Trend of -0.04 sd	Trend of +0.04 sd
Language	Instruction Time	0.187** (0.089)	0.239*** (0.091)	0.135 (0.090)
	First Lag of IT	-0.016 (0.061)	0.082 (0.063)	-0.113* (0.062)
	Second Lag of IT	-0.056 (0.055)	0.013 (0.057)	-0.125** (0.057)
Mathematics	Instruction Time	-0.107 (0.087)	-0.072 (0.095)	-0.142 (0.089)
	First Lag of IT	0.0067 (0.137)	0.160 (0.146)	-0.146 (0.138)
	Second Lag of IT	0.171** (0.086)	0.244*** (0.096)	0.098 (0.086)
IQ	Instruction Time	0.242** (0.101)	0.290*** (0.112)	0.194** (0.093)
	First Lag of IT	0.219*** (0.078)	0.295*** (0.087)	0.143** (0.072)
	Second Lag of IT	0.096 (0.111)	0.210* (0.124)	-0.018 (0.102)

Notes: Standard errors are in parentheses

*Significant at 10% level **Significant at 5% level ***Significant at 1% level

Coefficients report the effect of an extra hour of language instruction per week on the relevant test score (IV). The 'original effect' column shows the effects as they appear in tables 4.5 and 4.7. In the following column, we extract a linear negative trend in test scores of 0.04 of a standard deviation per two years. The last column does the same for a linear positive trend of 0.04. Based on other data sources, actual trends over this time period are approximated by -0.04 for math and language and +0.04 for IQ. The results pertaining to these 'appropriate' trends are in bold. For an explanation on the instrumentation, see table 4.5. No control variables are included.

The Labor Market Effects of Secondary School Track Assignment: A Regression Discontinuity Approach¹

5.1 Introduction

Heterogeneity in ability among students has strong implications for the optimal setup of the educational system. Different students may require different environments and different forms of instruction. Current educational systems generally split up students in different tracks according to ability. Tracking has become the most common method to deal with heterogeneity in ability among students. There are many ways in which tracking can be executed. Two main aspects are the age at which tracking takes place and the exact division of students to the available tracks. Past literature has generally focused on the former aspect. This chapter focuses on how students are allocated to school tracks. From an individual's point of view, the division of students across tracks is perfect when each student is sent to the track that maximizes his or her outcomes.

The aim of this chapter is to analyze whether students are sent to those tracks that maximize their individual outcomes. Parents generally like to see their children attend the higher of two tracks. But the question is whether attending this higher track will really be in the pupil's interest, or whether he or she will be better off in the environment of a lower track. We examine the Dutch educational system, where track assignment is based on scores on a selection test at the end of primary school. This test contains certain threshold scores, which indicate the required

¹This chapter is based on Borghans, Diris, Smits, and De Vries (2011).

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

ability level for a specific track. Thresholds are not fully enforced, but there is a strong discontinuity in the probability of track assignment around them. We estimate the difference in later-life outcomes that each track produces, by looking at discontinuities in outcomes around threshold scores. By using a fuzzy regression discontinuity design, we can exploit the discontinuity in treatment probability and thereby estimate which track would make people around this required ability level better off. We use data from Statistics Netherlands (CBS), which matches information from a Dutch secondary school study with job market information for the same individuals. We measure the effects of tracking on educational attainment and earnings. We find that students who are at the margin of the required ability level complete higher levels of educational attainment. Attending the higher track increases educational attainment by around 1 year for the lower threshold, and 1.5 years for the higher threshold. The return on this additional year of schooling in terms of long-run earnings is around 10% for the former and around 13% for the latter.

We define tracking as the selection of students into specific groups for all subjects within a school. The assignment to a track is based on a measure of overall ability. The literature employs different definitions of what constitutes tracking. It is often used interchangeably with ability-grouping or streaming, which group students for specific school subjects. Past literature has largely focused on the effect of tracking on inequality.¹ It has been argued that tracking reinforces differences in socioeconomic status, because it comes at the expense of lower-ability students. Our approach is different. We take the existence of tracking and its length as given and analyze the implications of track assignment for students that ‘should’ be indifferent between those tracks, given this system. If everyone goes to the track that maximizes his or her outcomes, given the assignment of the rest of the student population, this will result in a smooth system of outcomes along the distribution. If this is not the case, there will be discontinuities in outcomes around the thresholds, and individuals can be better off by switching tracks. This theoretical approach is described in Chapter 2.

The organization of this chapter is as follows. Section 5.2 gives a brief overview of relevant characteristics of the Dutch educational system. Section 5.3 presents a literature overview, in which we discuss previous research in the field. Section 5.4 discusses the data we use in the empirical analysis. Section 5.5 presents the methodology, while main results are discussed in section 5.6. Robustness analysis is discussed in section 5.7. Section 5.8 concludes.

¹See, e.g., Brunello and Checchi (2007), Figlio and Page (2002), Hanushek and Woessmann (2006).

5.2 Dutch educational system

5.2.1 Track assignment

The Dutch Educational System is characterized by relatively early tracking, namely at the age of 12. Students are selected into three main tracks: *vmbo* (vocational), *havo* (lower academic) and *vwo* (higher academic). They can still drop to a lower track in higher grades, in case of bad performance, but moving to higher tracks is only possible when the current track has been completed.¹ The current *vmbo* track used to consist of the separate *lbo* and *mavo* tracks. These tracks were joined in 1999, although there is still selection within *vmbo* between theoretical and practical education. Although these tracks were split for four out of five cohorts, we take them together when analyzing the effect of ‘*vmbo*’ versus *havo*, because they together represent one vocational track.² For the remainder of this chapter and the next, we refer to these tracks as T1 for the (lower) *vmbo* track (or T1a and T1b for *lbo* and *mavo*, respectively), T2 for the (intermediate) *havo* track and T3 for the (higher) *vwo* track.

The key determinant of track selection is the Cito test, which is employed by more than 80% of Dutch primary schools (Inspectie van het Onderwijs, 1997). The Cito score leads to an advice from the primary school for any of the three tracks. Students and their parents can decide to deviate from that advice.³ In the majority of cases the advice corresponding to the Cito score is adhered. There is additional room for students to be selected into a track that does not correspond to their Cito score, because selection is often postponed until the second or third year of secondary education. This occurs through the existence of comprehensive ‘bridge-grades’, where students of two or more tracks are still kept together. This is most common for Track 2 and Track 3 students, while students in Track 1 are generally selected early.⁴

We estimate the effect of being in a certain track for those students around the margin of two tracks, based on their score on the Cito-test. We estimate this for being assigned to Track 1, and being assigned to Track 3, both taken relative to being assigned in Track 2. Additionally, we split up Track 1 to estimate the effect of being assigned to Track 1a (relative to 1b). Although

¹Attending *havo* is only possible after completing the highest of four *vmbo* levels.

²Since we estimate effects at the margin, it should not be necessary to distinguish these tracks. It can only be an issue if they exhibit different slopes along the distribution, but our method of bandwidth selection takes this into account.

³If secondary schools allow for this, which is not always the case.

⁴In our sample, roughly 90% of T2 and T3 students is in a comprehensive grade for at least one year, and 35% for 2 years. About 75% of students in Track 1a and 1b is in a specific track at age 12 already, which contains almost all T1a students.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

this track does not formally exist anymore, its results can give an indication on whether it was beneficial to merge these two tracks, for those students around the margin.¹

5.2.2 Post-secondary education

The Dutch educational system has three levels of post-secondary education. The lowest level is *mbo*, which has a vocational orientation. *Mbo* has four main levels, which last one to four years. Higher education can be divided into two categories. *Hbo* provides higher professional education, while *wo* consist of university education. Students with a high school diploma from the T2 track or higher can enter *hbo*, while *wo* is only available for students who completed T3.² Hence, students who are selected to a specific T1 or T2 track lose the direct opportunity to attend *hbo* or *wo*, respectively. Completing the highest level of *mbo* makes one eligible for entering *hbo*. Hence, it is still possible for T1 students to complete higher education, although the route is less direct.

5.3 Literature

Previous literature on the effects of tracking generally focuses on its effects on inequality, often in relation to efficiency. We can distinguish studies that use across-country differences from those that use within-country variation in tracking over time. Cross-country studies generally make use of difference-in-difference estimation, by looking at differences in test score between countries, both before and after tracking has taken place. Hanushek and Woessmann (2006) use differences in tracking ages across countries and find that early tracking increases inequality, which is not compensated for by any gains in efficiencies. These studies make the rather strong assumption that differences would have remained stable when tracking would have occurred at the same age. Moreover, Jakubowski (2010) points out that there are differences in relative age levels between the studies Hanushek and Woessmann (2006) use, although he largely confirms their findings that tracking reduces mean performance and comes especially at the expense of students of lower ability. Ariga and Brunello (2007) contrast these findings. They use an IV-approach, employing reasons to drop out from school as an instrument, and find that the time spent in a track is positively related to performance on achievement tests.

¹This is only suggestive evidence, since the treatment of Track 1b is different than that from a joint T1 track.

²Completing *hbo* also gives access to university education.

A number of other studies has looked at variation in the design of a tracking system over time. Pekkarinen et al. (2009) use a change in the Finnish education system from multi-track to one track to assess the effect of tracking as a whole. The study uses variation in the exact timing of the policy change across municipalities to identify this impact. Pekkarinen et al. (2009) find no impact of tracking on several measures of educational attainment, test scores or income. Subgroup analysis showed that the policy change had beneficial effects for reading subscores and for those with low parental education. A similar policy change in England has been evaluated by Galindo-Rueda and Vignoles (2005). They find positive effects of tracking on high-ability students, and an absence of negative effects for the rest of the distribution. They use the political constituency of the area as an instrument, on the ground that Conservative areas are more likely to postpone conversion to the comprehensive system (which was done gradually across the country). However, Manning and Pischke (2006) question the validity of their IV approach, finding a similar impact with this method when they use test scores obtained before tracking, which naturally should be unaffected.

There has been special interest in whether tracking reinforces the effect of family background, thereby causing increased segregation and hampering equality of opportunity. Schnepf (2003) finds that family background, gender and degree of urbanization have a significant impact on the track someone ends up in, independent of ability. Waldinger (2007) finds that tracking reinforces the effects of family background on student performance, but he identifies that this is solely due to primary school selection and not exacerbated by secondary school tracking. This underlines the identification problem in estimating the impacts of (early) tracking. Possible reasons of negative effects of tracking, especially on students of lower ability, can occur for a number of reasons, including peer effects, teacher sorting and misallocation of students. As of yet, there is no clear consensus on which of these issues is the strongest contributor to any possible negative effects, since disentangling all these effects is troublesome.

In the literature, tracking is often used interchangeably with the more informal practice of ability-grouping, which refers to the splitting up of students according to ability, for certain school subjects. The overall findings do not differ strongly across definitions, although they appear to be slightly more favorable for ability-grouping. Brunello and Checchi (2007) look at long-term impacts and find that tracking has negative impacts on the intergenerational mobility of differences in educational attainment and early labor market outcomes, but positive impacts for literacy and participation in on-the-job training. Moreover, studies have identified positive

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

effects for high-ability students (Epple and Romano, 2002) and an absence of significant effects on inequality.¹

The focus of our study is on efficiency rather than equity.² Where other papers look at changes in the degree of tracking, either across countries or over time, our study is, to our knowledge, the first to estimate the effect of being in one track versus another from an individual's point of view. We estimate this treatment effect for those students who are at the margin of the required ability level for a certain track. This information is especially relevant for students and their parents who face this choice. Another unique point of our study is that we use a regression discontinuity approach to deal with the fact that track assignment is endogenous. Additionally, where other studies largely look at tracking in relation to test scores, we use longitudinal data that contain educational attainment and labor market information up to age 42.

5.4 Data

5.4.1 The sample

We employ data from the national database of Statistics Netherlands (CBS). The dataset contains a post-primary education study (VOCL) combined with administrative data, including demographic information and labor-market data. The individuals that participated in VOCL are retrieved in the administrative data. The available data include test scores at the start of secondary education and two years later, information on the educational career of each individual (secondary education track, years of schooling, highest degree etc.) and labor market status from 1999 until 2007 (employment status, earnings, etc.). Hence, we are able to examine what the effect of track assignment is on educational attainment and earnings after multiple years. Data are available for the VOCL cohorts of 1977, 1982, 1989, 1993 and 1999. The cohorts are named after the year in which the included individuals entered secondary education. This is when the VOCL study started, and students are 12 years old. Hence, we have labor market information ranging from those who just started to work up until people who are 42 years old (the 1977 sample in 2007). The 1989 and 1993 cohorts contain test scores from year three of secondary school. However, the exact tests differed across tracks which is why we do not employ them as outcome variables. Summary statistics are provided in table 5.1.

¹See e.g. Slavin (1990), Figlio and Page (2002).

²One could analyze whether there is a certain assignment of students to tracks that minimizes inequality, but is not possible to identify inequality by only looking at the margin, as in our RD design. One would have to look at changes in the assignment in tracking over time, which is beyond the scope of our setup.

The 1999 dataset includes individuals who are largely too young to have entered the labor market in any of the included years. Similarly, some individuals from the 1993 and 1989 cohorts have not entered the labor market yet in some of the years for which we have earnings data. This holds disproportionately more for those with higher degrees and therefore disfavors assignment to a higher track. Thus, these results should be interpreted in a different way than the results of ‘older’ cohorts. Moreover, the pattern of earnings over test scores could become more volatile and therefore more difficult to fit in a control function for these ages, which has some consequences for the viability of our RD design. We still report results for these ages, but they are less robust and should be interpreted with care.¹

We stated that assignment into tracks is largely based on an end-of-primary-school test; the Cito test. The specific results of this test are only available for the 1999 cohort. For the other cohorts, a very similar test is available which is taken at the very beginning of secondary education. This test serves as a proxy for the actual Cito test. We label this test as the *Entrance Test*, since it is taken just after entering secondary education. The density of test scores for the Entrance Test can be seen in figure 5.2. Although the use of the Entrance Test can lead to more disturbance than the actual Cito test, as long as it shows a strong enough discontinuity in treatment over its distribution, it is suitable for employing a fuzzy RD design. Moreover, since the 1999 cohort includes both tests, we can test how the ‘proxy test’ performs against the actual test in predicting the treatment effects of a certain track on 3rd-year test scores for this cohort.²

5.4.2 Graphical analysis

Figures 5.6 to 5.7 plot the development of our outcome variables over scores on the Entrance Test. The figures portray educational attainment for all cohorts, and earnings for the four most recent years (2007 to 2004) for the two oldest cohorts. The patterns are highly similar for earnings in 2001 to 2003, and for earnings in 1989. The 1993 cohort is naturally less smooth, as many individuals are still in education in the period from 2001 to 2007. Both the patterns for earnings (in logarithms) and educational attainment appear to be (log-)linear. There are minor jumps in mean outcomes at some points, but given that our design inhibits relatively high fuzzyness, we cannot draw any strong conclusions regarding treatment effects yet. The strong

¹The data show that the age of 26 is where the relationship between test score and earnings starts to become log-linear and where the fraction of people still in education is low. Hence, this is a good starting age from which on results are robust.

²Although test scores are not suited as outcome variables in the main analysis, they are still appropriate to use when we compare both instruments.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

linearity in the outcome patterns suggests that it is feasible to employ relatively large and largely linear bandwidths in our estimation. The selection of bandwidths for the actual analysis will be discussed in section 5.5.5.

5.5 Methodology

5.5.1 The regression discontinuity design

In this chapter, we estimate the effect of being assigned to a certain track. Tracks represent different treatments. Since treatment is not random, this does not represent a natural experiment. We can use a regression discontinuity design (RDD) instead. Assignment to treatment is partly determined by the value of a forcing variable, which the Entrance Test score. We compare individuals that are just below and just above a threshold level for being advised to attend a particular track. These individuals should be similar; the main difference being that one group is treated while the other is not (Imbens and Kalyanaraman, 2008).

The forcing variable we use is the Entrance Test score. This test has certain threshold scores that indicate if someone is eligible for a certain track or not. It does not contain a strict threshold between treatment and non-treatment. Hence, we cannot simply compare students above and below the threshold, because there is overlap between treated and non-treated observations. The theory on the RD design makes a distinction between a sharp regression discontinuity (SRD) and a fuzzy regression discontinuity (FRD) design. The former requires a sharp distinction between treatment and non-treatment based on a forcing variable. In the FRD design, the probability of treatment does not need to change directly from 0 to 1 at the threshold value of the forcing variable. The design allows for a smaller jump in the probability of treatment, as long as there is a strong enough discontinuity. Hahn et al. (2001) showed that the interpretation of the FRD estimator is equivalent to an IV estimator. One can employ this design by using eligibility of treatment as an instrument. Eligibility indicates whether an individual should have been assigned to treatment based on the forcing variable. The FRD allows for a correlation between treatment and unobservables, as long as there is no jump in the values of such unobservables at the cutoff point.¹

¹Van der Klaauw (2002) advocates using a propensity score, as an alternative instrument. This propensity score is based on a certain ability indicator, such as a test score or GPA average. We estimated treatment effects with this instrument as well, which gave very similar estimates. These results are available on request.

5.5.2 Predictive power

The FRD requires that the forcing variable is a strong enough predictor of treatment status and continues a discontinuity in treatment probability around the threshold. The predictive power of the Cito test is limited because deviations from the Cito advice are possible¹ and selection can be postponed. We calculate the predictive power from test score to final track assignment.² The fact that we use the Entrance test instead of the actual Cito score involves a small loss in predictive power. Where the translation from Cito score to advice is 91%, 77% and 71% for T1, T2 and T3, respectively, this is 78%, 68% and 61% for the Entrance Test score. When we look at the total predictive power of score to selection, we see that 81-86% of those that are eligible for track T1 also end up in T1. For T2 this is 30-35% and for T3 this is 70-75%. The lower predictive power for the small T2 track is mainly the result of upstream. About 30-35% of students with a T2-advice end up in Track 3 (only 20% fall down to T1), while 55-65% of those with a mixed T2-T3 advice go to T3. This suggests that advice towards T3 is conservative. The loss in predictive power due to deviations from test score to advice and from postponed selection are of roughly equal size. Figure 5.1 shows the division of students to all tracks by their Entrance Test score for the 1977 cohort.

Figures 5.3 to 5.5 show how the probability of treatment develops if we plot it against Entrance Test scores. Track T1 shows a smooth pattern with a discontinuity halfway the distribution. The picture is very similar for the probability of T3 treatment. The pattern for T1a will require a wider bandwidth to still give enough first stage power. The first stage for estimating the effect of Track 1a is strong enough for all cohorts when we use the full bandwidth, but not anymore for the 1977 cohort when we employ optimal bandwidths,³ as table 5.3 shows. Results would give very high standard errors, and such weak instrumentation can lead to several problems (Bound et al., 1995). This is why we do not estimate the effect of T1a treatment for this cohort. This also holds for estimation of T1 treatment in the 1993 cohort. For all other cohorts and treatment effects, the F-values are sufficiently large and estimation is appropriate.⁴

¹This can occur either through deviations from the test score to the formal advice given by the primary school or deviation from the formal advice to selection in year 1 of secondary school.

²This represents the strength of our instrument in the Fuzzy RD Design we apply.

³These bandwidths and their corresponding specifications are discussed in section 5.5.5.

⁴The bandwidths and corresponding F-values are larger when we estimate treatment effects on years of schooling, for T1 and T3. For T1a treatment, the first stage is too weak in any cohort for estimation of effects on educational attainment.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

5.5.3 The IV-model

We employ a two-stage model, where we use treatment eligibility (labeled w) as a predictor for actual treatment in the first stage and estimate the effect of (predicted) treatment (θ) on our outcome variable in the second stage. In our context, eligibility does not depend on a formal cutoff point, since track assignment is diffuse across a small range of test score values. Therefore, we take that Entrance Test score for which the first stage in our 2SLS model is strongest.¹ This represents the informal cutoff point for treatment eligibility. We will show whether small variation in the choice of this cutoff point changes our results. We control for the correlation between test scores and earnings on both sides of the cutoff, because the treatment effect could otherwise pick up on differences in the slope of these trends. This trend is estimated at the point where the threshold between the two tracks lies. Again, because there is no formal threshold, we choose the informal threshold, where the discontinuity in treatment is strongest.

One might consider to employ different control functions for all tracks. However, our method of selecting bandwidths ensures that larger bandwidths are rejected when there are strong differences in the slope of any two tracks. Another alternative would be to cancel out the track that lies at the other extreme. However, that could limit both the number of observations and the strength of the first stage considerably. We consider all these options as possible choices in our control function (including higher polynomials in each case) and let the cross-validation procedure that determines our optimal bandwidth decide which one is optimal in every separate treatment estimation. We will show if employing any of the alternatives leads to strong changes in our estimates.

Hence, our model becomes, in the simplest form (two linear trends):

$$\begin{aligned}\theta_i &= \beta_0 + \beta_1 w_i + \epsilon_i \\ Y_i &= \beta_0 + \beta_1 \theta_i + \beta_2 S_i + \beta_3 (S_i - \bar{s}) \theta_i + \epsilon_i\end{aligned}\tag{5.1}$$

This model is applied to three different treatments (T1, T3 and T1a treatment). Outcomes are measures of educational attainment and earnings. Score refers to the score on the Entrance test, which is our forcing variable. We deduct the threshold value $\bar{s}$ from the test score when

¹This is essentially the test score where the discontinuity is at its strongest. In practice, this corresponds to the point where the chance of being treated passes the 50% point.

we estimate the part of the control function that corresponds to the treatment. w is a dummy variable indicating if the specific test score makes the student eligible for the treatment or not.

5.5.4 The sample multiplier

The RDD is a quasi-experimental design. Due to collinearity between treatment status and the forcing variable, estimates are less precise than in a controlled experiment, given the same sample size (Goldberger, 1972). We need a larger sample size to achieve the same precision. The sample multiplier for a sharp RD design compared to an random assignment design is in between 2.75 and 4, depending on whether the distribution is normal (2.75) or uniform (4), or anything in between. This multiplier, or design effect, is larger for fuzzy RD designs. It is usually in the range of 3 to 5, but can be much larger for designs with high fuzzyness (Schochet, 2009). The exact multiplier depends on the cutoff point, the number of ‘compliers’ and ‘crossovers’ and the degree of correlation between forcing variable and treatment status. The sample multiplier is very relevant for this study, since the fuzzyness of our design is relatively high. The formula for calculating the design effect is shown below.

$$FuzzyDesignEffect = \frac{(q_t - q_c)^2}{(1 - \rho_{ts}^2)(p_t - p_c)^2}$$

where $(q_t - q_c)$ represents the treatment-control difference in service receipt rates for the full study population and $(p_t - p_c)$ for those just at the threshold. ρ_{ts} captures the correlation between treatment and the forcing variable.

Our multiplier is around 50, on average.¹ This number is lowest for Track 3 treatment and slightly higher for Tracks T1 and T1a.² The effect is also stronger for the 1977 cohort than in the other cohorts.³ High multipliers require a large number of observations to still ensure imprecision. Our final analysis contains 8,000 to 15,000 observations per cohort. This means that the precision of our estimates using a fuzzy regression discontinuity design corresponds to that of a controlled experiment with 200 to 300 observations.

¹The high multiplier is mainly due to the relatively small jump in mean treatment status at the exact threshold. Through the quadratic term in the denominator, this has a strong impact on the multiplier.

²Strong adherence to the test score for the sample as a whole leads to a high nominator for the former. Track 1a exhibits stronger collinearity between score and treatment status.

³The multiplier is around 70 in 1977, largely because of the larger range of the test score. It is around 40 for the other cohorts. The larger size of the 1977 dataset compensates for the higher multiplier there.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

5.5.5 Bandwidth

The exact bandwidth we employ in our estimation is important, because our design needs a large enough width to have a strong enough first stage and a large number of observations to compensate high fuzziness. Bandwidth choice represents a tradeoff between accuracy and power. Smaller bandwidths are more accurate close to the threshold while larger bandwidths imply a higher number of observations (Lee and Lemieux, 2010). We employ a cross-validation (CV) procedure to find the bandwidth for which this tradeoff is most favorable. This CV procedure checks which specific bandwidth minimizes the mean squared error of our control function (Imbens and Kalyanaraman, 2008). Theoretically, one should do this separately for both the first and second stage. In practice, employing different bandwidths for each stage is not advisable. The literature suggests that it is best to employ the smallest of the two suggested bandwidths when they differ (Imbens and Lemieux, 2007). The first stage is virtually indifferent between any bandwidth and our choice of bandwidth is fully driven by the second stage.

The bandwidth for the second stage can be limited if there are strong non-linearities in the relation between the forcing variable and the outcome variable. Figures 5.6 to 5.7 show that most outcome variables are linear for a large part of the distribution. However, there exists non-linearity at the extremes. This is most prominent for the upper part of the earnings distribution in the 1977 cohort. Figure 5.8 shows a linear prediction for the pattern of earnings in 2005 for the 1977 cohort. This confirms that a linear control function is a good fit for a large part of the distribution, but is imprecise at the upper end.¹ We execute the CV procedure for different specifications to see which function provides the best fit. Mean squared errors for linear functions are shown in figure 5.9, where the MSE of the smallest bandwidth is set to 0. The suggested bandwidths are smaller for T1 relative to T3 and for the 1977 cohort relative to the other two cohorts. For the 1977 cohort, the MSE is ever increasing and the very smallest bandwidth is the optimal one. This suggests strong non-linearity and makes the use of a linear control function not advisable. This also applies, to a lesser extent, for Track T1 in 1982. In general, the latter part of the trend exhibits a different slope. This difference in slope is more relevant when estimating T1 treatment, since we estimate the control function of the T3 track separately when estimating treatment effects for T3.

We analyze whether alternative specifications perform better. We include track-specific trends, leave out the track at the other extreme, include a quadratic term in the control func-

¹The frequency of high scores is low, which also adds to the large variation at the top.

tion, or use a combination of these. We use the control function that gives the lowest MSE for its optimal bandwidth. These results suggest: including a quadratic term in the control function for T3 estimation in 1977, dropping track T3 for T1 estimation in 1977 and introducing track-specific trends for T1 estimation in 1982. Linear trends are applied for T3 estimation in 1982 and both estimations in 1989. For track 1a, quadratic trends (after the threshold) are suggested for the 1977 and 1982 cohorts and a linear trend for 1989.¹ We will show sensitivity to changes in both the specification and the formally suggested bandwidths.

In the cross-validation procedure, we use the earnings from 2005 as a benchmark for all earnings outcomes for a specific cohort,² since it was not considered worthwhile to apply separate CV procedures and bandwidths for every single earnings variable. As a robustness check, we execute the same procedure for earnings in 2007, 2006 and 2004 for the 1977, 1982 and 1989 cohort. This gives almost identical bandwidths. This is not surprising, since the pattern of earnings over test scores is very similar (see figure 5.6).³

5.6 Results

We estimate the effect of being assigned to a T1 track (versus being in a higher track) and of being in a T3 track (versus being in a lower track) for students at the margin of the required ability level for those tracks. Our outcome variables are earnings and educational attainment. In addition, we look at the effect of track assignment for the students at the margin of the recently merged tracks of *lbo* (Track 1a) and *mavo* (Track 1b). We report results on both a sharp regression discontinuity design (OLS) and a fuzzy regression discontinuity design (IV). The former serves a comparative purpose. Our IV results use treatment eligibility as an instrument.

Table 5.4 reports results of T1- and T3-treatment on earnings for the OLS model. Tables 5.5 to 5.7 shows results for the IV estimation. One would expect the results from the OLS model to be biased in the direction of the upper track since the students in the higher track are expected to have better characteristics that are relevant for success in later life, conditional on Entrance Test

¹Bandwidths are smaller for this treatment.

²We use 2005 earnings, since earnings in 2006 and 2007 are calculated in the administrative records in a slightly different way (they include some welfare benefits as well), hence 2005 provides the most recent earnings data that are consistent with the majority of other reported earnings.

³The only deviation occurred for earnings in 2007 for T1 treatment, which is why a different (wider) bandwidth is applied there.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

scores. This is exactly what we see if we globally compare these tables.¹

5.6.1 Educational attainment

The first rows of tables 5.5 and 5.6 show how track assignment affects educational attainment. Higher tracks lead to more education, at both thresholds. This effect is higher for younger cohorts and centers around 1 to 1.5 years for T1 and 1.5 to 2 years for T3. In both tables, there is a significant jump in the treatment effect when comparing cohorts 1982 and 1989. Table 5.1 shows that educational attainment increased in this time frame as well, but not more so than between 1977 and 1982.² Between those two cohorts, the effect on educational attainment is stable. Hence, students on the right side of either threshold have had a more than proportionate increase in educational attainment in the 1980s.³ The first stage is not powerful enough to estimate the effects of T1a assignment on years of schooling.

5.6.2 Earnings

The other rows of tables 5.5 to 5.7 show the effects of track assignment on monthly earnings.⁴ The effect is negative for T1 and positive for T3, which mimics the results for educational attainment. The effects gradually increase with age. For those above the age of 30, they range from around 10% for Track 1 and 10 to 15% for Track 3. Estimates are very consistent. They naturally progress if we move from cohort '77 (ages 36-42) to cohort '82 (ages 37-31) to cohort '89 (ages 30-24). The treatment effects for T3 increase more strongly with age than those for T1. This can reflect the higher increase in educational attainment there. We also report results for the very youngest cohort, who are aged 20-26 from 2001 to 2007, for Track 3 treatment.⁵ This column shows that effects are (insignificantly) negative for the higher track in the early 20s, which is likely caused by a longer educational life when assigned to Track 3. Table 5.7 reports

¹On average, the difference is not very large. This confirms preliminary analysis (not shown) which indicated that the correlation between treatment and observables is low, when we control for the Entrance Test. This provides some indication that track assignment is not strongly based on other variables that correlate with our outcomes, conditional on the Entrance Test score.

²Selection became more strict between 1982 and 1989, but this change was also of the same magnitude as that between 1977 and 1982.

³We would have to speculate about the underlining reasons for this observation, but it could be that the worldwide increase in the return on skills in the 1980s has especially induced those students who are the least-able among those eligible for higher education, to pursue a higher degree.

⁴Earnings are corrected for part-time factor, topcoded at 10,000 euro's per month, and those who earn less than a minimum wage are excluded. Results are higher when we use raw reported earnings, but not by a large margin.

⁵Instrumentation is too weak for T1 treatment, largely because of a small optimal bandwidth.

result for track 1a. Again, results indicate that the higher track leads to higher earnings for the marginal student. Assignment to track 1a reduces earnings by 7 to 14% in the 1982 cohort, while the treatment effect is non-existent for the younger 1989 and 1993 cohorts.

Combining all estimates, we find that one extra year of education has a return of 10% on monthly earnings for T1. For T3, the payoff from 1.5 years of extra education is 9-14%, depending on the age group one applies. The fact that attending a higher track increases educational attainment is not surprising, since it gives students a more direct path towards forms of higher education. The labor market results show that students around the margin also experience a payoff from completing higher degrees. The return to schooling is generally assumed in the literature to be around 8 or 9%, which is largely based on Angrist and Krueger (1991)'s seminal paper and overviews such as those by Card (1999) and Harmon et al. (2003). Since these effects apply to individuals who are at the bottom of the distribution for their track and would not have taken up extra education had they been selected to another track, this is a more than reasonable return. However, we cannot strictly say whether the higher tracks makes students around the margin better off. Higher wages come at the expense of extra time spent in education. Many individuals do not pursue higher degrees even when expected returns are high. It will depend on the individual's utility function whether the estimated effects will be perceived as a gain or a loss.

5.7 Robustness

5.7.1 Sensitivity to bandwidth

Table 5.8 shows how sensitive our results are to modifications in the bandwidth choice. We show this for the 1982 cohort, since this is the most recent cohort where people have all settled in the labor market for the full range of earnings outcomes. Sensitivity for the 1977 and 1989 cohorts is of the same magnitude.¹ The first column shows the original results, estimated with the bandwidth suggested by the cross-validation Procedure. The next columns show the effects of a wide range of arbitrary but strong changes in that bandwidth. Changes are modest for the majority of results and signs are consistent. Especially results for T3 treatment are extremely robust to even very strong changes in bandwidth. The only strong variation occurs with very narrow bandwidths, when the first stage is not strong enough anymore. The effect for Years of

¹It is slightly lower for both cohorts, mainly because a larger number of observations and a stronger first stage provides lower standard errors, on average. These results are available on request.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Schooling for T1 treatment is vulnerable to expansion of the bandwidth at the top (from 22-48 to 22-60). However, such an effect was already suggested by the development of the MSE in our bandwidth estimation, which expanded rapidly for higher ranges at the top. This indicates strong non-linearity at the top, which is picked up by the coefficient for that bandwidth.

The threshold score we apply as the division point between two tracks is implicit. Although we believe that this implicit threshold is at the proper place, we still show the consequences of changes in their value. The last two columns of table 5.8 shows the consequences of either lowering or raising the score by a value of two. Again, results are extremely similar when we do this for T3,¹ but the sensitivity for T1 is relatively high. However, a shift in threshold score by two is large,² which is amplified for Track T1 by the fact that the frequency per test score on this part of the distribution is very high. Moreover, the original threshold was chosen where the discontinuity was strongest and deemed to be more proper than the ones used in the last columns of table 5.8. Hence, the estimates for T1 are more sensitive than those for T3, but are consistently negative.

5.7.2 Control function

We use several control function specifications, based on which one is optimal according to the cross-validation procedure. We analyze how sensitive our results are to using different specifications. Tables 5.9 and 5.10 show results for Track 1 and Track 3 treatment when we use track-specific control functions and when we drop tracks at the other end of the distribution. The former shows that using such track-specific trends increases our estimates somewhat for the 1977 and 1982 cohorts. A similar picture occurs in table 5.10, although sensitivity is lower here, and virtually absent for Track 3 in 1982 and 1989. These results could hint at a small downward effect from our approach of taking all tracks on one side of the threshold together. However, our CV-procedure recommended these specifications as the most accurate ones, so it is more likely that the initial results are more proper. Nonetheless, the bias is relatively small and our conclusions, in terms of positive estimates for the marginal student from going to the higher track, are not affected in any way.

¹The difference is somewhat higher if we decrease the threshold by 2, but standard errors double in that case, indicating that the threshold is at a place where the discontinuity in treatment is much weaker, leading to more imprecise results.

²The vast majority of T3 students are in the test score range from 48 to 55, and this range is even more narrow for the small T2 track.

We also conduct a specification where we instrument the control functions as well. We use the product of our instrument and the test score as an instrument. The instrumentation is too weak when we use eligibility as instrument, but is feasible when we use the method proposed in Van der Klaauw (2002), using a propensity score based on the Entrance Test. Results are portrayed in table 5.11. The conclusions are qualitatively the same, although such a specification might require more power than can be provided in this design, given the high standard errors and volatility of coefficients.

Additionally, we check whether our results are robust to the inclusion of controls. A strong influence of adding controls could indicate the existence of biased estimates and hence serves as a good robustness check (Hahn et al., 2001). A comparison is provided in table 5.12. Our controls are month of birth, gender, ethnicity, whether the student lives in one of the big four cities in the Netherlands (Amsterdam, Rotterdam, The Hague or Utrecht), parent education and social status.¹ On average, estimates are about 1 to 2 percentage points lower, with no strong swings for any specific coefficient.²

5.7.3 Cito vs. Entrance Test

The 1999 cohort contains results from the actual Cito test made at the end of primary education, in addition to the unofficial Entrance Test score. Hence, we can compare results when we use either test as our forcing variable. This allows us to see the implications of using a proxy test in stead of the genuine forcing variable. The earnings and educational attainment data are not available for this cohort, since they are only 26 years old in 2007. We employ test scores in year three as outcome variables in stead. Although these score are no reliable outcome measures because the tests differ slightly by track, they can still serve a comparative purpose here. Table 5.13 shows this comparison. Coefficients are extremely similar if we compare the two forcing variables used. Figures 5.10 and 5.11 show the probability of treatment patterns for both forcing variables. The figures illustrate that the difference in predictive power between both tests is small as well.

¹For parent education, we add one dummy variable indicating whether the highest educational attainment of the parents is in the highest category, and an equivalent dummy variable for the lowest category. Two similar dummy variables are added for the lowest and highest social status categories.

²The strongest change in any coefficient is 3.1 percent, for T1 treatment on earnings in 2007, for the 1977 cohort.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

5.7.4 Peer effects

Our results indicate that assignment to a higher track increases the educational attainment of students around the margin, which also provides substantial returns in the labor market. These returns can originate from more than these extra years of education. Students might effectively learn more in the higher tracks, through a better-adjusted curriculum, better teachers,¹ or higher peer quality. It is not possible with the data at hand to disentangle all these effects. We employ rough analyses to get an indication whether peer effects are a large explanatory factor for our estimates. First, we check whether adding mean peer ability as a control has strong impacts on the coefficients. If it does, this is a strong indication that peer effects are a large part of the treatment effect. We use the average score on the Entrance Test of all classmates as a proxy for average peer ability. Coefficients are only lowered by 1 to 2 percent when we do this, with the exception of the effect of T1a in cohort 1982, where the change is 6 percent. This suggests that the role of peer effects is not overwhelming. Another option is to check the correlation between peer level and earnings. We find that when the average peer rank (defined from 1 to 100) increases by 10, earnings increase by 1.5 percent. This effect is slightly higher when we restrict to low scoring individuals (1.7 percent) and lower for the highest scoring students (0.9 percent).² The difference in average rank for those at the margin is 22 for T1, 8 for T3 and 22 for T1a. This implies a predicted effect on earnings for the marginal student of 3.3 percent, 0.7-1.2 percent, or 3.3-3.7 percent,³ respectively. One could see these estimates of peer effects as upper bounds, since they pick up on one's own ability as well. However, Robertson and Symons (2003) use an IV-approach that produces larger estimates than with OLS, indicating that attenuation bias exceeds ability bias. Their IV-estimates are a factor 1.3 larger than OLS results, which would imply that the contribution of peer effects to our treatment effects is around 4 percent for the lower two thresholds and 1 to 1.5 percent for the higher threshold. These are substantial parts of the total effect for T1a and T1 (although they do not comprise the whole effect), but are only a marginal part of the effect for T3. This indicates that the effects we find occur through more than just higher peer quality.

¹Betts and Shkolnik (2000) provide evidence that teacher experience and education differ across ability groups in high school education in the U.S. However, no clear impact of these teacher dimensions has been found in the literature. Summers and Wolfe (1977) even provide evidence that teacher experience has a negative impact on low-ability students.

²This is consistent with recent evidence from Sund (2009), who finds that the magnitude of peer effects decreases for higher ability levels.

³The range is based on whether we use the correlation between earnings and peer effects for the sample as a whole (0.0015), for only students in T3 (0.00093), or for only students in T1a (0.0017).

5.8 Conclusion

In this study, we show that one can evaluate whether students are assigned to their most optimal school track, by looking at discontinuities in outcome variables around the thresholds. Discontinuities suggest that students on one side of the threshold can be better off by switching tracks. We employ a fuzzy regression discontinuity design, using eligibility of treatment based on a selection test as an instrument for actual treatment. Thereby, we estimate the effect of being in a specific track for those students around the margin of the required ability level for that track. Our results indicate that students in the Netherlands who are at this margin achieve higher educational attainment in the higher track. This increase is around 1 year when moving from the low to the medium track and around 1.5 years when moving from medium to high track. The return on this extra education in terms of earnings is around 10% in both cases. The long-term effects are slightly higher for the higher threshold, which reflects the higher increase in educational attainment there. Findings are consistent across years and cohorts. These are substantial returns, which might indicate that these students are not only induced to study longer, but effectively learn more in the higher track as well. A crude analysis on the role of peer effects suggests that these are part of the total treatment effect, but the size of their contribution is not overwhelming, especially for the higher threshold.

Our design inhibits relatively high fuzziness. We ensure high precision through a very extensive dataset (10,000 observations per cohort), precise estimation of the control function and a powerful first stage. We use a cross-validation procedure to select the optimal bandwidth and control function specification in every estimation. Our results are robust to changes in bandwidth, using difference specifications for the control function and adding controls. The results for Track 3 are extremely rigid with respect to any of these changes. The estimates for Track 1 treatment are slightly more sensitive, although they are consistently negative in any specification.

The results indicate that there are certain benefits for students who are just short of the required ability level when thresholds would become more lenient. Thresholds in the Netherlands are not strictly enforced. It is often up to the parents and the child whether they want to enroll their child in the higher track. This chapter indicates doing so can result in substantial returns in the labor market, at the expense of extra time spent in education. This does not imply that thresholds should be moved from the perspective of society. A lower threshold can negatively

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

affect the untreated, because peer quality will be reduced in tracks on both sides of the thresholds. Future research can focus on establishing the optimal assignment of students to tracks from the perspective of society as a whole, for example by looking at (exogenous) variation in assignment over time.

5.9 Figures and tables

Figure 5.1: Track assignment by test score

Notes: Graph shows the fraction of students assigned to a certain track for each score on the Entrance Test, for the 1977 cohort. The lower two graphs split up the T1 track (*vmbo*) into T1a (*lbo*) and T1b (*mavo*).

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Figure 5.2: Density of forcing variable

Notes: Graph shows the density of the Entrance Test score, which serves as forcing variable in our regression discontinuity design.

Figure 5.3: Fraction of students assigned to T1 (*vmbo*) by test score

Notes: Graph shows the fraction of students with a certain score on the Entrance Test that are assigned to the *vmbo* track (T1), for all cohorts.

Figure 5.5: Fraction of students assigned to T1a (*lbo*) by test score

Notes: Graph shows the fraction of students with a certain score on the Entrance Test that are assigned to the *lbo* track (T1a), for all cohorts.

Figure 5.4: Fraction of students assigned to T3 (*vwo*) by test score

Notes: Graph shows the fraction of students with a certain score on the Entrance Test that are assigned to the *vwo* track (T3), for all cohorts.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Figure 5.6: Mean earnings by test score: 2007-2004

Notes: Graphs show the log of earnings for every test score on the first-year Entrance Test, for earnings in 2007 until 2004, for both the 1977 and 1982 cohort.

Figure 5.7: Mean years of schooling by test score

Notes: Graphs show the mean number of years of schooling for every test score on the first-year Entrance Test, for all cohorts.

Figure 5.8: Linear prediction of earnings pattern

Notes: Graph shows the pattern of the log of mean earnings over first-year Entrance Test scores, in combination with a linear prediction over that same pattern. The portrayed pattern is for

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Figure 5.9: Mean squared errors

Notes: Graph shows the mean squared error (MSE) for upper and lower bandwidths, for all tracks, using a linear specification for the control function on either side of the threshold. MSE's are shown for both the 1977 and 1982 cohorts. Values are recalculated to put the lowest bandwidth on 0, for comparative purposes.

Figure 5.10: Probability of treatment: Cito

Notes: Graph shows the fraction of students with a certain Cito test score that are assigned to the T1 track, for the 1999 cohort.

Figure 5.11: Probability of treatment: Entrance Test

Notes: Graph shows the fraction of students with a certain score on the Entrance Test that are assigned to the T1 track, for the 1999 cohort.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Table 5.1: Summary statistics

	1977		1982		1989		1993	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Track 1a (<i>lbo</i>)	0.256	0.436	0.323	0.468	0.333	0.471	0.359	0.480
Track 1b (<i>mavo</i>)	0.422	0.494	0.367	0.482	0.381	0.486	0.323	0.468
Track 1 (<i>vmbo</i>)	0.678	0.467	0.690	0.463	0.714	0.452	0.682	0.466
Track 2 (<i>havo</i>)	0.140	0.347	0.142	0.349	0.135	0.341	0.152	0.359
Track 3 (<i>vwo</i>)	0.182	0.386	0.168	0.374	0.151	0.358	0.165	0.371
Entrance Test	42.17	12.11	33.32	10.17	34.23	11.33	35.20	11.42
Years of Schooling	12.27	3.29	12.64	3.54	13.05	3.83	12.99	3.84
Earnings 2007	3678.99	1792.22	3421.78	1552.66	2869.22	1065.80	2455.26	847.81
Earnings 2006	3590.35	1770.67	3270.15	1465.56	2696.38	999.41	2266.34	807.70
Earnings 2005	3237.45	1588.49	2895.01	1208.01	2426.07	811.36	2044.20	745.81
Earnings 2004	3147.07	1506.91	2787.86	1122.65	2316.85	738.55	1915.78	686.53
Earnings 2003	3067.49	1421.29	2689.48	1037.65	2209.92	683.26	1763.86	654.31
Earnings 2002	2947.18	1351.11	2561.65	986.39	2044.52	639.20	1544.61	614.55
Earnings 2001	2805.14	1284.92	2412.03	883.82	1868.88	628.76	1314.24	594.49
Female	0.503	0.500	0.504	0.500	0.481	0.500	0.485	0.500
Big 4 Cities	0.091	0.288	0.055	0.228	0.045	0.207	0.060	0.238
Non-Dutch	0.079	0.265	0.086	0.275	0.109	0.311	0.090	0.286
Lowest Social Class	0.306	0.461	0.291	0.443	0.088	0.284	0.088	0.284
Highest Social Class	0.115	0.318	0.107	0.310	0.146	0.353	0.153	0.360
Parent High Educ.	0.164	0.370	0.147	0.354	0.195	0.396	0.220	0.414
Parent Low Educ.	0.211	0.408	0.329	0.470	0.153	0.360	0.086	0.281
Intelligence Test					0.018	0.832	0.062	0.848
Early Tracking T1/T2					0.379	0.485	0.351	0.477
Early Tracking T2/T3					0.136	0.343	0.220	0.414
Higher Education					0.295	0.456	0.382	0.486
University					0.104	0.305	0.134	0.340
Dropout					0.126	0.332	0.129	0.335
Employment					0.870	0.336	0.847	0.360
Welfare					0.090	0.287	0.068	0.252

Notes: Table shows means and standard deviations of all relevant variables. Earnings are per month, pre-tax and in euro's. Reported earnings are corrected for part-time factor, topcoded at 10,000 euro's, and earnings below 1000 euro's are excluded. The variables Track 1a to Track 3 report the fraction of students that go to that track. Track 1 (*vmbo*) is the sum of Track 1a (*lbo*) and Track 1b (*mavo*). Years of schooling is coded to the highest degree obtained. Parent High Education is a dummy variable that takes value 1 when the highest education level of the parents is in one of the top two categories (six categories are defined). Parent Low Education takes value 1 when the highest education level of the parents is in one of the lowest two categories. Social class is based on the occupational status of the parents. 'Big 4 cities' indicates whether the child lived in one of the four big cities in the Netherlands. Higher education consists of higher professional education (*hbo*) and university education. Employment measures whether the individual earned at least a minimum wage in 2007. Welfare measures whether the individual ever received welfare benefits between 1999 and 2007.

Table 5.2: Thresholds along test score distribution

	1977 [0-70]	1982 [0-60]	1989 [0-60]	1993 [0-60]
Track 1a - Track 1b	31	27	26	30
Track 1 - Track 2	50	40	44	42
Track 2 - Track 3	57	48	49	48

Notes: Table reports the unofficial thresholds between tracks in terms of test score that need to be obtained to be eligible for a certain track. The values give the highest score that still makes one eligible for the lower of the two tracks. This means that, for example, a score of 58 is needed to be eligible for track T3 in 1977 and a score of 51 for track T2. These implied thresholds are based on the strength of the first stage using different thresholds, effectively measuring where the discontinuity in treatment probability is strongest.

Table 5.3: Strength of instrumentation: F-values

	1977	1982	1989	1993
Track 1a - Track 1b	744.93	404.54	159.27	4.91
Track 1 - Track 2	555.62	1098.53	1930.12	1132.55
Track 2 - Track 3	8.62	318.93	205.82	37.69

Notes: Table reports the strength of using eligibility of treatment as an instrument for treatment, represented by the corresponding F-score. Eligibility is determined by the student's score on the Entrance Test. F-scores are estimated using those bandwidths suggested by the cross-validation procedure, which are also the ones employed in the actual estimation.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Table 5.4: OLS estimates of track assignment on earnings and years of schooling

	1977	Track 1 1982	1989	1977	Track 3 1982	1989
Years of Schooling	-1.40*** (0.054)	-1.84*** (0.082)	-2.35*** (0.073)	1.60** (0.065)	2.04*** (0.100)	2.15*** (0.078)
Earnings 2007	-0.153*** (0.012)	-0.166*** (0.011)	-0.128*** (0.0081)	0.182*** (0.014)	0.168*** (0.014)	0.111*** (0.0091)
Earnings 2006	-0.158*** (0.012)	-0.169*** (0.010)	0.118*** (0.0079)	0.182*** (0.013)	0.148*** (0.014)	0.095*** (0.0090)
Earnings 2005	-0.161*** (0.012)	-0.161*** (0.010)	-0.095*** (0.0074)	0.190*** (0.013)	0.153*** (0.013)	0.081*** (0.0085)
Earnings 2004	-0.153*** (0.011)	-0.147*** (0.0092)	-0.084*** (0.0069)	0.188*** (0.013)	0.143*** (0.012)	0.072*** (0.0082)
Earnings 2003	-0.147*** (0.011)	-0.141*** (0.0086)	-0.062*** (0.0069)	0.176*** (0.012)	0.131*** (0.012)	0.053*** (0.0081)
Earnings 2002	-0.141*** (0.011)	-0.128*** (0.0084)	-0.061*** (0.0068)	0.164*** (0.012)	0.123*** (0.011)	0.048*** (0.0085)
Earnings 2001	-0.133*** (0.010)	-0.112*** (0.0081)	-0.035*** (0.0072)	0.156*** (0.012)	0.113*** (0.011)	0.028*** (0.0097)
Mean Earnings	-0.105*** (0.0092)	-0.143*** (0.0084)	-0.079*** (0.0063)	0.179*** (0.010)	0.136*** (0.012)	0.069*** (0.0074)

Notes: Table shows the effect of T1 and T3 treatment on educational attainment and the logarithm of earnings in a certain year, measured by using OLS (applying a Sharp RD design). The applied bandwidths and control functions are the ones suggested by the cross-validation procedure (and the same as in the IV model).

Table 5.5: IV estimates of Track 1 assignment on earnings and years of schooling

	1977 [10-61]	1982 [14-55]	1989 [32-60]
Years of Schooling	-0.901*** (0.216)	-0.861** (0.359)	-1.33*** (0.249)
Earnings 2007	-0.090** (0.038)	-0.109** (0.046)	-0.122*** (0.021)
Earnings 2006	-0.089** (0.039)	-0.089** (0.045)	-0.112*** (0.040)
Earnings 2005	-0.086** (0.038)	-0.091** (0.042)	-0.057 (0.038)
Earnings 2004	-0.081** (0.037)	-0.058 (0.040)	-0.080** (0.036)
Earnings 2003	-0.079** (0.036)	-0.056 (0.038)	-0.064* (0.037)
Earnings 2002	-0.103*** (0.035)	-0.061* (0.037)	-0.049 (0.036)
Earnings 2001	-0.110*** (0.034)	-0.057 (0.036)	-0.060* (0.036)
Mean Earnings	-0.105* (0.034)	-0.096*** (0.036)	-0.066** (0.031)
n	14468	10179	8364

Notes: Table shows the effect of T1 treatment on educational attainment and the logarithm of earnings in a certain year, measured by using IV (applying an Fuzzy RD design). We use treatment eligibility based on the Entrance Test score as an instrument and allow for separate control functions before and after the threshold. The applied bandwidths (which are listed at the top of the columns) and the specifications for the control function are the ones suggested by the cross-validation procedure. Estimates for cohort 1993 are not given, because the first stage is too weak to estimate earnings effects. The effect on years of schooling in 1993 (not shown in table) is equal to -1.76 (significant at 1% level). The bandwidths refer to estimation of earnings effects. Different bandwidths are used to measure effects on educational attainment.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Table 5.6: IV estimates of Track 3 assignment on earnings and years of schooling

	1977 [12-59]	1982 [14-55]	1989 [20-60]	1993 [28-58]
Years of Schooling	1.46*** (0.142)	1.46*** (0.197)	1.88*** (0.175)	1.87*** (0.161)
Earnings 2007	0.137*** (0.027)	0.099*** (0.030)	0.082*** (0.019)	0.045* (0.022)
Earnings 2006	0.138*** (0.028)	0.094*** (0.030)	0.056*** (0.019)	0.036 (0.024)
Earnings 2005	0.131*** (0.027)	0.090*** (0.028)	0.062*** (0.018)	-0.031 (0.026)
Earnings 2004	0.124*** (0.026)	0.112*** (0.026)	0.059*** (0.017)	-0.024 (0.028)
Earnings 2003	0.115*** (0.026)	0.105*** (0.025)	0.019 (0.017)	-0.0051 (0.032)
Earnings 2002	0.127*** (0.025)	0.070*** (0.024)	0.039** (0.018)	-0.026 (0.033)
Earnings 2001	0.106*** (0.025)	0.090*** (0.024)	0.018 (0.019)	0.049 (0.039)
Mean Earnings	0.128*** (0.024)	0.090*** (0.024)	0.044*** (0.015)	0.025 (0.020)
n	16795	10366	12344	10853

Notes: Table shows the effect of T3 treatment on educational attainment and the logarithm of earnings in a certain year, measured by using IV (applying an Fuzzy RD design). For an explanation on the instrumentation and estimation model, see table 5.5. The bandwidths refer to estimation of earnings effects. Different bandwidths are used to measure effects on educational attainment.

Table 5.7: IV estimates of Track 1a assignment on earnings

	1982 [10-48]	1989 [13-45]	1993 [18-31]
Earnings 2007	-0.105** (0.048)	-0.0098 (0.048)	0.090 (0.085)
Earnings 2006	-0.138*** (0.046)	-0.067 (0.047)	-0.030 (0.074)
Earnings 2005	-0.076* (0.043)	-0.060 (0.042)	0.017 (0.082)
Earnings 2004	-0.079* (0.041)	-0.035 (0.041)	0.064 (0.080)
Earnings 2003	-0.070* (0.039)	0.017 (0.041)	0.037 (0.084)
Earnings 2002	-0.047 (0.038)	-0.036 (0.041)	0.012 (0.088)
Earnings 2001	-0.039 (0.037)	-0.053 (0.041)	0.026 (0.117)
Mean Earnings	-0.078** (0.038)	-0.035 (0.035)	0.019 (0.064)
n	10139	10806	4410

Notes: Table shows the effect of T1a treatment on the logarithm of earnings in a certain year, measured by using IV (applying an Fuzzy RD design). For an explanation on the instrumentation and estimation model, see table 5.5. There are no results for the 1977 and 1993 cohorts, or for years of schooling, since the first stage is too weak.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Table 5.8: Sensitivity of results to bandwidth

Track 1										
	[15-49]	[15-60]	[5-60]	[25-49]	[30-49]	[35-49]	[35-45]	TH-2	TH+2	
Earnings 2007	-0.090** (0.038)	-0.081** (0.040)	-0.117*** (0.040)	-0.104** (0.042)	-0.171*** (0.061)	-0.637** (0.278)	-0.388*** (0.117)	-0.0045 (0.048)	-0.0018 (0.045)	
Earnings 2006	-0.089** (0.039)	-0.083*** (0.041)	-0.105*** (0.041)	-0.111*** (0.043)	-0.177*** (0.063)	-0.525* (0.279)	-0.295** (0.118)	-0.023 (0.047)	-0.022 (0.045)	
Earnings 2005	-0.086** (0.038)	-0.079** (0.040)	-0.109*** (0.040)	-0.101** (0.042)	-0.149** (0.061)	-0.419 (0.271)	-0.222* (0.114)	-0.038 (0.046)	-0.037 (0.043)	
	[22-48]	[22-60]	[5-60]	[25-48]	[30-48]	[35-48]	[30-45]	TH-2	TH+2	
Years of Schooling	-0.861** (0.359)	-0.081 (0.527)	-0.428 (0.315)	-0.548 (0.465)	1.50 (1.28)	15.49 (12.02)	0.015 (0.705)	-1.22*** (0.0437)	-2.03*** (0.342)	
Track 3										
	[12-68]	[12-70]	[12-65]	[12-60]	[20-60]	[30-60]	[40-60]	TH-2	TH+2	
Earnings 2007	0.137*** (0.027)	0.134*** (0.028)	0.136*** (0.026)	0.171*** (0.029)	0.175*** (0.030)	0.189*** (0.033)	0.212*** (0.041)	0.084*** (0.067)	0.098*** (0.031)	
Earnings 2006	0.138*** (0.028)	0.138*** (0.028)	0.139*** (0.026)	0.166*** (0.030)	0.171*** (0.031)	0.182*** (0.034)	0.186* (0.042)	0.121*** (0.066)	0.106*** (0.031)	
Earnings 2005	0.131*** (0.027)	0.130*** (0.028)	0.139*** (0.026)	0.169*** (0.029)	0.171** (0.030)	0.177*** (0.033)	0.190 (0.041)	0.156*** (0.064)	0.102*** (0.030)	
	[10-58]	[5-60]	[20-58]	[25-58]	[35-55]	TH-2	TH+2			
Years of Schooling	1.48*** (0.197)	1.50*** (0.196)	1.60*** (0.220)	1.58*** (0.253)	1.40*** (0.458)	-0.211 (0.954)	-3.47 (3.29)	1.65*** (0.202)	1.81*** (0.212)	

Notes: Table shows the sensitivity of the results from tables 5.5 and 5.6 to changes in the bandwidth and, in the last two columns, to the chosen threshold (TH) for treatment eligibility. The first entry shows the result for the official bandwidth (the one suggested by the cross-validation procedure). Other entries show the effect of a large range of arbitrary changes in that bandwidth, or in the threshold. For an explanation on the instrumentation and estimation model, see table 5.5.

Table 5.9: IV estimates of track assignment when dropping tracks

	Track 3 '77			Track 1 '82	
	Original	Drop T1a	Drop T1	Original	Drop T3
Earnings 2007	0.137*** (0.027)	0.190*** (0.033)	0.144 (0.116)	-0.109** (0.046)	-0.085* (0.049)
Earnings 2006	0.138*** (0.028)	0.182*** (0.034)	0.130 (0.121)	-0.089** (0.045)	-0.051 (0.048)
Earnings 2005	0.131*** (0.027)	0.165*** (0.033)	0.149 (0.120)	-0.091** (0.042)	-0.074* (0.044)
Earnings 2004	0.124*** (0.026)	0.157*** (0.031)	0.107 (0.109)	-0.058 (0.040)	-0.033 (0.042)
Earnings 2003	0.115*** (0.026)	0.148*** (0.031)	0.162 (0.112)	-0.056 (0.038)	-0.033 (0.039)
Earnings 2002	0.127*** (0.025)	0.154*** (0.031)	0.099 (0.118)	-0.061* (0.037)	-0.036 (0.038)
Earnings 2001	0.106*** (0.025)	0.128*** (0.030)	0.089 (0.114)	-0.057 (0.036)	-0.044 (0.037)
Mean Earnings	0.128*** (0.024)	0.160*** (0.029)	0.109 (0.109)	-0.096*** (0.036)	-0.057 (0.044)
Years of Schooling	1.46*** (0.142)	1.75*** (0.160)	1.59*** (0.451)	-0.861** (0.359)	0.052 (0.479)

	Track 3 '82			Track 3 '89		
	Original	Drop T1a	Drop T1	Original	Drop T1a	Drop T1
Earnings 2007	0.106*** (0.030)	0.103*** (0.035)	0.110 (0.071)	0.084*** (0.020)	0.095*** (0.022)	0.041 (0.052)
Earnings 2006	0.102*** (0.030)	0.105*** (0.035)	0.105 (0.072)	0.057*** (0.020)	0.059*** (0.022)	-0.017 (0.052)
Earnings 2005	0.096*** (0.028)	0.096*** (0.033)	0.058 (0.070)	0.064*** (0.018)	0.066*** (0.021)	0.062 (0.046)
Earnings 2004	0.116*** (0.026)	0.120*** (0.031)	0.129** (0.063)	0.060*** (0.017)	0.064*** (0.020)	0.048 (0.045)
Earnings 2003	0.110*** (0.025)	0.112*** (0.029)	0.073 (0.059)	0.020 (0.017)	0.020 (0.020)	-0.0013 (0.047)
Earnings 2002	0.075*** (0.024)	0.073*** (0.028)	0.023 (0.058)	0.040** (0.018)	0.033 (0.021)	0.028 (0.048)
Earnings 2001	0.093*** (0.023)	0.094*** (0.028)	0.092 (0.055)	0.018 (0.020)	0.0072 (0.022)	-0.016 (0.050)
Mean Earnings	0.090*** (0.024)	0.086*** (0.029)	0.084 (0.060)	0.044*** (0.015)	0.045*** (0.017)	0.033 (0.040)
Years of Schooling	1.46*** (0.197)	1.70*** (0.241)	1.37** (0.546)	1.88*** (0.175)	2.08*** (0.187)	1.21*** (0.466)

Notes: Table shows treatment effects when we drop one or two tracks at the other extreme of the distribution. For Track 1 treatment we can only do this for Track 3; for Track 3 treatment we can do this for both Track 1a and Track 1. The table lists the original results from tables 5.5 and 5.6 and the results when tracks are dropped. Dropping tracks for Track 1 in 1977 is not possible since we already excluded Track 3 in the initial analysis there. No results are shown for Track 1 in '89 because instrumentation fails when Track 3 is dropped. For an explanation on the instrumentation and estimation model, see table 5.5.

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Table 5.10: IV estimates of track assignment when using track-specific trends

	1977		1982	1989	
	Track 1	Track 3	Track 3	Track 1	Track 3
Earnings 2007	-0.169*** (0.031)	0.204*** (0.029)	0.160*** (0.031)	-0.122*** (0.021)	0.127*** (0.020)
Earnings 2006	-0.171*** (0.031)	0.206*** (0.029)	0.157*** (0.031)	-0.118*** (0.037)	0.095*** (0.020)
Earnings 2005	-0.179*** (0.031)	0.199*** (0.029)	0.149*** (0.029)	-0.067* (0.034)	0.092*** (0.018)
Earnings 2004	-0.184*** (0.030)	0.186*** (0.027)	0.167*** (0.027)	-0.084** (0.033)	0.084*** (0.017)
Earnings 2003	-0.181*** (0.029)	0.180*** (0.027)	0.152*** (0.026)	-0.068** (0.034)	0.037** (0.017)
Earnings 2002	-0.182*** (0.028)	0.187*** (0.027)	0.114*** (0.025)	-0.049 (0.033)	0.060*** (0.018)
Earnings 2001	-0.194*** (0.027)	0.167*** (0.026)	0.129*** (0.025)	-0.057* (0.034)	0.029 (0.020)
Earnings 2001	-0.195*** (0.027)	0.195*** (0.025)	0.145*** (0.025)	0.071*** (0.028)	-0.071** (0.015)
Years of Schooling	-0.918*** (0.213)	1.84*** (0.149)	1.98*** (0.205)	-1.27*** (0.254)	2.54*** (0.159)

Notes: Table shows treatment effects when we use separate control functions for every track. Results for Track 1 in 1982 are not shown, since the original specification was already conducted in this way. For an explanation on the instrumentation and estimation model, see table 5.5.

Table 5.11: IV estimates of track assignment when also instrumenting control function

	1977		1982		1989
	Track 1	Track 3	Track 1	Track 3	Track 3
Earnings 2007	0.074 (0.087)	0.083 (0.057)	-0.196 (0.149)	0.153*** (0.058)	0.091** (0.038)
Earnings 2006	0.049 (0.087)	0.097* (0.056)	-0.265* (0.142)	0.124** (0.058)	0.056 (0.038)
Earnings 2005	-0.089 (0.080)	0.185*** (0.054)	-0.0022 (0.145)	0.096* (0.055)	0.029 (0.035)
Earnings 2004	-0.025 (0.079)	0.136** (0.054)	-0.221 (0.143)	0.096* (0.052)	0.059* (0.035)
Earnings 2003	-0.024 (0.077)	0.121** (0.051)	-0.319*** (0.122)	0.121** (0.049)	0.058* (0.033)
Earnings 2002	-0.040 (0.075)	0.154*** (0.048)	-0.195 (0.120)	0.120** (0.049)	0.036 (0.036)
Earnings 2001	-0.026 (0.072)	0.153*** (0.048)	-0.238* (0.125)	0.134*** (0.046)	0.033 (0.038)
Mean Earnings	-0.021 (0.073)	0.146*** (0.049)	-0.165 (0.118)	0.136*** (0.046)	0.034 (0.029)
Years of Schooling	-0.295 (0.301)	1.18*** (0.312)	-0.327 (0.902)	0.485 (0.363)	1.48*** (0.274)

Notes: Table shows treatment effects when we also instrument the control function, using the product of the forcing variable and our main instrument. Results for Track 1 in 1989 are not shown, since the first stage is too weak. We use the propensity score given the score on the Entrance Test as our main instrument, as proposed by Van der Klaauw (2002). The first stage is too weak to employ eligibility as an instrument (the model does work there for years of schooling, which has a larger bandwidth. This effect is -1.17 and significant at the 1% level).

5. THE LABOR MARKET EFFECTS OF TRACKING: A REGRESSION DISCONTINUITY APPROACH

Table 5.12: IV estimates of track 1 assignment when adding controls

	1977		1982		1989	
Track 1						
Earnings 2007	-0.090*** (0.038)	-0.059* (0.035)	-0.109** (0.046)	-0.084* (0.045)	-0.122*** (0.021)	0.106*** (0.021)
Earnings 2006	-0.089*** (0.039)	-0.071** (0.035)	-0.089* (0.045)	-0.064 (0.044)	-0.112*** (0.040)	-0.109** (0.041)
Earnings 2005	-0.086*** (0.038)	-0.065* (0.035)	-0.091** (0.042)	-0.074* (0.040)	-0.057 (0.038)	-0.053 (0.038)
Earnings 2004	-0.081** (0.037)	-0.063* (0.034)	-0.058 (0.040)	-0.045 (0.040)	-0.080** (0.036)	-0.078** (0.037)
Earnings 2003	-0.079** (0.036)	-0.058* (0.033)	-0.056 (0.038)	-0.041 (0.038)	-0.064* (0.037)	-0.065* (0.038)
Earnings 2002	-0.103*** (0.035)	-0.080** (0.032)	-0.061* (0.037)	-0.044 (0.036)	-0.049 (0.036)	0.048 (0.037)
Earnings 2001	-0.110*** (0.034)	-0.091*** (0.032)	-0.057 (0.036)	-0.042 (0.036)	-0.060* (0.036)	-0.059 (0.037)
Mean Earnings	-0.105*** (0.034)	-0.078*** (0.030)	-0.096*** (0.036)	-0.076*** (0.035)	-0.066** (0.031)	-0.067** (0.036)
Years of Schooling	-0.901*** (0.216)	-0.881*** (0.218)	-0.861** (0.359)	-0.747** (0.370)	-1.33*** (0.249)	-1.26*** (0.252)
Controls	Yes		Yes		Yes	
Track 3						
Earnings 2007	0.137*** (0.027)	0.112*** (0.026)	0.099*** (0.030)	0.086*** (0.029)	0.082*** (0.019)	0.069*** (0.019)
Earnings 2006	0.138*** (0.028)	0.115*** (0.026)	0.094*** (0.030)	0.084*** (0.029)	0.056*** (0.019)	0.046** (0.019)
Earnings 2005	0.131*** (0.027)	0.102*** (0.026)	0.090*** (0.028)	0.075*** (0.028)	0.062*** (0.018)	0.053*** (0.018)
Earnings 2004	0.124*** (0.026)	0.095*** (0.024)	0.112*** (0.026)	0.100*** (0.026)	0.059*** (0.017)	0.055*** (0.017)
Earnings 2003	0.115*** (0.026)	0.086*** (0.024)	0.105*** (0.025)	0.093*** (0.025)	0.019 (0.017)	0.014 (0.017)
Earnings 2002	0.127*** (0.025)	0.107*** (0.024)	0.070*** (0.024)	0.056** (0.024)	0.039** (0.018)	0.037** (0.018)
Earnings 2001	0.106*** (0.025)	0.089*** (0.024)	0.090*** (0.024)	0.080*** (0.023)	0.018 (0.019)	0.020 (0.020)
Mean Earnings	0.128*** (0.024)	0.107*** (0.022)	0.090*** (0.024)	0.076*** (0.024)	0.044*** (0.015)	0.040*** (0.015)
Years of Schooling	1.46*** (0.142)	1.41*** (0.147)	1.46*** (0.197)	1.45*** (0.199)	1.88*** (0.175)	1.79*** (0.155)
Controls	Yes		Yes		Yes	

Notes: Table shows treatment effects of assignment to Track 1 and Track 3, with and without controls. The added controls are gender, month of birth, whether the student lives in one of the four big cities in the Netherlands, ethnicity, social status and parental education. For an explanation on the instrumentation and estimation model, see table 5.5.

Table 5.13: Cito score versus Entrance Test score as forcing variable

	Track 1		Track 3	
	Language	Math	Language	Math
Entrance Test	-0.098*** (0.024)	-0.123*** (0.027)	0.117*** (0.023)	0.171*** (0.019)
Cito Test	-0.106*** (0.016)	-0.123*** (0.027)	0.124*** (0.019)	0.130*** (0.014)

Notes: Table shows a comparison of treatment effects when using the unofficial Entrance Test score as a forcing variable (as in all other tables) versus the formal Cito score, for language and math scores in the 3rd year of secondary education. For an explanation on the instrumentation and estimation model, see table 5.5.

6

The Impact of Early Tracking on Later-life Outcomes: An Instrumental Variable Approach¹

6.1 Introduction

The previous chapter has looked at whether the thresholds in a country's tracking system maximize the outcomes of each individual student. The exact age at which students are selected into tracks constitutes another important source of variation among tracking systems. The choice of early versus late tracking is a choice between a more directed school curriculum and a more homogeneous peer group versus the possible positive interactions from having a peer group of heterogeneous ability. This chapter estimates the effect of early tracking on several measures of educational attainment and labor market outcomes. We use an instrumental variable model where we employ the relative supply of early tracking schools in the municipality as an instrument for early tracking. We use Dutch data on secondary and tertiary education and the labor market, which is the same as in the previous chapter. We find that early tracking has a strong negative impact on educational attainment, and this is translated into lower earnings for some subgroups as well. The negative effect of tracking on these outcomes reduces when we move upward in the distribution. However, it only becomes positive for the very highest ability level. The effect of early tracking on earnings can be attributed to a decrease in employment and hours

¹This chapter is based on Borghans, Diris, Smits, and De Vries (2012).

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

worked, rather than to a decrease in the hourly wage. We employ several robustness checks that provide evidence in favor of the validity of our instrument.

Developed nations track students as early as age 10, while some have a comprehensive system throughout secondary school. The Dutch educational system is in between these extremes, although it is generally seen as an early tracking system. Children in the Netherlands enter secondary school at age 12, but actual track assignment can still be postponed for one or two years. This induces variation in the tracking age within the Netherlands that can be exploited. Where other studies rely on between-country differences¹ or changes in a country's tracking system over time,² we look at differences between students in the same country at the same point in time.

This chapter employs an Instrumental Variable approach to estimate the effect of tracking age, using regional variation in the supply of schools that track early. As such, it has similarities with the study of Van Elk et al. (2011). They use the relative supply of early tracking schools by degree of urbanization, which is divided into nine categories, as an instrument for early tracking. They estimate the effect of early tracking on completion of higher education, for the subgroup with a secondary school advice for Track 1b. We use the relative supply of early tracking on the level of the municipality as an instrument. We estimate effects for several measures of educational attainment and labor market outcomes, for the full distribution of students. An advantage of this study over other papers that examine the effect of the tracking age, which we share with Van Elk et al. (2011), is that we are looking at variation in the tracking age *within* a country, at the same point in time. Thereby, we avoid relying on differences in tracking across countries, which has a large potential for confounding factors. Moreover, we examine a long-established tracking system, in contrast to studies that look at a recent reduction or expansion in the number of tracks.

This chapter is organized as follows. Section 6.2 discusses the relevant characteristics of our data. The methodological approach of our analysis is explained in section 6.3. Section 6.4 presents and elaborates on the main results, after which robustness issues are discussed in section 6.5. Section 6.6 concludes. Previous literature in the field of educational tracking and additional explanation of the data are located in the previous chapter.

¹See e.g. Hanushek and Woessmann (2006); Jakubowski (2010).

²See e.g. Galindo-Rueda and Vignoles (2005); Pekkarinen et al. (2009)

6.2 Data

We employ the same data in this chapter as in the previous one. We only use the 1989 and 1993 cohorts, since the instrument is only available for those waves.¹ Additionally, we focus more on educational attainment as an outcome than in Chapter 5. Our main indicator for educational attainment is completion of higher education. As explained in Chapter 5, this constitutes both completion of higher professional education (*hbo*) or university education (*wo*). We also employ completion of university education as a separate outcome variable, as well as dropout, which refers to only completing primary education. We only use the 1989 cohort for our estimation of labor market outcomes. Individuals in the 1993 cohort are aged 26 and younger. Some observations are still in education, or at the very start of working life. Including them will undervalue the earnings potential of the most well-educated individuals and thereby not give a proper picture of the true impacts of early tracking. The 1989 cohort is aged 27 to 30 during the years we include in our earnings estimation.² We use the raw reported monthly earnings as an outcome, as well as corrected earnings. The latter correct the monthly wage for the part-time factor, exclude wages below the minimum wage and topcode wages at 10,000 euro's per month. Other labor market outcomes we use are: a dummy 'employment' variable which takes the value of 1 if the individual earns at least a minimum (monthly) wage, and a dummy that indicates whether the individual was ever on welfare from 1999 to 2007.³ We use both cohorts when we estimate impact on welfare benefits, because the difference in the gradient between those who attain higher levels of education should not be an issue here.

We have information on the exact school type each individual attends for all years of secondary education and thereafter. We can identify when students enter a non-comprehensive track for the first time. This constitutes their tracking age. A student is in a non-comprehensive track if he or she is separated from the other track under examination. When we are looking at tracking between the low (T1) and medium (T2) track, this implies that a T1 student is separated from T2 students and vice versa. The T2 student can still be in a comprehensive class together with T3 students. The same reasoning applies for tracking between medium and high track. We also observe the type of education students are in after they finish secondary education. Thereby, we

¹Our instrument is based on national data on the relative supply of early tracking schools from 1997. Going back as far as 1982 would lead to a too inaccurate instrument.

²We use the earnings data from 2004 to 2007, since the vast majority of individuals in the 1989 cohort have finished education by then.

³This includes unemployment benefits and disability benefits.

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

can follow the educational levels they complete until the year 2007, when the 1989 cohort is 30 years old and the 1993 cohort is 26 years old.

Summary statistics can be seen in table 5.1.¹ Government reports show that the share of students aged 25-34 that have a higher educational degree is 34 percent in 2007 (Ministerie van Onderwijs Cultuur en Wetenschap, 2011). Hence, the overall sample is very representative of the total Dutch population in this age group, although 1989 is slightly less and 1993 slightly more affluent. Figure 6.2 shows where each advice group ends up in terms of educational attainment. It shows that higher education (the sum of the two right most bars) is completed by students of all subgroups, but that this strongly increases by advice group. The prevalence of higher education completion is more than twice as high in the subgroup with T2 advice than for those with T1b advice. This reflects, besides the higher level of T2 students, also the easier access to higher education for T2 students.

6.3 Methodology

6.3.1 Defining early tracking

We estimate the effect of being tracked early or late on different levels of the Dutch educational system. Since the Dutch system has four school tracks, there are three levels at which students can be tracked or held together: between track 1a and 1b; track 1b and track 2; and track 2 and 3. For ease of notation we refer to tracking between tracks 1b and 2 as Track 1/Track 2 tracking, or T1/T2 tracking.² We are not able to estimate the effects of T1a/T1b tracking since the variation in tracking age is too low at that level to ensure a powerful first stage. The vast majority of students in these two tracks are already separated from each other at age 12, and only few thereafter. Tracking at age 12 is very rare in the case of T2/T3 tracking. Only 9% of students who end up in T2 or T3 are already in a specific track by age 12. Hence, early tracking is defined at this level as at age 13 or earlier. The definition of early tracking is less straightforward for T1/T2 tracking. There is a fairly even division of tracking at ages 12, 13 and 14. Choosing a division at age 12 or thereafter provides a stronger first stage for low-ability subgroups, but choosing age 13 or thereafter produces strong F-values for a larger part of the distribution.³

¹Since the data are the same as in the previous chapter, table 5.1, located at the end of the previous chapter, lists all relevant variables for both chapters.

²Correspondingly, the other levels of tracking are labeled as Track 1a/Track 1b tracking and Track 2/Track 3 tracking, or T1a/T1b and T2/T3 tracking.

³This is mainly because the T1/T2 students of relatively higher ability are often tracked in T1/T2/T3 comprehensive classes, which generally do not track at age 12.

Since we are looking at effects for all subgroups, we employ the latter. Hence, we define early tracking at the T1/T2 level as before age 14, or the third year of secondary education.¹

6.3.2 Endogeneity

Whether you attend a school that tracks early is, to a large extent, a free choice of the student and his parents. Hence, it is an endogenous variable which could be strongly correlated with several other characteristics that affect outcome variables, such as educational attainment and earnings.

Table 6.1 shows the means of the groups that are tracked early and those who are not, for tracking on both the T1/T2 and T2/T3 level. The fourth and eighth column of table 6.1 give the p-value of a two-sided Wald test on whether these means are equal. The table clearly shows that those who are tracked early have less favorable observables in case of T1/T2 tracking, and more favorable in the case of T2/T3 tracking. This is reflected by significant differences in Entrance Test scores, parental education, social class and ethnicity. The difference are more pronounced for T1/T2 tracking. Table 6.1 shows that high-ability students and their parents have a preference for being in a separate T3 track, while the relatively highly-able students at the T1/T2 level prefer to be in a mixed class to be able to make it to the T2 track. Since the T1 track is much larger, the majority of early tracked students in the T1/T2 sample is tracked to a T1 track and hence of lower ability.² The table also lists the outcome variables we employ. They show that the early tracked students are generally worse off for both levels of tracking. For T2/T3 tracking, these students are also of higher ability, suggesting that the actual effect is even stronger. The differences for T1/T2 tracking are often severe, but given that non-tracked students are of better ability, the true effects can be substantially lower.

6.3.3 The IV model

Hence, one needs a source of exogenous variation to identify the effect of early tracking. We employ Instrumental Variable estimation to correct for endogeneity. We construct separate instruments for T1/T2 tracking and T2/T3 tracking. When there is at least one student in the school that is tracked separate from the other track under examination, this school is classified as a tracking school at that specific grade level. Our instrument measures the fraction of *students*

¹Van Elk et al. (2011) choose age 12 or thereafter as a division point. This produces stronger instrumentation for the T1b subgroup, which is the subpopulation that comprises their sample.

²Lower ability refers to less favorable observables, as shown in table 6.1, which suggests relevant unobservables are less favorable as well.

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

in a municipality that are in a school where that specific track is offered separately before age 14. This is the municipality where the school resides rather than the student. These values are not based on the sample at hand, but determined on a national level. Graphs 6.3 and 6.4 show that there are multiple municipalities where the instrument takes either value 0 (no tracking anywhere) or 1 (all schools track), but there are many values in between that range as well. Our explicit assumption is that the relative supply of early tracking schools on the municipal level is not influenced by demand side variables. Figure 6.5 shows the relation between mean Entrance Test scores on the municipal level and our instrument, for tracking on both levels. The figure shows that there is no relation between a municipality's average student level and the degree of early tracking in that municipality.¹

We employ a long list of control variables. We compare results with and without controls to get an indication of the validity of our instrument. Our control variables are: month of birth, gender, ethnic origin,² social class, highest education level of the parents,³ whether the child lived in one of the big 4 cities in the Netherlands,⁴ the score on an intelligence test taken in year 1 of secondary school and the scores on all three test domains of the Entrance Test, also taken at the beginning of secondary education.⁵ Our model then becomes:

$$D_i = \beta_0 + \beta_1 w + \beta_2 X' + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X' + \epsilon_i \quad (6.1)$$

D indicates whether the child is tracked early or not. The variable w is our instrument, indicating the relative supply of early tracking schools at the municipal level. X' is a vector of controls (described above) and ϵ is our error term. We employ robust standard errors, which are clustered on a school level. Y is our outcome variable. These outcomes are measures of educational attainment or labor market outcomes.

In addition to overall effects, we estimate effects by subgroup. We categorize these subgroups by the track advice they receive at the end of primary school. This advice is strongly

¹This is confirmed by formal estimation of this correlation, which proves to be insignificant.

²Defined as a dummy variable that takes the value of 1 if the child or his parents were not born in the Netherlands.

³We employ two dummy variables. One takes the value 1 if the highest educational level of the parents is in the top two categories, the other when it is in the lowest two categories (six categories are defined).

⁴Amsterdam, Rotterdam, The Hague and Utrecht

⁵These domains are language, mathematics and information processing. The scores are entered separately for each domain.

based on the Cito score, but can deviate from that in some cases. This occurs when the teacher believes this score does not reflect the students true ability. Using these groups allows us to see the impacts of early tracking on different types of students, which are essentially ranked according to ability. Schools and teachers can give specific track advices or mixed advices. Figure 6.1 shows where the students of these different advice groups end up at their point of actual track selection. Adherence to the T1 and T3 advices is very strong. This is lower for the smaller T2 track, where there is mainly upstream towards T3. This is also the case for the mixed T2/T3 advice, which most often leads to assignment towards T3.

Table 6.2 shows the F-values of a Wald test on the instrument, measuring the strength of the instrument. We report this separately for educational attainment and labor market outcomes. The rows list the values for the overall samples and the subgroups according to secondary school advice. We only use the 1989 cohort in our estimation of labor market effects, which naturally makes these F-values smaller. Our instrumentation is strongest for T2/T3 tracking, especially for educational attainment. The informal critical value¹ of 10 is not reached in T1/T2 tracking for some subgroups with respect to educational attainment. The overall sample does not pass the critical value for labor market outcomes and only one subgroup does. We report the effect for these estimations nonetheless, but they are very imprecise and carry less weight.

6.4 Results

We present results of the effect of early tracking on the level of T1/T2 and on the level of T2/T3. We estimate the effect of being in a separate T1 or T2 track versus being grouped together with student of both tracks, and of being in a separate T2 or T3 track versus a comprehensive T2/T3 track. We are also interested in how effects of early tracking differ for students of different ability. The general belief is that tracking hurts those at the bottom of the ability distribution, but could be beneficial for the highest-able students.

We estimate the effect of early tracking on several outcome variables, which we organize in two categories: educational attainment and labor market effects. In the first category, we look at completion of higher education, completion of university and school dropout (not finishing secondary education). Labor market outcomes consist of two different measures of earnings, employment and use of welfare. One measure of earnings consists of those earnings reported to

¹First proposed by Staiger and Stock (1997).

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

the tax authorities. In a second measure, we exclude earnings below the minimum wage, correct for part-time factor and topcode at 10,000 euro's per month.

Results using OLS are portrayed in table 6.3, for comparative purposes. We expect them to be positively biased in the case of T2/T3 tracking and negatively biased for T1/T2 tracking, based on the comparison of observables in table 6.1. The OLS estimates for T2/T3 tracking are largely insignificant. We can observe a clear pattern of negative effects for the low ability groups and positive effects for the higher ability groups. The effects for T1/T2 tracking are significant and negative, especially for the low advice groups. The effect of adding controls is in some cases substantial, mainly for the higher advice groups. This confirms our belief that early tracking is endogenous,¹ which is partly picked by our control variables.

6.4.1 Educational attainment

Table 6.4 shows the effects of early tracking on multiple measures of educational attainment, both for the sample as a whole as for subgroups according to ability. Early tracking reduces the chance of completing a higher education degree, both on the T1/T2 and on the T2/T3 level. The reduction is approximately 16 percentage points for T1/T2 tracking and 8 percentage points for the T2/T3 group. This negative effect originates primarily from students who got a secondary school advice for the T1 track or a mixed advice for T1/T2 for the former. For T2/T3 tracking, the effects are most prominent for the T1 (which is relatively small), T1/T2 and T2 advice groups. The fact that the impact is larger for the T1/T2 sample can be driven by the fact that early tracking closes off some paths towards higher education for those tracked to a T1 track early. The vast majority of early tracked students in the T1/T2 sample end up in a T1 track. For the subgroup with a T2/T3 advice, the effect is insignificant but still negative. The only positive effect occurs for the highest advice group, but is far from significant.² The effect of early tracking gradually decreases for higher ability levels, which is in line with previous literature (e.g. Hanushek and Woessmann (2006)). The coefficients in the next two columns suggest that this reduction in the completion of higher education can be attributed to higher completion rates of higher professional education, rather than completion of university. Effects are negative here as well, but never reach significance. The sign change occurs one category earlier than for higher

¹The fact that this is also present in our subgroup estimates indicates that there is also endogeneity within these groups. Hence, the better students with T3 advice are more likely to be tracked early than students of relatively lower ability with the same advice.

²This can be partly due to the fact that the vast majority of students in this group attains high education anyway. Moreover, the T3 advice group is relatively small.

education. The last two columns suggest that part of the negative impact occurs through an increase in dropout. Although the coefficients are not significant, they are consistently positive and close to significance for the overall sample and the T2 advice group for T2/T3 tracking and the T1/T2 advice group for T1/T2 tracking.

If we compare the columns with and without controls, we observe a very strong consistency in coefficients, for all outcome variables and both levels of tracking. All differences are neglectable and stay within half of a standard error even in the most extreme cases. This is suggestive evidence that our instrument is indeed exogenous, since it does not correlate strongly with those determinants of the outcome variable that are observable. There can still be a correlation between the instrument and unobservables, but given that our range of control variables is vast and offers multiple proxies for relevant unobservables, the fact that adding these controls has such a small effect over such a wide range of coefficients is a strong indication of instrument validity.

The estimates of Van Elk et al. (2011) are not easily comparable to our results. They employ a different age of what constitutes early tracking¹ and focus on one specific subgroup for which our instrumentation is weak.² Their IV-estimates suggest a decrease of 13 percentage points in the completion of higher education for the T1b advice group. These estimates are even higher in the robustness analyses, with a maximum effect of 21 percentage points. This is roughly equal to our overall estimate for T1/T2 tracking, as well as our subgroup effect for the T1b advice group. However, the instrumentation is weak for the latter. We can estimate the effect for this subgroup in a less robust way, constructing the same instrument using only the sample data. This results in an effect of early tracking for those with T1b advice of -0.066,³ or a lowering of higher education completion of 6.6 percentage points. This estimate is, although still negative, clearly lower than the one in Van Elk et al. (2011).⁴

¹We employ before age 14 as early tracking, while Van Elk et al. (2011) use age 13.

²Because the vast majority of students in this particular subgroup is tracked early, the instrumentation is weak both when we choose age 12 as a cutoff or age 13. This is exacerbated by the discrepancy between the frequency of early tracking in the sample, and the nationwide average where our instrumentation is based on. Van Elk et al. (2011) construct the instrument with the sample at hand, which increases their first stage power.

³Significant at the 1% level.

⁴The mean difference between both groups is 10 percentage points for the subgroup with T1b advice. Observables are slightly in favor of comprehensive students which suggests that the ATE will be below 10 percentage points. Still, the LATE could exceed this.

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

6.4.2 Labor market outcomes

Table 6.5 presents results for several outcomes relating to labor market performance. The effects of early tracking on these outcomes are largely insignificant, although we do observe some statistically significant coefficients. Impacts are stronger for T2/T3 tracking than for T1/T2 tracking. For the former group, there is an effect on reported earnings and on employment, but not on corrected earnings. This indicates that the effect of early tracking is mainly due to higher employment and hours worked, since it disappears when we exclude low earnings and correct for part-time factor. The effect originates mainly from the subgroup of students with a T2 advice. The earnings of the T2/T3 advice group are negatively affected by early tracking as well, while the impacts on educational attainment are low and insignificant for this subgroup. This could be indicative evidence of the effect of peer quality on later earnings. We can say this subgroup is clearly worse off, since the lower earnings are not compensated by a gain in time from a shorter educational life.

The labor market effects are much less pronounced for T1/T2 tracking. The effects on educational attainment do not seem to have translated (yet) in earnings at age 30. However, first stage power is too low to make strong statements with respect to the labor market impacts of early T1/T2 tracking. The results again show that the only positive coefficients are present for the very highest subgroup, but these effects do not reach significance. We also estimate the effect of early tracking on earnings in the latest reported year (2007). These estimates (not shown) are highly similar as those for the mean earnings over 2004 to 2007. They are slightly more negative for both levels of tracking (with the exception of the highest subgroup in each case, where the effects become more positive), which is what we would expect when age increases. We also see here that adding controls has a very small impact on coefficients, for all outcome variables and both levels of tracking.

We also examine whether someone was ever on welfare from 1999 to 2007. The effect for T1/T2 tracking is negative, but insignificant. Table 6.1 showed that the mean difference was significant and higher for the comprehensive group. Given the fact that comprehensive students are of higher ability, this suggests that welfare receipt is higher when tracked late, which is in contrast to results for educational attainment.¹ The overall effect size is insignificant for T2/T3 tracking but this hides a significantly positive effect for the T2 advice group² and a significantly

¹The fact that IV estimates are insignificant is mainly driven by relatively high standard errors. The effect sizes are larger than the mean difference in table 6.1, as one would expect given the direction of the bias.

²Although the significance disappears when controls are added.

negative effect for the T3 advice group. This confirms earlier patterns of how the effect of early tracking develops over the distribution, although this is the only variable that is significantly affected for the highest advice group.

Our estimation of labor market outcomes is weakened by the fact that we do not use the 1993 cohort in the previous analyses. We also estimated results when both cohorts are included. We only use earnings for the most recent year, 2007, in this estimation. The 1993 observations are 26 years old then. We obtain estimates (not shown) that are lower in magnitude than when we only use cohort 1989. This is not surprising, since the students from comprehensive classes have a higher probability of still being in education. Still, the effect for the T1/T2 advice group in T2/T3 tracking is significantly negative. The effect for the highest advice group is positive and close to significance. Given the fact that the early tracked students are *more* likely to still be in education here,¹ this effect could become even larger once all students have entered the labor market. All other subgroups are not significantly affected, but this could change when individuals mature in the labor market.

6.4.3 Magnitudes

Compared to the OLS estimates from table 6.3, the IV-estimates are more negative for T2/T3 tracking. This is exactly what we would expect when high-ability students are more likely to go to a specific track early on. On the other hand, the IV estimates are more negative for T1/T2 tracking as well, where the observable characteristics of the comprehensive students are clearly better. However, OLS estimates refer to an average treatment effect, while the IV coefficients are the local average treatment effect (Imbens and Angrist, 1994). This is the effect of early tracking on those students who would respond to a change in the supply of schools that track early. It is expected that the students who have the ambition to enter higher education, are the ones that would respond strongly to an increase in comprehensive classes. These classes increase the opportunity of attending high education. Even considering that our effects are LATE, the impact of early tracking on the T1/T2 level on high education are large. A large part of this effect could be driven by the institutional design. For students in the T1b track, which constitutes the large majority of T1/T2 students, it is less straightforward to enter higher education.² The path to higher education is more direct when they first go to a comprehensive track. We have seen

¹Based on the insignificant but positive estimates from table 6.4.

²They can still achieve higher education, if they follow the T2 track after the T1b track, or when they complete the highest level of lower vocational tertiary education. This ensures that still 22 (1989 cohort) to 27 (1993 cohort) percent of T1b students ultimately completes higher education.

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

that the advices given at the end of primary school are conservative; a lot of students that go to comprehensive classes end up in higher tracks.

We also saw that there is large upstream for students with a T2 advice from primary school. This can explain why the results for T2/T3 tracking are especially large for this subgroup (along with the T1/T2 subgroup). Early tracking has strong negative effects for this group on completion of higher education, earnings and employment, while the treatment effects for completion of university, dropout and welfare receipt are not far from being significant. It is likely that a large part of these effects can be attributed to misallocation from early selection. More than 30% of students with a T2 advice still go to Track 3. Naturally, such upstream is less likely when selected early.

We do not find significantly positive estimates of early tracking for any subgroup. However, our results are indicative that early tracking at the T2/T3 level increases outcomes for the T3 advice group. This subgroup is small ($n=724$) and the first stage is not strong. The effects for these students are consistently positive and close to significance in some cases. Moreover, they become larger when we use more recent earnings estimates and estimates are significant for welfare receipt, which has a more powerful first stage. Since these are highly educated individuals, the labor market gains from early tracking can strongly increase after age 30 for this subgroup.

6.5 Robustness

We have seen that adding control variables has a very small effect on our estimates, which is evidence in favor of the validity of our instruments. In addition, we check whether our model finds any estimates for variables that should not be affected. We use all three domains of the Entrance Test at the beginning of secondary education and whether one of the parents completed one of the highest two educational attainment levels.¹ We find that the effects of early tracking on these variables is far from significant, as it should be. The t-statistics are all lower than 1, both with and without controls and for T1/T2 and T2/T3 tracking.

One possible concern is that the supply of early tracking schools can be driven by the demand for tracking. Schools in municipalities with students of a higher than average level might be induced to offer early tracking if these better able students and their parents have a preference for early tracking. We include two variables in our model that are indicators of such differences

¹Naturally, these variables cannot be used as controls in the respective regressions.

in the demand for early tracking schools. These are the mean Entrance Test score and mean parental education level, both defined per municipality. Table 6.6 shows that the effect of adding these variables on coefficients is virtually non-existent.

As an alternative, we use the supply of early tracking schools by municipality of residence, rather than municipality where the school resides, as an instrument. We employ the latter as an instrument in our main analysis, because there are multiple municipalities without any school that offers the specific track. Hence, using municipality of residence weakens our first stage. Potentially, it could be that individuals of a specific ability level are more likely to commute to a school outside of their municipality of residence. This can induce a correlation between the instrument and unobserved ability. If such behavior occurs in practice, this should lead to differences in estimates from using these different instruments. Estimates with the alternative instrument are portrayed in columns 2 and 6 of table 6.6. Since instrumentation is weaker in this case, standard errors are higher. The coefficients are comparable, only the estimates for higher education completion for T1/T2 tracking are higher under the alternative specification. This indicates that mobility between municipalities does not drive our results.

Van Elk et al. (2011) address robustness by adding the change in the supply of early tracked schools over time as an extra control. This change might reflect changes in the demand for these schools. However, our instrument is based on national figures from 1997 and has the same value for both cohorts. We can construct this variable based on the supply from the sample at hand. This reduces the number of observations and first stage power considerably, because not all municipalities occur in both cohorts. Therefore, the coefficients from column 4 in table 6.6 are not the same as the original estimates. Comparing to this benchmark, the effect of adding this ratio is very minimal.

Our instrument proves to very robust to any robustness analysis employed. This is somewhat in contrast with other studies that use Instrumental Variables when estimating the effects of tracking. Manning and Pischke (2006) show that the use of political constituency as an instrument in Galindo-Rueda and Vignoles (2005) is invalid, since the Conservative areas are relatively more wealthy and have better performing students. Moreover, the approach used in Galindo-Rueda and Vignoles (2005) also gives significant estimates on predetermined variables. This is not the case for our instrument, which does not lead to significant estimates for variables that should be unaffected. Moreover, it is not correlated with any of the control variables we employ, while the estimates from Van Elk et al. (2011) change significantly when controls are

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

added. This suggests that areas with different density or political constituency will differ more systematically than municipalities with a different relative supply of early tracking schools.

6.6 Conclusion

We have identified that early tracking leads to negative impacts on educational attainment and labor market outcomes. These impacts are more pronounced for those with lower ability. Treatment effects of early tracking decrease in magnitude as we move up in the distribution, but they are only positive for the very highest subgroup. Both tracking at the T1/T2 level as well as on the T2/T3 level lead to negative treatment effects. The effects are stronger for educational attainment for the former, but our instrumentation at the T1/T2 level is not powerful enough to make a similar claim with respect to labor market effects.

We use an instrumental variable approach where we exploit difference in tracking age across municipalities to identify the effect of early tracking. The relative supply of early tracking schools proves to be a strong predictor of early tracking and multiple analyses strongly suggest that it is not correlated with any other determinants of our outcome variables. The effect of adding a large range of controls is virtually non-existent, as is the adding of the mean level of parental education and mean Entrance Test score on the municipal level. Our model also gives very insignificant results when we employ pre-determined outcome variables. The strong robustness of our instrumentation is in contrast with other studies that use IV-approaches in relation to tracking.

The fact that early tracking negatively affects educational attainment does not immediately imply it has a detrimental overall effect. By construction, early tracking puts some subgroups of students already in a track where the path towards completion of higher education is more complicated. However, the reason to put them in such tracks is because they are not deemed fit for a path that leads to higher education. Although the payoffs of attending higher education are generally large and positive, they do not have to be so for every single individual, and maximizing the fraction of higher education degrees should not be a goal in itself. Still, our estimates suggest that the impacts on educational attainment are translated into lower earnings as well, at least for T2/T3 tracking. This works mainly through a lowering of employment and hours worked. Whether higher earnings compensate sufficiently for the extra time spent in school depends on the utility function of the individual. The effects we find can occur through changes in peer quality, or a different allocation of students towards tracks when selection is postponed.

Our estimates for earnings at the T1/T2 level lack the statistical power to draw strong conclusions with respect to the labor market effects for this specific level of tracking. Moreover, we do not have earnings data beyond age 30. Future studies could focus more on estimating effects of the tracking age on earnings, for a larger sample of individuals and a larger age range, to get a better indication of the lifetime consequences of early tracking. Additionally, our study cannot distinguish whether the estimated impacts of early tracking should be attributed to peer effects, curriculum effects or teacher sorting. Nonetheless, the results indicate that providing more comprehensive classes will increase the share of students that obtains a higher education degree, which translates into higher earnings as well.

6.7 Figures and tables

Figure 6.1: Ultimate track assignment per advice group

Notes: The graph shows the ultimate track assignment per track advice received for secondary education. Advices are given by the students' primary school at the end of primary education. The indicators on the horizontal axis indicate the track that was advised for the particular student. Advices can be direct or mixed.

Figure 6.2: Ultimate educational attainment per advice group

Notes: The graph shows the highest educational level completed per track advice received for secondary education. The indicators on the horizontal axis indicate the track that was advised for the particular student.. Advices are given by primary school at the end of primary education. Data on educational attainment are available until age 26 and 30 for the 1993 and 1989 cohorts, respectively. Advices can be direct or mixed.

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

Figure 6.3: Distribution of early tracking instrument: T1/T2

Notes: The graph shows the distribution of instrument values, for tracking at the T1/T2 level. The instrument represents the relative supply of schools that track early (before age 14) in the municipality where the school resides.

Figure 6.4: Distribution of early tracking instrument: T2/T3

Notes: The graph shows the distribution of instrument values, for tracking at the T2/T3 level. The instrument represents the relative supply of schools that track early (before age 14) in the municipality where the school resides.

Figure 6.5: Value of instrument by mean test score in municipality

Notes: The Graph represents the relation between the instrument and the mean Entrance Test score in the municipality, for both tracking at the T1/T2 level and T2/T3 level. The Entrance Test is taken at the beginning of secondary education.

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

Table 6.1: Comparison of variables for different age of tracking

	T1/T2				T2/T3			
	Mean Comp.	Mean ET	Difference	P-value	Mean Comp.	Mean ET	Difference	P-value
Female	0.505	0.530	-0.025	0.003	0.520	0.521	-0.0015	0.881
Big 4 Cities	0.058	0.030	-0.028	0.000	0.051	0.058	0.0067	0.123
Non-Dutch	0.101	0.091	0.010	0.040	0.099	0.092	0.0067	0.240
Parent High Education	0.258	0.191	0.067	0.000	0.360	0.423	-0.063	0.000
Parent Low Education	0.074	0.098	-0.024	0.000	0.053	0.034	0.019	0.000
Lowest Social Class	0.105	0.099	0.0069	0.178	0.094	0.085	0.0097	0.078
Highest Social Class	0.178	0.140	0.037	0.000	0.239	0.282	-0.043	0.000
Intelligence Test Score	0.166	0.028	0.139	0.000	0.339	0.363	-0.024	0.136
Language Subscore	13.36	12.31	1.05	0.000	14.83	15.40	-0.567	0.000
Math Subscore	13.05	11.60	1.45	0.000	15.04	15.59	-0.554	0.000
Information Subscore	13.45	12.33	1.11	0.000	15.05	15.54	-0.489	0.000
Track 1	0.418	0.807	-0.389	0.000	-	-	-	-
Track 2	0.581	0.192	0.389	0.000	0.555	0.404	0.151	0.000
Track 3	-	-	-	-	0.445	0.596	-0.151	0.000
Mean Earnings (logs)	7.444	7.449	-0.0049	0.618	7.452	7.389	0.062	0.000
Earnings 2007 (logs)	7.671	7.627	0.044	0.000	7.738	7.708	0.030	0.015
Corrected Mean Earnings (logs)	7.719	7.703	0.016	0.002	7.762	7.747	0.015	0.032
Higher Education	0.457	0.299	0.157	0.000	0.692	0.706	-0.014	0.113
University	0.081	0.045	0.036	0.000	0.280	0.362	-0.083	0.000
Dropout	0.065	0.073	-0.0083	0.056	0.060	0.063	-0.0027	0.559
Employment	0.884	0.862	0.022	0.000	0.892	0.862	0.031	0.000
Welfare	0.086	0.069	0.016	0.000	0.088	0.040	0.048	0.000

Notes: The table shows the means of comprehensive (Comp.) and early tracked (ET) students, for both T1/T2 as T2/T3 tracking. The table lists p-values of a two-sided Wald test conducted on the difference between means. ‘Big 4 cities’ indicates whether the child lived in one of the big 4 cities in the Netherlands. Parent High Education is a dummy variable that takes value 1 when the highest education level of the parents is in one of the top two categories (six categories are defined). Parent Low Education takes value 1 when the highest education level of the parents is in one of the lowest two categories. Social class is based on the occupational status of the parents. All four test scores are taken at the start of the first year of secondary education. The Track variables indicate the fraction of students that end up in this specific track. Earnings are in logs. Corrected earnings are corrected for part-time factor, topcoded at 10,000 euro’s, and earnings below 1000 euro’s are excluded. Higher education consists of higher professional education (*hbo*) and university education. Employment measures whether the individual earned at least a minimum wage in 2007. Welfare measures whether the individual ever received welfare benefits between 1999 and 2007.

Table 6.2: First stage results (F-values)

	Educational Attainment		Labor Market Outcomes	
<u>T1/T2 Tracking</u>				
All	21.73	26.33	5.03	6.68
T1a/T1b	9.01	8.64	2.66	2.61
T1b	7.53	8.94	0.42	0.89
T1/T2	13.73	14.26	0.72	0.87
T2	33.52	36.85	12.72	14.74
T2/T3	11.65	12.49	2.74	3.33
Controls		Yes	Yes	
<u>T2/T3 Tracking</u>				
All	64.94	67.14	30.13	34.67
T1b	20.92	23.13	4.34	5.06
T1/T2	31.67	32.47	11.45	12.78
T2	70.36	73.62	39.47	43.84
T2/T3	42.18	42.06	33.26	37.48
T3	12.34	13.16	5.31	7.87
Controls		Yes	Yes	

Notes: The table shows the F-values of a Wald test of the instrument in the first stage of the regression, either with or without controls. The test is conducted for tracking between the low and the medium track (T1/T2), and medium and high track (T2/T3). The left column indicates the subgroup for which the effect is estimated. The indicators refer to track advice they received at the end of primary school, which can be either specific or mixed. The degree of Early Tracking in the municipality is used as an instrument for Early Tracking. Control variables are: month of birth, gender, ethnic origin, social class, highest education level of the parents, whether the child lived in one of the big 4 cities in the Netherlands, the score on an intelligence test taken in year 1 of secondary school and the scores on all three test domains of the Entrance Test. F-values are lower for labor market outcomes, since estimation only includes the 1989 cohort. The treatment effects on use of welfare contains both cohorts and, correspondingly, F-values are equal to those for the educational attainment columns.

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

Table 6.3: OLS estimates of early tracking

	High Education		Mean Earnings	
T1/T2 Tracking				
All	-0.155*** (0.015)	-0.097*** (0.012)	-0.053*** (0.016)	-0.021 (0.014)
T1a/T1b	-0.026 (0.033)	-0.013 (0.035)	0.019 (0.045)	0.0098 (0.042)
T1b	-0.094*** (0.020)	-0.082*** (0.017)	-0.035 (0.026)	-0.016 (0.026)
T1/T2	-0.042* (0.023)	-0.028 (0.022)	-0.0065 (0.036)	-0.0086 (0.035)
T2	-0.104*** (0.026)	-0.086*** (0.024)	0.045 (0.031)	0.044 (0.028)
T2/T3	-0.030 (0.036)	-0.017 (0.034)	-0.025 (0.038)	-0.038 (0.038)
Controls	Yes		Yes	
T2/T3 Tracking				
All	0.010 (0.014)	-0.010 (0.011)	-0.011 (0.017)	-0.021 (0.017)
T1b	-0.084* (0.046)	-0.076* (0.040)	-0.017 (0.065)	0.0026 (0.069)
T1/T2	-0.041 (0.030)	-0.033 (0.030)	-0.081 (0.049)	-0.061 (0.046)
T2	-0.029 (0.021)	-0.032 (0.019)	-0.026 (0.030)	0.021 (0.029)
T2/T3	0.00061 (0.017)	-0.0076 (0.017)	-0.051** (0.024)	-0.071*** (0.023)
T3	0.033 (0.022)	-0.016 (0.021)	0.123*** (0.039)	0.118** (0.040)
Controls	Yes		Yes	

Notes: Table shows the OLS estimates of early tracking on the specified outcome variables, either with or without controls. Estimation is conducted for tracking between the low and the medium track (T1/T2), and medium and high track (T2/T3). High education refers to the completion of either hbo (higher professional education) or university education. Earnings are the average of reported monthly pre-tax earnings from age 27-30 (2004 to 2007). Effects of tracking on earnings are estimated on the 1989 cohort only. See table 6.2 for a description on the indicators in the left column and the controls included.

Table 6.4: IV estimates of early tracking: educational attainment

	High Education		University		Dropout	
<u>T1/T2 Tracking</u>						
All	-0.163** (0.066)	-0.172*** (0.050)	-0.0078 (0.026)	-0.012 (0.021)	0.037 (0.026)	0.032 (0.023)
T1a/T1b	-0.121 (0.211)	-0.211 (0.215)	-0.038 (0.043)	-0.041 (0.045)	0.057 (0.102)	0.074 (0.114)
T1b	-0.121 (0.136)	-0.192* (0.115)	-0.018 (0.042)	-0.026 (0.039)	0.000035 (0.067)	0.020 (0.055)
T1/T2	-0.147 (0.091)	-0.145* (0.090)	0.021 (0.046)	0.017 (0.044)	0.071 (0.045)	0.068 (0.044)
T2	-0.117 (0.067)	-0.091 (0.061)	0.00055 (0.042)	-0.0050 (0.040)	0.027 (0.032)	0.012 (0.029)
T2/T3	-0.047 (0.103)	-0.092 (0.099)	0.044 (0.065)	0.041 (0.056)	0.030 (0.051)	0.043 (0.050)
Controls	Yes		Yes		Yes	
<u>T2/T3 Tracking</u>						
All	-0.082*** (0.030)	-0.082*** (0.025)	-0.0025 (0.041)	-0.018 (0.033)	0.022 (0.015)	0.017 (0.013)
T1b	-0.221** (0.106)	-0.173* (0.097)	-0.157** (0.067)	-0.143** (0.058)	0.023 (0.055)	0.0016 (0.045)
T1/T2	-0.139* (0.078)	-0.134* (0.079)	-0.0027 (0.043)	-0.017 (0.041)	0.024 (0.029)	0.030 (0.029)
T2	-0.120*** (0.043)	-0.100*** (0.037)	-0.048 (0.038)	-0.039 (0.036)	0.041 (0.026)	0.026 (0.023)
T2/T3	-0.030 (0.035)	-0.042 (0.037)	0.024 (0.047)	0.0014 (0.048)	0.0073 (0.019)	0.011 (0.019)
T3	0.018 (0.075)	0.018 (0.070)	0.015 (0.105)	-0.010 (0.092)	0.013 (0.029)	0.0077 (0.027)
Controls	Yes		Yes		Yes	

Notes: Table shows the IV estimates of early tracking on several measures of educational attainment, either with or without controls. Estimation is conducted for tracking between the low and the medium track (T1/T2), and medium and high track (T2/T3). See table 6.1 for an explanation of outcome variables. See table 6.2 for a description on the indicators in the left column, the method of instrumentation and the controls included. The first stage is too weak for estimation of subgroups T1a/T1b and T1b for T1/T2 tracking (see table 6.2), but these effects are reported for completeness.

6. THE IMPACT OF EARLY TRACKING ON LATER-LIFE OUTCOMES: AN INSTRUMENTAL VARIABLE APPROACH

Table 6.5: IV estimates of early tracking: labor market effects

	Earnings		Corrected Earnings		Employment		Welfare	
T1/T2 Tracking								
All	-0.0085 (0.126)	0.026 (0.090)	0.081 (0.078)	0.081 (0.060)	0.0086 (0.069)	0.0068 (0.059)	-0.028 (0.028)	-0.036 (0.025)
T1a/T1b	-0.405 (0.380)	-0.448 (0.469)	-0.197 (0.216)	-0.258 (0.237)	-0.182 (0.256)	-0.154 (0.263)	0.048 (0.117)	0.042 (0.122)
T1b	0.972 (1.57)	0.630 (0.823)	0.801 (1.13)	0.513 (0.551)	0.735 (1.28)	0.423 (0.604)	-0.143 (0.097)	-0.137 (0.087)
T1/T2	0.371 (0.658)	0.070 (0.341)	-0.133 (0.281)	-0.035 (0.170)	0.203 (0.276)	0.269 (0.286)	-0.013 (0.037)	-0.023 (0.036)
T2	0.048 (0.089)	0.036 (0.082)	0.078 (0.057)	0.078 (0.055)	-0.051 (0.059)	-0.053 (0.059)	-0.0015 (0.026)	-0.013 (0.025)
T2/T3	-0.218 (0.306)	-0.177 (0.235)	0.066 (0.149)	0.043 (0.131)	-0.193 (0.163)	-0.160 (0.122)	-0.022 (0.041)	-0.026 (0.041)
Controls	Yes		Yes		Yes		Yes	
T2/T3 Tracking								
All	-0.091** (0.040)	-0.077** (0.036)	-0.0087 (0.028)	-0.00025 (0.026)	-0.054** (0.021)	-0.053*** (0.019)	-0.0010 (0.010)	-0.0062 (0.0098)
T1b	-0.160 (0.241)	-0.101 (0.203)	0.021 (0.097)	0.031 (0.092)	0.035 (0.117)	0.054 (0.126)	-0.019 (0.045)	-0.039 (0.044)
T1/T2	-0.208 (0.158)	-0.181 (0.136)	-0.043 (0.067)	-0.020 (0.060)	-0.104* (0.060)	-0.095* (0.055)	-0.0044 (0.022)	-0.013 (0.021)
T2	-0.106* (0.059)	-0.095* (0.055)	-0.0054 (0.034)	-0.00029 (0.033)	-0.085*** (0.029)	-0.083*** (0.028)	0.026* (0.014)	0.017 (0.013)
T2/T3	-0.101 (0.067)	-0.110* (0.061)	-0.022 (0.046)	-0.030 (0.042)	-0.041 (0.031)	-0.048 (0.030)	0.0037 (0.017)	0.0035 (0.018)
T3	0.310 (0.187)	0.298 (0.178)	0.104 (0.099)	0.111 (0.100)	0.094 (0.081)	0.090 (0.074)	-0.067* (0.041)	-0.066* (0.040)
Controls	Yes		Yes		Yes		Yes	

Notes: Table shows the IV estimates of early tracking on several measures of labor market success, either with or without controls. Estimation is conducted for tracking between the low and the medium track (T1/T2), and medium and high track (T2/T3). Effects are estimated on the 1989 cohort only. See table 6.1 for an explanation of outcome variables. See table 6.2 for a description on the indicators in the left column, the method of instrumentation and the controls included. The first stage is too weak for estimation of all but the T2 subgroup for T1/T2 tracking and for the T1b and T3 subgroups for T2/T3 tracking (see table 6.2). These estimates are reported for completeness. For welfare, this only applies to subgroups T1a/T1b and T1b for T1/T2 tracking.

Table 6.6: Robustness to alternative specifications

	Original Estimates	Residence Municipality	Add municipal ability	Comparison Supply Ratio	Add change Supply Ratio
<u>T1/T2 Tracking</u>					
Higher Education	-0.082** (0.025)	-0.086*** (0.029)	-0.085*** (0.024)	-0.039 (0.092)	-0.040 (0.087)
University	-0.018 (0.033)	0.0065 (0.0036)	-0.020 (0.032)	-0.143 (0.149)	-0.140 (0.147)
Dropout	0.017 (0.013)	0.014 (0.015)	0.018 (0.013)	0.019 (0.043)	0.019 (0.043)
Earnings	-0.077** (0.036)	-0.101* (0.053)	-0.077** (0.034)	-0.417 (0.303)	-0.463 (0.369)
Corrected Earnings	-0.00025 (0.026)	-0.017 (0.033)	-0.00054 (0.026)	-0.109 (0.100)	-0.131 (0.116)
Employment	-0.053*** (0.019)	-0.063** (0.026)	-0.053*** (0.019)	-0.169 (0.121)	-0.163 (0.128)
Welfare	-0.0062 (0.0098)	0.0067 (0.011)	-0.0050 (0.0096)	-0.029 (0.052)	-0.029 (0.052)
<u>T2/T3 Tracking</u>					
Higher Education	-0.172*** (0.050)	-0.276*** (0.079)	-0.178*** (0.050)	-0.095 (0.085)	-0.068 (0.090)
University	-0.012 (0.021)	-0.027 (0.031)	-0.011 (0.020)	-0.042 (0.032)	-0.040 (0.033)
Dropout	0.032 (0.023)	0.034 (0.034)	0.029 (0.022)	0.055 (0.051)	0.046 (0.057)
Earnings	0.026 (0.090)	0.170 (0.230)	0.034 (0.091)	-0.240 (0.347)	-0.131 (0.205)
Corrected Earnings	0.081 (0.060)	0.074 (0.126)	0.081 (0.059)	0.093 (0.204)	0.056 (0.156)
Employment	0.0068 (0.059)	0.120 (0.162)	-0.0045 (0.059)	0.0029 (0.190)	0.039 (0.157)
Welfare	-0.036 (0.025)	-0.010 (0.036)	-0.036 (0.024)	0.00058 (0.042)	-0.016 (0.043)

Notes: Table shows the robustness of the coefficients from tables 6.4 and 6.5 to various alternative specifications. The columns report, in order: the original estimates; estimates when using the relative supply of Early Tracking schools by municipality of residence (in stead of municipality where the school resides); estimates when adding mean Entrance Test Score and mean Parental Education by municipality; estimates with the original model on a sample of municipalities that are prevalent in both cohorts; estimates on the same sample, adding the change in the supply-ratio from 1989 to 1993 cohort. All reported coefficients are based on specifications using all controls. See table 6.2 for a description on the original method of instrumentation and the controls included. See table 6.1 for an explanation of outcome variables.

The Impact of Grade Retention: An IV approach to Estimating Heterogeneous Treatment Effects¹

7.1 Introduction

Grade retention is employed in a number of countries to deal with students that lack the skills they should master at a certain grade level. There has always been a lot of debate on whether retention really benefits the student in the long run. Students could benefit from further developing more basic skills before they progress to a next grade. On the other hand, one can question whether a student will really catch up if he is simply required to do the same material again. Moreover, there can be negative psychological consequences from the stigma of being a repeat student and being separated from peers. There are strong voices both in favor and against employing grade retention. But the effect of being retained will be very different for different type of students. The relevant question is therefore not whether retention is good or bad, but who exactly benefits and loses from its implementation.

The aim of this study is to estimate the effects of grade retention and delayed starts in kindergarten and see how these effects differ for different people. We use an instrumental variable approach to estimate the impact of both these practices and combine IV with a quantile regression model to estimate effects by percentile of the distribution. We use PISA data from the 2003 and 2009 cohorts to estimate this. These waves include information on whether students were re-

¹This chapter is based on Diris (2012).

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

tained during formal education (Grade Retention) or simply entered first grade a year later (Late Starts), which is a crucial distinction. We find that both practices lead to negative effects for the sample as a whole, but retention during formal grades is much more harmful. Grade retention has negative impacts for the whole distribution while late starts can have positive effects for those near the bottom. Negative effects are larger the later retention is executed. We also show that these results are not driven by differences in the curriculum and that those that experience Grade Retention perform relatively worst at the easiest questions.

Previous research has generally estimated the overall effects of retention. However, the idea of the whole concept is that it is supposed to be effective for those pupils that are lagging behind. Overall effects are informative because they tell us the effect for the average repeat student, which serves as a good indication whether retention is exercised too often. However, negative overall effects do not justify abandoning the whole practice. If retention becomes more harmful the more we move upward in the distribution, one ideally wants to find that student for which retention starts to be harmful. Effects can also differ by the timing of retention. We especially make a distinction between retention in kindergarten and thereafter. But also the distinction between repeating in primary school or (early) in secondary school is important.

Retention does not occur by random assignment and can be highly correlated with the child's ability level and a number of other variables that also affect test scores. Previous research has consistently found negative impacts for grade retention. However, these studies generally add control variables to an OLS model or use matching techniques to estimate treatment effects. This is unlikely to fully remove endogeneity, and presumably leads to a negative bias in the estimated coefficients. We use an instrumental variable approach, where we employ the fraction of students within a certain month of birth that were retained (net of one's own value) as an instrument for actual retention.

This chapter proceeds as follows. Previous literature on this topic is discussed in section 7.2. Section 7.3 concerns the PISA data we use in our estimation. Section 7.4 specifies the methodology, focusing on endogeneity and heterogeneity. Results are discussed in section 7.5; further robustness issues in section 7.6. Section 7.7 concludes.

7.2 Literature

The general trend in the empirical literature is that retention is not a fruitful approach for dealing with learning difficulties and that it often leads to negative outcomes in terms of student perfor-

mance and behavior.¹ Evidence of any positive (long-term) effects on academic performance is virtually absent. Studies differ in the comparison group they employ. The vast majority of studies use both same-grade and same-age comparisons. Same-age comparisons generally produce significantly negative effects. Same-grade comparisons often produce positive effects in the short-run (although not always significantly positive) but these effects diminish or disappear over time (Peterson et al., 1987). Some studies still find negative effects on a grade level (although often insignificant) (McCoy and Reynolds, 1999; Pagani et al., 2001b). There is some evidence for positive effects of retention in the domain of perceived competence (Reynolds, 1992). However, pupils are likely to rate that competence with respect to their same-grade peers rather than their same-age peers. Studies on grade retention generally use matching to partially control for endogeneity. The variables they match on and the specific matching techniques differ by study. Matching does not completely remove all bias when retained and promoted students differ in unobservables. Results will still be biased towards the direction of negative retention effects.

There is no clear consensus on how treatment effects differ by timing of retention. Peterson et al. (1987) are one of the few that find more positive effects for older children, while McCoy and Reynolds (1999) find no significant difference between early and late retention. The majority of studies points towards better, or less harmful, implications the earlier retention takes place. For example, Mantzicopoulos and Morrison (1984) show convincing evidence that retaining in kindergarten produces effects that are lower compared to retaining during formal education. Retention in kindergarten even leads to significantly positive effects in the short-run. This confirms our assumption from section 2.4 that earlier retention has more potential to be effective, which could be driven by complementarity, self-productivity and lower adverse psychological effects. We estimate effects of retention for different ages, and specifically make a distinction between retention in kindergarten and retention thereafter, throughout our empirical analysis.

Clearly, the empirical evidence rejects retention as a productive practice in raising pupils' performance. Then why is it still so prevalent in almost every educational system? The empirical evidence does not seem to match with the beliefs of schools, parents and teachers, who in the end decide whether a child is retained or not. Studies by Tomchin and Impara (1992) and Smith and Shephard (1988) provide a good overview on teachers' beliefs towards retention. There is a lot of diversity in beliefs among teachers. There is some consensus that retention can be justifiable in the early grades, but there is a lot of disagreement on its effectiveness in later grades. On

¹See Holmes and Matthews (1984); Jackson (1975); Karweit (1991) for an overview.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

the other hand, beliefs are very similar within the same school. This could indicate an influence of school boards on teachers' beliefs or that teachers with similar beliefs select into the same schools. Tomchin and Impara (1992) report that parents believe that teachers underestimate the negative effects that retention can bring. Teachers that are in favor of promotion give a number of reasons for this view. Not only do they think that retention is a good practice for those who are not deemed ready to move on, they also see the threat of retention as an effective motivator. Moreover, retention can be behaviorally oriented rather than academically.

There are a lot of institutional factors that can increase the frequency of retention. Primarily, retention can be used to inflate average test scores for a school or a class (Karweit, 1991). By keeping a child in a specific grade for an additional year, he or she will generally score better on test scores because of the additional year of schooling and maturity. Moreover, teachers' beliefs towards retention tend to become self-fulfilling prophecies (Tomchin and Impara, 1992). If a retained student performs well after retention, this is seen as a confirmation that retention has helped; if the student performs badly, this is seen as confirmation that retention was necessary. Since we can never know how the student would have performed when promoted, there is no correction mechanism. Tomchin and Impara (1992) report that teachers tend to see retention as the safe choice whenever there is doubt about a students' performance.

7.3 Data

The data we use in this study stem from the international assessment test PISA, the same test on 15-year-olds that we employ in Chapter 3. We again include those countries that contain students who are born within one year of each other, in the twelve months after the formal cut-off date.¹ The countries included in our analysis are Belgium, Canada, Denmark, England, Finland, France, Italy, New Zealand, Poland, Portugal, Spain and Sweden. Table 7.1 shows mean rates of retention for all countries.

We use standardized achievement test scores on math, science or reading as dependent variables in our analysis, which are the mean of the five plausible values reported in the data. These scores are standardized to have a mean of 0 and a standard deviation of 1.² All PISA cohorts

¹This is a necessary condition for a country to be selected for our analysis. Countries who, for example, contain students from a specific calendar year while the cutoff date is not January the 1st are excluded. Our estimation requires that we look at cohorts that would be in the same grade in absence of any retention.

²PISA itself uses a mean of 500 and a standard deviation of 100 in its standardization. Since our sample of countries is slightly more affluent and less heterogeneous, our initial mean is higher and standard deviation lower.

also contain a large range of background questions (gender, country of origin, language spoken at home etc.).

There are two main reasons why the PISA data are chosen for our analysis. First of all, it is comprised of students who are born within one year of each other, which ensures we can compare repeaters and non-repeaters at the same age level, as is proper in this context. Secondly, PISA contains information on the stage where students were retained. The PISA data allow us to identify both repeaters in kindergarten (who we label as Late Starters) and those who repeat after (who we label as Grade Repeaters). We can identify the latter by looking at the answer to the repeat question in PISA, and the former by looking at those who are one grade behind, but were not being retained in any formal grade. The data also distinguish between retention during primary education (ISCED I) and early secondary education (ISCED II). Only students in the 2003 and 2009 cohorts report this information, hence we only use those two cohorts in our analysis.

7.4 Methodology

7.4.1 Comparison group

The fact that we do not observe one person in two different states, in this case retained and non-retained, means we need to compare retainers with some other group of non-retained students. One can conduct this comparison on an age level or on a grade level. Previous research differs strongly in this respect. Some compare with students in the same grade, some with students of the same age, while the majority does both.¹ It is obvious that choosing a grade-level comparison favors retained students compared to age-level comparisons. Retained students are older and have had more schooling in the former case. If one wants to know the effect of retention, one needs to find out the difference in performance when retained, compared to the performance when that same student would not have been retained. This comparison should take place at the same age, since the counterfactual should be of the same age as the actual observation.²

The comparability of test scores presents another issue. For the same reason that we would like to compare students at the same age level, we would also like to compare their performance on the same test questions. Some studies compare performance of non-retained students in grade

¹See Holmes and Matthews (1984); Jackson (1975); Karweit (1991) for an overview.

²A grade-comparison does not necessarily have to be improper. Only, when it is carried out, one has to weigh any effects against the cost it involves,. These consist of both extra educational expenditures and opportunity costs of losing a year in time (see Chapter 3).

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

X with performance of retained students in grade X-1 on grade-specific tests (e.g. Peterson et al. (1987)). Although these students are of the same age, the tests they make are not the same. The tests of the retained students are taken at a lower grade level and therefore easier. Essentially, the age-comparison becomes a grade-comparison since the tests pupils take are designed with respect to their grade-level peers. This biases results in the direction of positive retention effects.

Therefore, we should compare students of the same age, in different grades, taking the exact same test. One potential issue with this choice is that non-retained students, who are in a higher grade, might have discussed specific topics already which retained students of the same age have not. If such topics come back in the test, this can provide a disadvantage for retained students that can be seen as unfair. This is unavoidable, if we want to avoid the (presumably stronger) bias from differences in age and years of schooling that grade-level comparisons inhibit.¹ A student cannot be of the same age, have had the same material and still have a different treatment status for retention. Moreover, this issue is also better controllable, since we can often identify which topics have been discussed for everyone and can see whether effects differ for different types of questions.

7.4.2 Selection

The ‘assignment’ of being retained is not a random process. Retained students are of lower ability than their promoted peers, which means we cannot compare these groups one-to-one. Ignoring this issue will bias results towards rejecting retention as a productive practice. Many previous studies use matching to correct for this. Such matching is done by controlling for basic demographical variables, usually combined with pre-retention test scores.² Although this can reduce bias from endogeneity, it cannot correct for correlation between retention and unobservables. There should still be differences in unmatched variables, since treatment status would not be different if the two individuals are perfectly similar. Studies that either ignore ability bias or use matching to correct for it can overestimate the (negative) effects of retention on performance.

We estimate the effect of grade retention by using an Instrumental Variable method, that corrects for the endogeneity of retention. The nature of the endogeneity problem is that the decision to retain someone is based on factors that are also related to the future performance of the child. Month of birth is a strong determinant of retention, while we can assume it is exogenous to other

¹One might argue if this is really a bias or part of the genuine effect of grade retention. Retained students are always ‘behind’ in some way. Whenever they catch up, their non-retained peers have learned something new already, or built up work experience etc. This is simply a consequence of being retained.

²See e.g. Gomes-Neto and Hanushek (1994) and Coffield and Blommers (1956).

determinants of test scores. Month of birth essentially measures the maturity at the moment of a test. Children who are born in a later month and are, correspondingly, younger have a higher chance of being retained, *ceteris paribus*. Figures 7.1 to 7.3 show the pattern of retention by month of birth. There is a clear non-linear, increasing pattern in retention rates. This pattern is most prominent for Late Starts (LS). This is not surprising since the discussion to withhold a child in kindergarten for a year is most often carried out for those children that are born just before the school cutoff-date. For Grade Retention (GR), there is an increasing pattern, but it is weaker.¹

Hence, we use the fraction of students within a certain month of birth and within each country, that was retained (or started late) as an instrument for actual retention (or late start). Thus, each country-month combination has a specific value for the instrument (labeled w). These mean values are calculated net of one's own value for retention. The instrument is a strong predictor for Grade Retention and Late Starts, but should not be correlated with the error term. We also include month and country dummies in this model:

$$\begin{aligned}
 Retention_i &= \beta_0 + \beta_1 w_i + \sum_j \gamma_j Country_j + \sum_k \delta_k Month_k + \epsilon_i \\
 Score_i &= \beta_0 + \beta_1 Retention_i + \sum_j \gamma_j Country_j + \sum_k \delta_k Month_k + \epsilon_i \quad (7.1)
 \end{aligned}$$

Where *Retention* can either mean Grade Retention or Late Starts and β_1 is our coefficient of interest. We see first stage results in table 7.2. Because the pattern for Grade Retention is more constant than that for Late Starts, instrumentation is weaker there. Although, it passes the traditional criteria for first stage power, we merge both PISA cohorts to increase the strength of the instrument. F-values for the merged dataset are listed in the last row of table 7.2 and are stronger.²

7.4.3 Heterogeneity

Some papers (e.g. Jackson (1975)) have advocated that the only reliable studies of grade retention are experimental studies, where retention is randomly assigned. Although methodologically sound, we have to make strong reservations with respect to the interpretation of such studies.

¹Especially taking into account the smaller scaling on the y-axis.

²This also means that model 7.1 will be expanded by controlling also for PISA-cohort, plus interactions between PISA-cohort and month of birth, plus interactions between PISA-cohort and country-code.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

When assignment is truly random, these studies estimate the effect of retention for the average student.¹ If such effects are negative, this does not imply that retention is negative for every student in the class. The idea of retention is that we give the least-able students extra time to master the material for that grade. Even the strongest proponents of retention will not argue that half of a class should be retained.² The effects of retention are not universal. We would expect the most positive (or least negative) impacts to occur at the bottom of the distribution and a decrease in coefficients when we move upward.

We estimate how the effects of retention differ along the distribution. We give students a rank based on their score (relative, compared to everyone else in their country) and examine how retention affects that ranking for a range of deciles or percentiles of the distribution. This gives us a clear picture of any possible differences in effects for those with poor scores, who are expected to benefit more (or suffer less) from retention, versus those with good scores. Ranks are based on the fraction of students in that country that had a similar or higher score. Hence, if 60 percent of students scored equal or better, the corresponding rank is 0.4. Since the rank is based on the exact fraction of students that performs similar or better, it is very specific and not rounded to the nearest percentile. We estimate a quantile regression model, similar to Chapter 4 and based on Koenker and Bassett (1978). Everyone who is below the percentile for which we estimate the effect gets the value of the lowest threshold, and anyone above that percentile gets the value of the higher threshold. This means that if we, for example, estimate the effect for the tenth percentile, those with a rank below 0.09 get a value of 0.09, those above 0.10 get value of 0.10 and those in between 0.09 and 0.10 get their precise rank. We give an example of this model below, for the case where we assign 10 deciles:

$$\begin{aligned} Score1_i &= \min(0.1, score_i) \\ Score2_i &= \max(0.1, \min(0.2, score_i)) \\ &\cdot \\ &\cdot \\ Score10_i &= \max(0.9, score_i) \end{aligned}$$

¹This is the average student in the sample population, hence for a more deprived sample this could still be very valuable information. But it still neglects the strong heterogeneity in treatment effects for retention.

²This would simply mean the curriculum is improper.

To make coefficients interpretable, we have to rescale the group-effects. The method ‘divides’ the total effect over all separate quantiles. All individual effects add up to the total effect. If retention would be equally divided over the distribution, it would simply be a matter of multiplying by 10 for deciles and 100 for percentiles. This is not the case, which is why we instead need to divide by the number of retained students located in a certain quantile as a fraction of the total amount of retained students in the whole sample.¹

7.5 Results

7.5.1 Overall results

We first present OLS estimates, for comparative purposes, in tables 7.3 and 7.4. We standardized scores to have a mean of 0 and a standard deviation of 1. OLS gives effects of around -1.0 of a standard deviation for Grade Repeaters (GR) and -0.5 for Late Starts (LS). This reduces to -0.8 and -0.3, respectively, when we add controls (gender, parental education, country of birth of father and self, language spoken at home, number of books at home). When we use our IV model, effects are reduced to -0.5 for GR and are in the range of -0.05 to -0.20 for Late Starters (table 7.5). We also observe clear differences by subject. GR is especially bad for reading and less so for math, where LS is most harmful for mathematics and even insignificant (but still negative) for reading. The effects are consistent by cohort, with a modest exception for science scores.

The fact that IV estimates are clearly lower than OLS coefficients could be seen as a confirmation of the endogeneity problem discussed before, and underlines that adding controls does not solve this issue. However, IV estimates provide Local Average Treatment Effects (LATE) while OLS represents an average of all treatment effects (ATE) (Imbens and Angrist, 1994). IV shows the impact for those students whose treatment status would respond to a change in their month of birth. These students are the relatively higher-able repeaters.² Since negative treatment effects are larger for individuals higher in the distribution, the LATE will be above the ATE.

Table 7.6 provides IV estimates with the same controls as in table 7.4. Coefficients remain very stable, especially there where the instrumentation is strong. The estimates are extremely

¹I.e. if 50% of all repeaters are located in the lowest decile, we have to divide the initial coefficient for the first decile by 0.5. If only 2% of all repeaters are in the 7th decile, we have to divide by 0.02 for that decile, etc.

²They are not the very top segment of repeaters because they consist of those at the margin of being retained for every birth month. These are individuals very low in the distribution for early birth months and relatively high in later birth months. Still, the average percentile of all these observations is above that of the average repeater.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

similar for Late Starts especially.¹

7.5.2 Heterogeneity by percentile

These results suggest that there is a negative payoff for the repeater or late starter at the margin, but do not preclude that certain groups benefit from treatment. Retention is believed to be especially helpful for students at the bottom of the distribution. To address heterogeneity, we look at effects by decile and percentile. These effects are displayed in tables 7.7 and 7.8, for both Grade Retention and Late Starts, and for all test subjects. Figures 7.4, 7.5 and 7.6 show the effects by percentile for the three test topics for both independent variables graphically, in a more intuitively interpretable manner. The graphs plot the coefficients and 95% confidence intervals. We only plot the estimates for Grade Retention until the 20th percentile, since only very few retained students are located in higher percentiles, which makes these estimates very imprecise. This is much less the case for late starters, which is based to a stronger extent on month of birth and represented by pupils all over the distribution. Grade Retention is negative for the whole range of the distribution. The only positive coefficient occurs for the first percentile in reading, but it is very low and insignificant. From the second percentile on, all effects are significantly negative.

Late Starts benefit certain individuals. Reading skills are positively affected for a large group at the bottom of the distribution. Effects are positive until the 20th percentile and significantly positive up until the 10th percentile. Effects for science are positive for the lowest decile, although significance is low. We only observe significant negative effects beyond the 10th percentile for math, from the 30th percentile on for science and from the 45th-50th percentiles on for reading.

Coefficients show the effect that Grade Retention or Late Starts have on the within-country ranking for that specific topic, for those students that actually were retained. Hence, if the 4th percentile has a coefficient of -0.060, it means that Repeaters (or Late Starters) in that percentile would have been in the 10th percentile in the absence of repeating. We can see that effects for Grade Retention are very substantial, leading to hefty drops in the ranking. Effects are modest in magnitude for Late Starts.

We estimate effects of retention on an 'Overall Score', which is the mean of the three other scores. This indicates from which percentile on retention is harmful for an 'overall' skill level.

¹We employ several methods to deal with the relatively large range of missing observations in control variables. All produced similarly low variation.

These results are portrayed in table 7.9. GR is significantly harmful from the second percentile on, while LS is positive for the lowest 11 percentiles and significantly so up until the 5th percentile.¹ One should not draw any final conclusions from these results with respect to what the exact percentile is where one should stop with retaining students. Since our ranking is based on test scores, which are influenced by treatment status themselves, distortions are present. Nonetheless, the results provide strong evidence for the overall negative impact of Grade Retention and the strong difference in implication with retention in kindergarten, as well as the presence of strong heterogeneity across the distribution. Moreover, they show that positive effects are present for specific individuals, when retention is executed early.

7.5.3 Heterogeneity by demographics

Here, we analyze group effects by demographic characteristics, being gender and social class. These results, portrayed in table 7.10, again underline the difference in implications between Grade Retention and Late Starts. Grade Retention turns out to be especially bad for boys, which is most prominent for math skills. This is an interesting finding since the average rank for a boy repeater is higher than that of a girl repeater (simply because boys repeat more often). In other words, retention has more adverse effects for boys, even though boy repeaters are on average of higher ability. Effects by social class do not differ between the low and medium category. However, they are (insignificantly) positive for the highest social class, for science and reading scores.² For math, the middle group is most negatively affected by grade retention. The pattern is very different for Late Starts. It is more harmful for girls, in all topics. Late starts are less harmful for the lower social classes, which is again consistent for all test fields. This confirms the belief that it can be better for disadvantaged students to mature longer before they enter formal education (although the overall effect is still negative for this group).

7.5.4 Effects on confidence

We also analyze the effect of the two different forms of repeating on attitude- or confidence-related variables (table 7.11). We construct variables using factor analysis, for several measures of confidence and attitude (such as attitude towards school, towards teacher, confidence in math skills etc.). Each latent factor is based on the answers to at least 4 questions, on which students

¹The very first percentile is an exception here. It is probable that this is due to a distorting influence of extreme outliers here.

²Instrumentation is weak here, and the fraction of retained students in this class is low.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

indicate their agreement with a certain statement, on a 1-4 scale. The latent factors are standardized with a mean of 0 and a standard deviation of 1. Most effects are insignificant. We observe positive and significant effects for late starters in their attitude towards school¹ and their attitude towards school subjects. There is a significantly negative effect for ‘confidence towards math topics’, while the coefficient for ‘confidence in math skills’ is positive. This means that students perceive their skills as slightly better. This is likely because they compare themselves to the rest of the class.² However, if their confidence towards a specific subtopic, such as calculating an area or dealing with quadratics, is questioned, it is reported as lower when one is a Late Starter. The fact that these topics might not have been (extensively) discussed for those who are a grade behind could be a factor here. We observe that the effect on the ‘attitude towards lessons’ is negative and significant, indicating that late starters are sooner bored. Effects for Grade Repeaters are not significant, but the instrumentation is weak there.³ Overall, the effects appear to be neutral, at least for Late Starts. For Grade Retention, standard errors are too high to make strong statement towards its possible psychological effects.

7.5.5 Early versus late retention

In this section, we distinguish between early (ISCED 1) or late (ISCED 2) repeating.⁴ These separate effects are displayed in table 7.12. Results exhibit high standard errors, since the instrumentation with Grade Retention as a whole is already not extremely powerful and weakens further. There is a pattern of higher effects for late retention versus early retention. The effects for early retention are close to those of late starters, with the exception of reading. We have to be careful in interpreting these estimates, since those students who effectively repeat early or late might be very different groups (for example, it could be that more concerned parents let their children repeat more early).⁵ It appears to be far more harmful to repeat late, perhaps because at these ages students are especially in danger of losing their interest in school. Moreover, the sense of failure might be greater since retention is more often based on formal requirements at this stage.

¹Measures if students deem school valuable and helpful in their lives, or find it a waste of time.

²This finding is similar as that in Reynolds (1992).

³We only have this information for PISA 2003, hence instrumentation is weaker than in the merged sample.

⁴Students could indicate in the PISA questionnaire if they were being retained, for both periods separately. The distinction we make is directly based on the answer on this question.

⁵A similar concern could exist in our comparison of Grade Retention and Late Starts, but our heterogeneous analysis of estimating effects for every percentile of the distribution makes them more comparable.

7.6 Robustness

7.6.1 Curriculum bias

There could be an influence from the fact that those who lag a grade behind did not have any instruction on some specific topics that are in the tests. The results suggest that this could be an issue for Late Stars, but not for Grade Retention, because results for reading are of equal size for the latter group while they are clearly lower for the former. Since reading scores are not likely to be affected by the offered curriculum in that specific grade, these results are expected to be more ‘clean’. We analyze this ‘curriculum bias’ further by looking at the effects by difficulty level of the questions, and by looking at results from a problem solving test.

7.6.1.1 Effects by difficulty level

We categorize questions by difficulty level according to the fraction of students that answer them correctly. We use raw scores, by giving one point for every right answer. We find that Grade Repeaters score relatively worse on the easy questions than on the tough questions (table 7.13). For Late Starts, the effects are more negative for tough questions and are insignificant for the easy questions.¹ Results are very consistent between PISA cohorts 2003 and 2009.² Any bias from questions that would not have been discussed yet for repeat students (which are presumably the more advanced or difficult questions) could play a role in the effects for Late Starts, but cannot be a strong determinant of the estimated effects for Grade Retention. Those who are retained in formal grades score bad on basic questions. We do not have information on whether those easy questions are indeed those questions on topics that have been discussed early. It could be that students score well on exactly those questions that have recently been discussed. However, this is not likely given the coefficient for late starters on easy questions. They score relatively well on these test questions, while they have the same disadvantage from the curriculum that real repeaters do.

7.6.1.2 Problem solving

In addition, we look at the performance of students on a specific problem solving test that was taken in PISA 2003. Table 7.14 shows the results from this analysis. The table also lists results

¹They are (insignificantly) positive in some cases, depending on how we treat missing answers and the cohort we look at.

²The table displays results for the merged sample. Estimates by cohort are extremely similar and available on request.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

for mathematics for this same sample, as a comparison. The problem solving test is aimed at students' general way of reasoning. Questions are on how to design the quickest or cheapest way between two subway stations, which gates have to be opened up in order to let the water flow from beginning to end, etc. Any influence from the curriculum in the highest grade is very unlikely here. Coefficients are very similar to those for mathematics.¹ This confirms that the negative effects that we find originate from lacking basic skills. It also indicates that the PISA tests do not suffer from a 'recent curriculum bias' but measure a very general level of skills. These results also show that there is still a negative effect for Late Starters when we obtain results that should be clean of differences in the curriculum.

7.6.2 Exogeneity of month of birth

Recent research challenges the exogeneity of month of birth, arguing that it is correlated with indicators of social class.² We analyze this, using education of the father and number of books at home as indicators of social status. There is a statistically significant (at a 5% level) relationship between month of birth and these indicators. However, it is very weak in magnitude and is only significant if we take the merged sample, which includes 240,000 observations. If we take both PISA studies separate, only the 2003 study exhibits significant estimates. Moreover, the effects for education of the father and number of books at home have opposite signs.³ Similar inconsistencies are visible if we compare countries. All country-level estimates are insignificant. For number of books, 8 country-estimates are negative and 4 are positive. For education of the father, this is exactly opposite. Magnitudes are extremely small. The effect of moving from the lowest social class to the highest,⁴ decreases the predicted month of birth by 0.05 (which effectively is 1.5 days earlier). Hence, we do find a relation between social status and month of birth, but this effect is extremely weak, only significant in very large samples and inconsistent by year, by country and even by indicator. If one can show results that are consistent over time, or by country, possible endogeneity of month of birth should not invalidate results. The fact that our results are extremely similar when comparing the 2003 and 2009 cohorts provides strong support for the robustness of our estimates.

¹They are slightly lower for GR and slightly higher for LS.

²E.g. Buckles and Hungerman (2008)

³For education of the father, the separate PISA studies have opposite signs as well.

⁴Based on number of books, meaning 0-10 books versus more than 500 books.

7.7 Conclusion

In this chapter, we estimate the effect of retention in kindergarten (Late Starts) and retention thereafter (Grade Retention) on achievement and show how these effects differ for different students. We establish that Grade Retention is consistently negative, while Late Starts can have benefits for those near the bottom of the distribution. We use an Instrumental Variable method, where the fraction of students within a certain month of birth is used as an instrument for retention. This value is defined on a country-level and net of one's own value. We combine this with a quantile regression model to identify effects for each percentile in the distribution. Our model corrects for endogeneity and takes into account that the effects of retention are heterogeneous across the distribution; two issues which are not fully addressed in most of the current literature.

Grade Retention leads to especially poor performance on easy questions. Retained students are lacking basic skills. This is very different for Late Starters, where there is no negative effect for the easiest questions. This lends support to our theoretical framework, which specifies that early retention can have beneficial effects because it can better develop fundamental skills. The negative effects of late retention are stronger than those of early retention. Boys and those from lower social classes are hurt more by Grade Retention, while those same groups suffer less from Late Starts than girls and those with higher social status. Additional analysis suggests that our results are not based on the fact that retained students were not taught certain topics yet. Results on a general problem solving test are very similar to effects for mathematics, also suggesting that the PISA tests aim at testing general knowledge and skills rather than questions about very specific and recently-taught sub-topics.

Our effects by percentile are based on grouping students according to their test scores, which are influenced by the treatment themselves. One should therefore not draw any final conclusions towards the exact percentile for which holding back students starts to be harmful. Additionally, our analysis does not take into account changes in peer effects for those who are not being retained, as well as the possible effect of the sole threat of retention on non-repeated students. Future research can focus on possible effects of retention on the untreated. However, our results clearly indicate that the majority of students that are being retained are worse off because of it. There are specific groups that could benefit from it, if retention is executed early enough.

7.8 Figures and tables

Figure 7.1: Frequency of retention by month of birth

Notes: Graph shows the average level of repeating over the month of birth of the corresponding students. Month of birth is counted from the official school cutoff date on, 1 being the first month after the cutoff date (so not necessarily January) and 12 the month just before the cutoff date for the next year. All students who lag (at least) one grade behind their same-age peers are defined as repeaters. No distinction is made between Grade Retention and Late Starts.

Figure 7.2: Frequency of Grade Retention by month of birth

Notes: Graph shows the average level of Grade Retention over the month of birth of the corresponding students. Month of birth is counted from the official school cutoff date on, 1 being the first month after the cutoff date (so not necessarily January) and 12 the month just before the cutoff date for the next year. All students that indicated in the PISA questionnaire that they repeated at least once during their educational career are defined as Grade Repeaters.

Figure 7.3: Frequency of Late Starts by month of birth

Notes: Graph shows the average level of Late Starts over the month of birth of the corresponding students. Month of birth is counted from the official school cutoff date on, 1 being the first month after the cutoff date (so not necessarily January) and 12 the month just before the cutoff date for the next year. All students who lag at least one grade behind their same-age peers and have not answered that they repeated are defined as Late Starters.

Figure 7.4: Effect of Grade Retention and Late Starts on science scores by percentile

Notes: Graph shows how the effect of Grade Retention (in black) and Late Starts (in grey) on science scores develops over the distribution. Effects are estimated by percentile, using a quantile regression model. Coefficients show the treatment effect on the ranking, which is defined on a country-level, for that particular test subject. The solid lines give the estimates, the dotted lines form 95% confidence intervals.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

Figure 7.5: Effect of Grade Retention and Late Starts on math scores by percentile

Notes: Graph shows how the effect of Grade Retention (in black) and Late Starts (in grey) on math scores develops over the distribution. Effects are estimated by percentile, using a quantile regression model. Coefficients show the treatment effect on the ranking, which is defined on a country-level, for that particular test subject. The solid lines give the estimates, the dotted lines form 95% confidence intervals.

Figure 7.6: Effect of Grade Retention and Late Starts on reading scores by percentile

Notes: Graph shows how the effect of Grade Retention (in black) and Late Starts (in grey) on reading scores develops over the distribution. Effects are estimated by percentile, using a quantile regression model. Coefficients show the treatment effect on the ranking, which is defined on a country-level, for that particular test subject. The solid lines give the estimates, the dotted lines form 95% confidence intervals.

Table 7.1: Degree of retention by country

	Degree of Grade Retention		Degree of Late Starts	
	PISA 2003	PISA 2009	PISA 2003	PISA 2009
Belgium	0.255	0.276	0.054	0.055
Canada	0.076	0.049	0.092	0.084
Denmark	0.020	0.038	0.073	0.118
England	-	0.0031	-	0.037
Finland	0.027	0.024	0.023	0.023
France	0.332	0.350	0.019	0.025
Italy	0.135	0.139	0.023	0.033
New Zealand	0.017	0.018	0.051	0.044
Poland	0	0.029	0	0.0041
Portugal	0.227	0.232	0.028	0.014
Spain	0.232	0.237	0.014	0.015
Sweden	0.026	0.0031	0.0063	0.037

Notes: Table lists all countries included in our sample and the corresponding fractions of Repeat Students and Late Starters, for both PISA cohorts (age 15). England is not included in our PISA 2003 estimation, because it did not contain a complete birth cohort. Countries are included when the sample contains those students who were born in the 12 months after the cutoff date. Countries without any retention are dropped as well.

Table 7.2: First stage results: F-values

	Grade Retention	Late Start
PISA 2003	103.87	6604.63
PISA 2009	67.72	9012.92
Merged	169.20	15646.36

Notes: Table reports the strength of using the fraction of students within a certain month of birth who has either repeated or started late as an instrument for the corresponding treatment. This instrument is defined on a country level and net of one's own value. Strength is represented by the F-value of a Wald test on the instrument.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

Table 7.3: Effect of Grade Retention and Late Starts: OLS

	Grade Retention	Grade Retention 2003	Grade Retention 2009	Late Start	Late Start 2003	Late Start 2009
Science Score	-0.944*** (0.0064)	-1.02*** (0.010)	-0.896*** (0.0082)	-0.508*** (0.011)	-0.528*** (0.017)	-0.493*** (0.014)
Math Score	-1.01*** (0.0064)	-1.01*** (0.010)	-1.01*** (0.0081)	-0.508*** (0.011)	-0.544*** (0.016)	-0.480*** (0.014)
Reading Score	-1.01*** (0.0064)	-1.01*** (0.010)	-1.00*** (0.0084)	-0.533*** (0.011)	-0.576*** (0.016)	-0.499*** (0.014)
Overall Score	-1.04*** (0.0064)	-1.06*** (0.010)	1.02*** (0.0082)	-0.541*** (0.011)	-0.576*** (0.016)	-0.515*** (0.014)

Notes: Table reports the effect of Grade Retention (retention after kindergarten) and Late Starts (retention in kindergarten) on test scores, using OLS estimation. Test scores are standardized with a mean of 0 and a standard deviation of 1. ‘Overall Score’ is the average of the scores in the three other subject.

Table 7.4: Effect of Grade Retention and Late Starts: OLS with controls

	Grade Retention	Grade Retention 2003	Grade Retention 2009	Late Start	Late Start 2003	Late Start 2009
Science Score	-0.795*** (0.0066)	-0.867*** (0.011)	-0.667*** (0.0082)	-0.313*** (0.011)	-0.330*** (0.018)	-0.283*** (0.014)
Math Score	-0.888*** (0.0065)	-0.878*** (0.011)	-0.815*** (0.0082)	-0.325*** (0.011)	-0.349*** (0.017)	-0.289*** (0.014)
Reading Score	-0.813*** (0.0064)	-0.805*** (0.010)	-0.734*** (0.0082)	-0.297*** (0.011)	0.313*** (0.016)	-0.262*** (0.014)
Overall Score	-0.872*** (0.0065)	-0.891*** (0.010)	-0.774*** (0.0082)	-0.327*** (0.011)	-0.347*** (0.017)	-0.292*** (0.014)

Notes: Table reports the effect of Grade Retention (retention after kindergarten) and Late Starts (retention in kindergarten) on test scores, using OLS estimation including a range of control variables. Test scores are standardized with a mean of 0 and a standard deviation of 1. Controls included are gender, education of mother and father, country of birth of self and father, language spoken at home and number of books at home. ‘Overall Score’ is the average of the scores in the three other subject.

Table 7.5: Effect of Grade Retention and Late Starts: IV

	Grade Retention	Grade Retention 2003	Grade Retention 2009	Late Start	Late Start 2003	Late Start 2009
Science Score	-0.545** (0.230)	-0.496 (0.306)	-0.577 (0.354)	-0.146*** (0.041)	-0.217*** (0.063)	-0.088 (0.054)
Math Score	-0.432* (0.229)	-0.450 (0.300)	-0.375 (0.359)	-0.204*** (0.041)	-0.224*** (0.062)	-0.188*** (0.055)
Reading Score	-0.619*** (0.228)	-0.600** (0.292)	-0.613* (0.363)	-0.054 (0.041)	-0.075 (0.061)	-0.036 (0.056)
Overall Score	-0.558** (0.228)	-0.541* (0.296)	-0.548 (0.359)	-0.141*** (0.041)	-0.181*** (0.061)	-0.108** (0.055)

Notes: Table reports the effect of Grade Retention (retention after kindergarten) and Late Starts (retention in kindergarten) on test scores, using IV estimation. Test scores are standardized with a mean of 0 and a standard deviation of 1. IV uses the fraction of students within a particular month that has repeated a grade (or started late) as an instrument. This instrument is defined on a country level and net of one's own value. 'Overall Score' is the average of the scores in the three other subject.

Table 7.6: Effect of Grade Retention and Late Starts: IV with controls

	Grade Retention	Grade Retention 2003	Grade Retention 2009	Late Start	Late Start 2003	Late Start 2009
Science Score	-0.528** (0.225)	-0.424 (0.303)	-0.708 (0.344)	-0.138*** (0.040)	-0.216*** (0.062)	-0.070 (0.053)
Math Score	-0.424* (0.224)	-0.376 (0.297)	-0.571* (0.344)	-0.194*** (0.040)	-0.222*** (0.060)	-0.165*** (0.053)
Reading Score	-0.543** (0.222)	-0.519* (0.286)	-0.501 (0.349)	-0.053 (0.040)	-0.079 (0.059)	-0.035 (0.053)
Overall Score	-0.524** (0.223)	-0.461* (0.292)	-0.623 (0.348)	-0.134*** (0.040)	-0.181*** (0.060)	-0.094* (0.053)

Notes: Table reports the effect of Grade Retention (retention after kindergarten) and Late Starts (retention in kindergarten) on test scores, using IV estimation including a range of control variables. Test scores are standardized with a mean of 0 and a standard deviation of 1. See table 7.5 for an explanation on the instrumentation. See table 7.4 for a list of control variables employed. 'Overall Score' is the average of the scores in the three other subject.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

Table 7.7: IV estimates of Grade Retention and Late Starts: per decile

	Grade Retention			Late Starts		
	Science	Math	Reading	Science	Math	Reading
[0-10]	-0.073*** (0.016)	-0.065*** (0.015)	-0.069*** (0.016)	0.0029 (0.0034)	-0.00059 (0.0033)	0.0088* (0.0034)
[10-20]	-0.177*** (0.042)	-0.170*** (0.039)	-0.134*** (0.039)	-0.0058 (0.0093)	-0.022** (0.0093)	0.0067 (0.0092)
[20-30]	-0.200*** (0.062)	-0.144** (0.062)	-0.195*** (0.060)	-0.025* (0.014)	-0.052*** (0.014)	-0.0061 (0.014)
[30-40]	-0.150* (0.091)	-0.096 (0.095)	-0.176** (0.087)	-0.044** (0.017)	-0.090*** (0.019)	-0.016 (0.017)
[40-50]	-0.100 (0.119)	0.0061 (0.124)	-0.155 (0.121)	-0.069*** (0.022)	-0.107*** (0.021)	-0.034 (0.021)
[50-60]	-0.159 (0.154)	0.091 (0.184)	-0.071 (0.171)	-0.087*** (0.025)	-0.117*** (0.025)	-0.044* (0.024)
[60-70]	-0.109 (0.224)	0.011 (0.228)	-0.083 (0.232)	-0.093*** (0.028)	-0.112*** (0.026)	-0.061** (0.026)
[70-80]	0.0062 (0.287)	0.034 (0.319)	-0.254 (0.315)	-0.104*** (0.025)	-0.127*** (0.027)	-0.068** (0.028)
[80-90]	0.159 (0.363)	0.039 (0.441)	-0.290 (0.492)	-0.090*** (0.024)	-0.110*** (0.025)	-0.065*** (0.025)
[90-100]	0.618 (0.486)	-0.150 (0.554)	-0.149 (0.637)	-0.089*** (0.019)	-0.069*** (0.017)	-0.065*** (0.020)

Notes: Table reports the effect of Grade Retention and Late Starts on the ranking of the student in that particular test subject, per decile. We use IV estimation in combination with a quantile regression model. This method sets the values for those that are outside the relevant decile at the lowest or highest threshold of that decile (e.g. at 0.2 and 0.3 for the 3rd decile), while those within the decile keep their own value. Original Effects are rescaled by dividing through the fraction of total Repeaters that is in that specific decile. Ranking is defined on a country level and ranges from 0 to 1. As an example, a coefficient of -0.200 for decile 3 means that the repeaters located there would be located two deciles higher had they not repeated. See table 7.5 for an explanation on the instrumentation.

Table 7.8: IV estimates of Grade Retention and Late Start: per percentile

	Science	Math	Reading		Science	Math	Reading
Grade Retention							
[1]	-0.0063 (0.0046)	-0.0018 (0.0040)	0.00021 (0.0044)	[9]	-0.120*** (0.028)	-0.114*** (0.026)	-0.122*** (0.029)
[2]	-0.031*** (0.010)	-0.024*** (0.0082)	-0.016* (0.0096)	[10]	-0.143*** (0.032)	-0.146*** (0.033)	-0.124*** (0.029)
[3]	-0.049*** (0.013)	-0.038*** (0.012)	-0.044*** (0.013)	[15]	-0.162*** (0.043)	-0.166*** (0.039)	-0.117*** (0.039)
[4]	-0.061*** (0.016)	-0.049*** (0.016)	-0.058*** (0.016)	[20]	-0.228*** (0.056)	-0.225*** (0.061)	-0.189*** (0.064)
[5]	-0.068*** (0.019)	-0.055*** (0.018)	-0.071*** (0.019)	[30]	-0.178** (0.070)	-0.109* (0.063)	-0.220*** (0.078)
[6]	-0.081*** (0.022)	-0.077*** (0.022)	-0.085*** (0.021)	[40]	-0.145 (0.133)	-0.061 (0.128)	-0.167 (0.112)
[7]	-0.092*** (0.024)	-0.091*** (0.023)	-0.102*** (0.023)	[50]	-0.146 (0.157)	0.045 (0.103)	-0.247 (0.195)
[8]	-0.110*** (0.028)	-0.109*** (0.025)	-0.130*** (0.028)	All	-0.121* (0.068)	-0.081 (0.068)	-0.138** (0.067)
Late Starts							
[1]	-0.00049 (0.00075)	0.00029 (0.00066)	-0.000025 (0.0070)	[9]	0.0022 (0.0069)	-0.0077 (0.0069)	0.015** (0.0066)
[2]	0.0020 (0.0018)	0.0022 (0.0018)	0.0024 (0.0017)	[10]	0.0019 (0.0073)	-0.0090 (0.0075)	0.017** (0.0077)
[3]	0.0041* (0.0023)	0.0032 (0.0024)	0.0070** (0.0028)	[15]	-0.0063 (0.0096)	-0.022** (0.0093)	0.0080 (0.0086)
[4]	0.0050 (0.0031)	0.0033 (0.0036)	0.0084*** (0.0031)	[20]	-0.015 (0.012)	-0.030** (0.012)	0.0024 (0.016)
[5]	0.0051 (0.0039)	0.0019 (0.0040)	0.0095** (0.0043)	[30]	-0.028* (0.015)	-0.065*** (0.015)	-0.015 (0.018)
[6]	0.0041 (0.0045)	-0.00062 (0.0048)	0.011** (0.0047)	[40]	-0.074*** (0.025)	-0.112*** (0.022)	-0.020 (0.020)
[7]	0.0040 (0.0055)	-0.0018 (0.0052)	0.016*** (0.0055)	[50]	-0.100*** (0.026)	-0.101*** (0.021)	-0.036 (0.022)
[8]	0.0019 (0.0073)	-0.0040 (0.0062)	0.017*** (0.0064)	All	-0.044*** (0.012)	-0.064*** (0.012)	-0.021* (0.012)

Notes: Table reports the effect of Grade Retention and Late Starts on the ranking of the student in that particular test subject, per percentile. We use IV estimation in combination with a quantile regression model. See table 7.5 for an explanation on the instrumentation, and table 7.7 for an explanation on the quantile regression model.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

Table 7.9: IV estimates of Grade Retention and Late Start: Overall Score (percentiles)

	Retention	Late Start		Retention	Late Start
[1]	-0.0019 (0.0041)	-0.00021 (0.00068)	[13]	-0.161*** (0.035)	-0.0032 (0.0083)
[2]	-0.022** (0.0084)	0.0029* (0.0016)	[14]	-0.183*** (0.040)	-0.0040 (0.0097)
[3]	-0.047*** (0.013)	0.0073*** (0.0024)	[15]	-0.186*** (0.042)	-0.0034 (0.010)
[4]	-0.059*** (0.015)	0.0068** (0.0045)	[16]	-0.169*** (0.039)	-0.0045 (0.0081)
[5]	-0.070*** (0.017)	0.0066* (0.0038)	[17]	-0.159*** (0.039)	-0.010 (0.010)
[6]	-0.082*** (0.019)	0.0068 (0.0045)	[18]	-0.194*** (0.050)	-0.011 (0.012)
[7]	-0.091*** (0.022)	0.0073 (0.0058)	[19]	0.194*** (0.046)	-0.011 (0.011)
[8]	-0.098*** (0.022)	0.0052 (0.0054)	[20]	-0.201*** (0.048)	-0.0092 (0.013)
[9]	-0.108*** (0.025)	0.0036 (0.0066)	[30]	-0.171*** (0.062)	-0.021 (0.013)
[10]	-0.115*** (0.028)	0.00065 (0.0069)	[40]	-0.058 (0.103)	-0.070*** (0.021)
[11]	-0.153*** (0.032)	0.00042 (0.0093)	[50]	-0.114 (0.130)	-0.049*** (0.018)
[12]	-0.165*** (0.034)	-0.0013 (0.0093)	All	-0.121* (0.068)	-0.044*** (0.012)

Notes: Table reports the effect of Grade Retention and Late Start on the ranking of the student on his ‘Overall Score’ (the mean of the scores in Science, Math and Reading), per percentile. We use IV estimation in combination with a quantile regression model. See table 7.5 for an explanation on the instrumentation, and table 7.7 for an explanation on the quantile regression model.

Table 7.10: IV estimates of Grade Retention and Late Starts by demographic group

	Overall	Females	Males	Lowest Social Class	Medium Social Class	Highest Social Class
Grade Retention						
Science Score	0.545** (0.230)	-0.375 (0.354)	-0.700** (0.302)	-0.763** (0.377)	-0.728* (0.381)	0.250 (0.433)
Math Score	-0.432* (0.229)	-0.179 (0.357)	-0.680** (0.296)	-0.250 (0.383)	-0.965** (0.378)	0.442 (0.442)
Reading Score	-0.619*** (0.228)	-0.464 (0.340)	-0.559* (0.300)	-0.865** (0.382)	-0.851** (0.377)	0.210 (0.428)
Overall Score	-0.558** (0.228)	-0.357 (0.351)	-0.678** (0.302)	-0.660* (0.373)	-0.888** (0.373)	0.341 (0.433)
Late Starts						
Science Score	-0.146*** (0.041)	-0.202*** (0.060)	-0.099* (0.056)	-0.015 (0.085)	-0.147** (0.069)	-0.217*** (0.056)
Math Score	-0.204*** (0.041)	-0.286*** (0.060)	-0.129*** (0.055)	-0.118 (0.085)	-0.223*** (0.069)	-0.256*** (0.056)
Reading Score	-0.054 (0.041)	-0.072 (0.058)	-0.048 (0.056)	0.089 (0.087)	-0.060 (0.069)	-0.129** (0.056)
Overall Score	-0.141*** (0.041)	-0.196*** (0.060)	-0.096* (0.056)	-0.015 (0.085)	-0.150** (0.068)	-0.210*** (0.056)

Notes: Table reports the IV effects of Grade Retention and Late Starts on test scores, for different demographic groups. Test scores are standardized with a mean of 0 and a standard deviation of 1. 'Overall Score' is the average of the scores in the three other subject. See table 7.5 for an explanation on the instrumentation.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

Table 7.11: IV estimates of Grade Retention and Late Start on confidence-related variables

	Repeating (no distinction)	Grade Retention	Late Start
Attitude towards school (2009)	0.157*** (0.057)	-0.518 (0.381)	0.162*** (0.049)
Attitude towards school (2003)	0.068 (0.051)	-0.022 (0.358)	0.030 (0.045)
Attitude towards teacher	-0.036 (0.052)	0.157 (0.356)	-0.046 (0.047)
Attitude towards school subjects	0.169*** (0.055)	0.499 (0.381)	0.115** (0.049)
Confidence towards math topics	-0.130*** (0.047)	-0.229 (0.333)	-0.110*** (0.043)
Confidence in math skills	0.067 (0.053)	-0.024 (0.351)	0.067 (0.047)
Attitude towards peers	0.0076 (0.043)	0.414 (0.299)	-0.011 (0.038)
Attitude towards lessons	-0.076* (0.042)	-0.061 (0.288)	-0.080** (0.037)
Attitude towards education	-0.052 (0.065)	-0.318 (0.639)	-0.017 (0.058)
Overall confidence and attitude	0.0074 (0.026)	0.052 (0.180)	-0.00086 (0.023)

Notes: Table shows the effect of Grade Retention and Late Starts on confidence-related variables. All estimates are from the 2003 PISA wave, with the exception of the first entry. The dependent variables are constructed using factor analysis on 4 or more questions. These variables are standardized with a mean of 0 and a standard deviation of 1. The first result column gives effects when we do not distinguish between kindergarten retention and retention thereafter. The later two columns show results when we do. See table 7.5 for an explanation on the instrumentation.

Table 7.12: IV estimates of Early (ISCED 1) versus Late (ISCED 2) Grade Retention

	Early Retention	Late Retention
Science Score	-0.274 (0.238)	-0.720** (0.366)
Math Score	-0.226 (0.237)	-0.496 (0.364)
Reading Score	-0.243 (0.233)	-0.916** (0.365)
Overall Score	-0.260 (0.237)	-0.747** (0.364)

Notes: Table reports the IV effects of early (ISCED 1, but after kindergarten) versus late (ISCED 2) Grade Retention on test scores. 'Overall Score' is the average of the scores in the three other subject. See table 7.5 for an explanation on the instrumentation.

Table 7.13: IV estimates of Grade Retention and Late Start by difficulty level of questions

	Not reached is wrong		Not reached is missing	
	Grade Retention	Late Start	Grade Retention	Late Start
Overall	-0.553** (0.231)	-0.099** (0.041)	0.560** (0.230)	-0.079* (0.041)
[0-0.25]	-0.751*** (0.233)	-0.0077 (0.041)	-0.693*** (0.232)	0.035 (0.042)
[0.25-0.5]	-0.276 (0.236)	-0.099** (0.041)	-0.341 (0.234)	-0.078* (0.041)
[0.5-0.75]	-0.582** (0.232)	-0.089** (0.041)	-0.572** (0.231)	-0.070 (0.041)
[0.75-1]	-0.312 (0.235)	-0.137*** (0.041)	-0.432* (0.233)	-0.134*** (0.041)

Notes: Table shows the effect of Grade Retention and Late Starts on test scores (taking all subjects together), for groups of questions that differ in difficulty level. The order in the first column goes from the 25% easiest to the 25% toughest questions. Difficulty is defined by the percentage of students that had the question correct. We do not use standardized PISA scores, but give one point for each correct question and divide that by the number of total questions. In the table, we make the distinction in treating a question that is not reached by the student as either a wrong answer, or a missing. Scores are standardized with mean 0 and standard deviation 1. See table 7.5 for an explanation on the instrumentation.

7. THE IMPACT OF GRADE RETENTION: AN IV APPROACH TO ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

Table 7.14: IV estimates of Grade Retention and Late Start: problem solving (percentiles)

	Problem Solving		Mathematics	
	Retention	Late Start	Retention	Late Start
[Overall]	-0.632** (0.292)	-0.209*** (0.061)	-0.450 (0.300)	-0.224*** (0.062)
[1]	0.0038 (0.0052)	0.00013 (0.0017)	0.0015 (0.0058)	0.00058 (0.0016)
[2]	-0.0083 (0.010)	0.0017 (0.0022)	-0.012 (0.010)	0.0021 (0.0029)
[3]	-0.013 (0.014)	0.0013 (0.0029)	-0.021 (0.014)	0.0040 (0.0048)
[4]	-0.022 (0.017)	0.00084 (0.0033)	-0.032* (0.019)	0.0047 (0.0061)
[5]	-0.033* (0.019)	0.00033 (0.0046)	-0.032 (0.023)	0.00029 (0.0057)
[6]	-0.052** (0.025)	-0.0040 (0.0054)	-0.043 (0.028)	-0.0038 (0.0068)
[7]	-0.068** (0.029)	-0.0017 (0.055)	-0.060** (0.030)	-0.0041 (0.0090)
[8]	-0.087*** (0.029)	-0.0029 (0.0070)	-0.072** (0.030)	-0.0063 (0.011)
[9]	-0.125*** (0.036)	-0.0035 (0.0082)	-0.110*** (0.030)	-0.012 (0.011)
[10]	-0.172*** (0.043)	-0.0032 (0.0066)	-0.144*** (0.042)	-0.010 (0.011)
[20]	-0.282*** (0.065)	-0.018 (0.016)	-0.321*** (0.082)	-0.014 (0.014)
[40]	-0.332** (0.151)	-0.078* (0.043)	-0.171 (0.155)	-0.119*** (0.040)

Notes: Table shows the effect per percentile of Grade Retention and Late Start on the ranking of the student on his problem solving score (2003 cohort), using IV estimation in combination with a quantile regression model. See table 7.5 for an explanation on the instrumentation, and table 7.7 for an explanation on the quantile regression model. Effects on the ranking for math are added for comparative purposes.

8

Summary and conclusion

8.1 Summary

This thesis uses economic theory to analyze different aspects of the educational system. Thereby, we employ statistical techniques to establish causal links in a research field characterized by inherent endogeneity. In a society where the return to skills is increasing, and where the majority of social challenges are related to a lack of socially productive skills, it is of crucial importance that we get better insights into the process of skill development within the educational system. With this knowledge, we can make the school curriculum more optimal, which will lead to a more educated and more productive work force and can have spillovers towards many other domains of social progress.

We look at four specific aspects of the school curriculum: the age of school entry, the allocation of instruction time, the design of tracking systems, and grade retention. We provide contributions to the discussion of these topics by using economic theory to approach the problems at hand. We address the crucial concept of complementarity in skill development, which shows the importance of measuring impacts at different points in time. We also look at economic implications by executing a cost-benefit analysis and using measures for educational attainment and earnings as outcome variables, in a field that generally relies on test scores as measures of effectiveness. Moreover, we rely on exogenous variation in the curriculum, where studies on these topics generally employ OLS and matching models. The methodology of this thesis also uses specific methods for estimating heterogeneity and discusses a correction method which shows it is possible to use data samples on a grade level in the presence of grade retention and delayed starts. Finally, this thesis designs new theoretical approaches to several issues relating to the

8. SUMMARY AND CONCLUSION

design of educational systems and shows that economic models such as Cunha and Heckman (2007)'s Technology of Skill Formation can be used to approach core questions in education.

We develop a theoretical framework for all of the four aspects of the school curriculum that we analyze. These are summarized in Chapter 2. They all build on Cunha and Heckman (2007)'s Technology of Skill Formation. Their study specifies that skill development is based on incumbent skills, parental characteristics and investments. These concepts are strongly related, both to each other as well as over time. The return on investment will be higher if there is a larger set of incumbent skills. In turn, this implies that future investments will be more productive when early investments have been made in the past. This complementarity will be stronger when interactions between skills are stronger. Initial effects can develop in very different ways and measuring impacts at different points in time is absolutely critical if we want to evaluate the success of any change in policy. In Chapter 2, we specify that we can analyze the effect of differences in school starting ages by variation in month of birth. Differences in month of birth lead to differences in pre-school endowments, while it holds other relevant factors constant. We use the same framework to estimate the impact of the quantity of language instruction time. Instruction time is a form of investment, and its long-term effect depends on its initial impact and the degree of complementarity over time. We also describe that, when we look at changes within the typical school week, we should take into account where this time is taken from and what the productive value is of that activity. The theoretical approaches for the topics of tracking and grade retention have a strong overlap. Both lead to a different investment path, which consists of, primarily, a different curriculum and different peers. We specify that when all students are selected to the track that maximizes their individual outcome, this results in a smooth system of outcomes. When there are discontinuities in outcomes around thresholds between tracks, this indicates that students around that threshold will be better off when switching tracks. For grade retention, our framework recognizes that retention is more beneficial for those near the bottom of the distribution and when executed early.

Chapter 3 contains the empirical analysis of school starting ages. We use within-country differences in month of birth to estimate how learning productivity is related to the age of school entry. This estimates the slope of the Skill Technology Function along a one year range. By executing this for a wide range of countries, we approximate the curvature of this function. Older students learn faster and perform better on achievement tests throughout school. This Birth Month Effect stabilizes relatively early. However, it does not compensate for the foregone time that postponement of school entry causes. Therefore, younger students reach a higher skill

level at a particular exact age. We show that this leads to substantial economic returns, which more than compensate the extra educational costs. The economically optimal starting age lies around the age of 5, and is not very sensitive to changes in parameters.

Chapter 4 analyzes the implications of having more hours of language instruction during primary school. We focus on how initial effects develop over time and whether more language instruction affects skills other than language as well. These two issues depend on the degree of complementarity within a skill over time, and across skills, respectively. We find that complementarity is low in the former case and high in the latter; initial effects of language instruction on performance fade away quickly, while early language instruction has a significant effect on performance in mathematics and environmental studies at the end of primary education. It appears that the average performance of students in language plateaus at a certain point. More language instruction only accomplishes that we reach that plateau at an earlier point in time, but does not improve long-term performance. These results are not driven by a trend in score over time or other changes in school policy.

We examine the impact of differences in the design of the tracking system in Chapters 5 and 6. Chapter 5 examines the exact division of students to the available tracks. We estimate the long-run implications of track assignment for the students that are around the margin of the required ability level for a certain track. Track assignment in the Netherlands is largely based on scores on a selection test at the end of primary education. We exploit the fact that there are discontinuities in treatment probability at certain scores of this test. Using a Fuzzy regression discontinuity design, we find that students who are just short of the required ability level for a certain track would achieve higher educational attainment in that track. This results in an effective return on earnings of around 10% per extra year of education. Our design inhibits relatively high fuzzyness and requires a relatively large bandwidth to ensure first stage power. This makes it important to properly control for the effect of the forcing variable on the outcome variable. We employ a cross-validation procedure that finds the most optimal control function among a wide range of bandwidths and specifications. Sensitivity analysis indicates that our results are not sensitive to changes in this control function.

Chapter 6 estimates the impact of the age of tracking. We analyze how early tracking affects educational attainment and labor market outcomes. Early tracking is defined as taking place before the age of 14 (third year of secondary education in the Netherlands). The control group is tracked at age 14 or thereafter. In the Netherlands, there is within-country variation in this tracking age. This variation is not exogenous, since specific types of students are more likely to

8. SUMMARY AND CONCLUSION

end up in either comprehensive or tracked classes. We use the relative supply of schools in the municipality that track early as an instrument for early tracking. With this IV-approach, we find that early tracking has negative impacts on educational attainment and labor market outcomes. The effect on educational attainment is large and most apparent for tracking at the level of the low and medium track. The impacts of early tracking on labor market outcomes is present for specific subgroups. It can be attributed to a lowering of employment and hours worked, rather than changes in the hourly wage. The impacts of early tracking weaken when we move up in the distribution, but only change sign for a small group of students at the very top of the distribution.

Chapter 7 measures the effect of grade retention on school performance. We develop an IV-approach where we use the fraction of students in a specific month of birth as an instrument for retention. The instrument is defined on a country-level and is taken net of one's own value for retention. We examine how the impact of retention differs by age of execution and thereby make a specific distinction between holding a child back one year in kindergarten (Late Starts) and retention that takes place after the child has entered the first grade of primary school (Grade Retention). We are especially interested in how the effects of retention differ across the distribution. The impact of grade retention on student performance, measured at age 15, is strong and negative for the sample as a whole. However, Late Starts have positive effects for the lowest deciles of the ability distribution. Similarly, retention in early secondary school is more harmful than retention during primary education. Students retained in formal grades perform especially bad on easy questions, while the effect of Late Starts on these same questions is insignificant. This illustrates that early retention can help some individuals because it develops more fundamental skills in a better way.

8.2 Policy implications

We have provided evidence on how certain aspects of the educational system influence skill development. Our results have several implications for policy in the field of education. First of all, results suggest that there are strong benefits from starting early in a child's life with teaching literacy. The optimal age lies around the age of 5, when we take all relevant factors into account. For the majority of developed nations, this corresponds to a lower starting age than currently employed. Such a change would bring large economic payoffs. The predicted economic returns will not be the same for every country, and the optimal starting age can vary across nations as well. However, we indicate that changing parameters does not have strong

implications. Moreover, the method we employ can be applied to any specific case. In order to get the optimal starting age and corresponding economic returns for a specific country, one only needs to estimate and fill in the proper parameters. The fact that we find that earlier starting ages increase inequality, and the fact that Chapter 7 finds positive effects from delayed starts for students of lower ability, indicates that earlier starting ages should go hand in hand with flexible cutoff dates. As such, students who appear to be lagging behind have the opportunity to mature in kindergarten for an additional year.

The results from Chapter 4 indicate that if one wants to improve the long-term performance of students in language, quantitative increases are not a proper tool. However, more language instruction does have value for other topics, especially in early primary school. These spillovers are not present for language instruction at the end of primary school, which questions the overall value of extensive language instruction in these grades. Our results do not preclude that qualitative changes can have sustained impacts on language performance. Machin and McNally (2008) indicate that more focused language instruction or better classroom management could have long-run payoffs on language scores.

The implications of our results on educational tracking are less straightforward. We find that attending the higher track leads to significantly different outcome for students around the required ability threshold. Changing thresholds will raise educational attainment and earnings for this group of students. This provides valuable information to students and parents who face the choice of attending a higher track or not. However, changing thresholds will affect more than the marginal student and will not necessarily improve overall efficiency. Since allowing students at the margin to enter the higher track reduces peer quality in *both* tracks, the effect on the untreated is likely to be negative. On the other hand, crude analysis on peer effects suggests that they do not comprise a large share of the total treatment effect, especially for the highest track. The share of students in the high track is steadily increasing in the Netherlands and our results are indicative that this might have led to increases in educational attainment as well. We also show that postponement of tracking will increase the share of high educated people. Since increasing the number of high educated students should not be a goal in itself, it is important to estimate whether this has labor market implications as well. These effects are apparent for tracking at the level of the medium and high track, but our results for tracking at the level of the two lowest tracks are inconclusive. We require more and longer-term data at this level to draw strong conclusions. Nonetheless, early tracking lowers both educational attainment and earnings

8. SUMMARY AND CONCLUSION

for the sample as a whole, and these outcomes are only positively affected for a small subgroup of students of very high ability.

Finally, positive effects of grade retention are only identified for those at the bottom of the distribution and only when retention is executed at early ages. Retention after kindergarten is harmful for the whole distribution and the negative effects are large for higher percentiles and even more so when retention is executed in secondary education. For several developed nations, the share of students that were ever retained exceeds 25 percent. Our results indicate that this is a far too frequent use of retention. The Technology of Skill Formation tells us that early skill levels are extremely important for the development of skills throughout the life cycle. Retention at later ages can occur when the origin of the students' bad school performance might lie in very fundamental skills. The fact that retained students perform especially bad on easy questions confirms that view. This is in line with evidence that later remediation is difficult when early investment is not undertaken (Cunha et al., 2010).

8.3 Reservations and future research

The findings from this thesis are not universal. The results from chapters 4, 5 and 6 are based on Dutch data. Those from chapters 3 and 7 are based on international data and represent the average effect for a large range of countries. The implications of the timing of school entry, or of letting students repeat a grade, depend on other aspects of the educational system. Examples are the degree of reading emphasis in kindergarten, the method of teaching, whether students simply have to repeat a grade or receive remedial education, et cetera. These aspects can differ by country. Nonetheless, the group of countries we include are fairly homogeneous, and we provide indications that our results are not sensitive to changes in parameters. In addition, the educational system of the Netherlands is representative for many other developed nations. More importantly, we show new approaches in dealing with these issues, which could be applied to (other) country-specific cases.

Although we provide numerous robustness checks and sensitivity analyses in all empirical chapters, some potential issues can still affect estimates. We cannot rule out that the relative age position in the class has an impact on performance that can affect the estimation of our Birth Month Effect. Although we provide evidence against this, future research can elaborate more on the impact of relative age in the class on school performance. Moreover, the heterogeneous analysis of grade retention does not strictly indicate from which percentile on we should stop

with holding back students. The variable we base our ranking on is influenced by the treatment itself, which means this ordering is not perfect. Future studies can contribute when they contain test results or indications of ability before treatment takes place. Additionally, although we employ multiple checks that confirm the validity of the instruments, there exists no formal test for instrument validity in exactly identified systems.

The chapters in this thesis fill some of the gaps in the literature of the economics of education. However, the topics discussed still have many black spots. We give evidence on the effect of language instruction on several skills, but an optimal allocation of instruction time requires that we know the productive value of instruction in all school subjects. Moreover, the payoff of each of these skills towards later-life outcomes is still largely unexplored. Designing the optimal tracking system requires still more research as well. We provide a distinction between tracking before age 14 and at age 14 or thereafter. The fact that the latter outperforms the former does not imply that age 14 is the optimum. Besides looking at larger variation in ages, additional aspects of educational tracking can be examined. The most prominent examples are the optimal number of tracks, and the optimal division of students to tracks from the point of society as a whole. For policy purposes, it is crucial to establish causal links between changing the school curriculum and various outcomes. Since aspects of the school curriculum are endogenous, and since experimental variation is rare in this field of research, it is crucial that we rely on (other) sources of exogenous variation when identifying these effects.

References

- Abadie, A., J. D. Angrist, and G. Imbens (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70(1), 91–117. 3, 211
- Alexander, K. L., D. R. Entwisle, and S. L. Dauber (2002). *On the Success of Failure: A Reassessment of the Effects of Retention in the Primary Grades*. Cambridge, UK: Cambridge University Press. 24
- Almlund, M., A. Duckworth, J. J. Heckman, and T. Kautz (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Amsterdam: Elsevier. 1, 3, 8, 20, 66
- Anderson, M. (2007, July). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool and early training projects. Unpublished manuscript, University of California at Berkeley, Department of Agricultural and Resource Economics. 67
- Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106(4), 979–1014. 1, 62, 117, 209
- Angrist, J. D. and V. Lavy (1999). Using maimonides rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114(2), 553–575. 3, 211
- Angrist, J. D. and V. Lavy (2002). New evidence on classroom computers and pupil learning. *Economic Journal* 112, 735–765. 62
- Ariga, K. and G. Brunello (2007). Does secondary school tracking affect performance? Evidence from IALS. *IZA Discussion Paper* 2643. 106
- Barnett, W. S. (1995, Winter). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children* 5(3). Special volume on Long-Term Outcomes Of Early Childhood Programs. 67
- Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, With Special Reference to Education*. New York: National Bureau of Economic Research. 9, 31
- Bedard, K. and E. Dhuey (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics* 121(4), 1437–1472. 9, 33
- Bell, J. F. and S. Daniels (1990). Are summer-born children disadvantaged? The birthdate effect in education. *Oxford Review of Education* 16(1), 67–80. 33
- Bell, M. L. and C. W. Davidson (1976). Relationships between pupil-on-task-performance and pupil achievement. *The Journal of Educational Research* 69(5), 172–176. 65
- Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy* 75(4, Part 1), 352–365. 9
- Betts, J. R. and J. A. Shkolnik (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review* 19, 1–15. 120
- Black, S. E., P. K. Devereux, and K. G. Salvanes (2008). Too young to leave the nest? the effects of school starting age. *IZA Discussion Paper* 3452. 33
- Blackmore, S. and U. Frith (2005). *The learning brain: lessons for education*. Blackwell Publishing. 9, 34
- Blau, D. and J. Currie (2006). Preschool, daycare, and afterschool care: Who's minding the kids? In E. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Volume 2 of *Handbooks in Economics*, Chapter 20, pp. 1163–1278. Amsterdam: North-Holland. 67
- Borghans, L. and R. Diris (2010a). Allocating instruction time: How language instruction can affect multiple skills. Discussion Paper. 7, 61
- Borghans, L. and R. Diris (2010b). An economic analysis of the optimal school starting age. Discussion Paper. 7, 31
- Borghans, L., R. Diris, W. Smits, and J. De Vries (2011). The labor market effects of secondary school track assignment: A regression discontinuity approach. Discussion Paper. 7, 103
- Borghans, L., R. Diris, W. Smits, and J. De Vries (2012). The impacts of early tracking on later-life outcomes: An instrumental variable approach. Discussion Paper. 7, 143
- Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90(430), 443–450. 111
- Brinch, C. N. and T. A. Galloway (2012). Schooling in adolescence raises IQ scores. *Proceedings of the National Academy of Sciences* 109(2), 425–430. 67
- Bruer, J. T. (1999). *The myth of the first three years: A New Understanding of Early Brain Development and Lifelong Learning*. New York: The Free Press. 34
- Brunello, G. and D. Checchi (2007). Does school tracking affect equality of opportunity? new international evidence. *Economic Policy* (52), 781–861. 104, 107
- Buckles, K. and D. M. Hungerman (2008). Season of birth and later outcomes: Old questions, new answers. *NBER Working paper* (14573). 49, 182
- Card, D. (1993). Using geographic variation in college proximity to estimate the returns to schooling. *NBER Working Paper* 4483. 62
- Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 5, pp. 1801–1863. New York: North-Holland. 62, 117
- Ceci, S. J. (2001). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology* 27(5), 703–722. 67
- Centraal Bureau voor de Statistiek (2010). *Jaarboek onderwijs in cijfers*. Jaarrapport. 82
- Coffield, W. H. and P. Blommers (1956). Effects of non-promotion on educational achievement in the elementary school. *The Journal of Educational Psychology* 47, 235–250. 174
- Commissie Evaluatie Basisonderwijs (1994). *Zicht op kwaliteit: Evaluatie van het basisonderwijs*. Eindrapport, De Meern: Inspectie van het basisonderwijs. 64
- Cooley, W. W. and G. Leinhardt (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis*. 65

REFERENCES

- Costa, P. T. and R. R. McCrae (1994). Set like plaster? evidence for the stability of adult personality. In T. F. Heatherton and J. L. Weinberger (Eds.), *Can Personality Change?*, pp. 21–40. Washington, DC: American Psychological Association. 66
- Costa, P. T., R. R. McCrae, and I. C. Siegler (1999). Continuity and change over the adult life cycle: Personality and personality disorders. In C. R. Cloninger (Ed.), *Personality and Psychopathology*. Washington, D. C.: American Psychiatric Publishing, Inc. 66
- Cunha, F. and J. J. Heckman (2007). The technology of skill formation. *American Economic Review* 97(2), 31–47. 4, 7, 8, 9, 17, 23, 26, 33, 39, 198
- Cunha, F. and J. J. Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources* 43(4), 738–782. 12
- Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* 78(3), 883–931. 12, 26, 76, 80, 202, 215
- Currie, J. (2001). Early childhood education programs. *Journal of Economic Perspectives* 15(2), 213–238. 67
- Daniels, A. and E. Haller (1981). Exposure to instruction, surplus time, and student achievement: A local replication of the Harnischfeger and Wiley research. *Educational Administration Quarterly* 17, 48–68. 65
- Devlin, B., S. Fienberg, D. Resnick, and K. Roeder (1997). *Intelligence, genes and success: Scientists respond to The Bell Curve*. New York: Springer. 20
- Diris, R. (2012). The impact of grade retention: An iv approach to estimating heterogeneous treatment effects. Discussion Paper. 7, 169
- Eden, G., D. Flowers, L. Gareau, P. Turkeltaub, and T. Zeffiro (2003). Development of neural mechanisms for reading. *Nature Neuroscience* 81(6(7)), 767–773. 35
- Elder, T. E. and D. H. Lubotsky (2009). Kindergarten entrance age and children's achievement impacts of state policies, family background, and peers. *Journal of Human Resources* 44, 641–683. 47
- Epplé, D., E. N. and R. Romano (2002). Ability tracking, school competition, and the distribution of educational benefits. *Journal of Public Economics* (83), 1–48. 108
- Figlio, D. N. and M. E. Page (2002). School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics* (51), 497–514. 104, 108
- Flynn, J. R. (1987). Massive iq gains in 14 nations: What iq tests really measure. *Psychological Bulletin* 101, 171–191. 85
- Fredriksson, P. and B. Ockert (2005). Is early learning really more productive? the effect of school starting age on school and labor market performance. *Working paper* (1659). 33, 47
- Galindo-Rueda, F. and A. Vignoles (2005). The heterogeneous effect of selection in secondary schools: Understanding the changing role of ability. CEP discussion paper, Centre for Economic Performance, London School of Economics. 107, 144, 155
- Goldberger, A. S. (1972). Selection bias in evaluating treatment effects: Some formal illustrations. *Discussion Paper*, 129–172. 113
- Gomes-Neto, J. B. and E. A. Hanushek (1994). Causes and consequences of grade repetition: Evidence from brazil. *Economic Development and Cultural Change* 43, 117–148. 174
- Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69, 201–209. 110, 119
- Hanushek, E. A. and L. Woessmann (2006). Does educational tracking affect performance and inequality? differences-in-differences evidence across countries. *The Economic Journal* 116, 63–76. 104, 106, 144, 150
- Harmon, C., H. Oosterbeek, and I. Walker (2003). The returns to education: Microeconomics. *Journal of Economic Surveys* 17, 115–141. 62, 117
- Harnischfeger, A. and D. Wiley (1976). The teaching-learning process in elementary schools: A synoptic view. *Curriculum Inquiry* 6(1), 5–43. 64
- Heckman, J. J. (2000, March). Policies to foster human capital. *Research in Economics* 54(1), 3–56. 1, 3, 8
- Heckman, J. J. and D. V. Masterov (2006, August). The productivity argument for investing in young children. Discussion paper, Early Childhood Research Collaborative, University of Minnesota, Minneapolis, MN. 12
- Heckman, J. J., S. H. Moon, and R. Pinto (2010). The effects of early intervention on abilities and social outcomes: Evidences from the Carolina Abecedarian study. Unpublished manuscript, University of Chicago. 67
- Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics* 1(1), 1–46. First draft, September, 2006. 67, 74
- Heckman, J. J., J. Stixrud, and S. Urzua (2006, July). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* 24(3), 411–482. 1
- Herrnstein, R. and C. Murray (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press. 20
- Holmes, T. C. and K. M. Matthews (1984). The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research* 54, 297–315. 171, 173
- Huttenlocher, P. and A. Dabholkar (1997). Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology* 387(2), 167–178. 35
- Imbens, G. and K. Kalyanaraman (2008). Optimal bandwidth choice for the regression discontinuity estimator. *NBER Working Paper* (14726). 110, 114
- Imbens, G. and T. Lemieux (2007). Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142, 615–635. 114
- Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica* 62(2), 467–475. 153, 177
- Inspectie van het Onderwijs (1997). Aansluiting voortgezet onderwijs op het basisonderwijs. Technical report, Utrecht: Inspectie van het Onderwijs. 105
- Jackson, G. B. (1975). The research evidence on the effects of grade retention. *Review of Educational Research* 45(4), 613–635. 24, 171, 173, 175
- Jacobson, K. (1980). The relationship of individual student time allocation to reading and mathematics achievement. Technical report, Research and Development Center for Individualized Schooling, Wisconsin University, Madison, Wisconsin. 65
- Jaeggi, S., M. Buschkeuhl, J. Jonides, and W. Perrig (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, USA* 105, 6829–6833. 66, 75
- Jakubowski, M. (2010). *Institutional Tracking and Achievement Growth: Exploring Difference-in-Differences Approach to PIRLS, TIMSS, and PISA Data*. 106, 144

- Karoly, L. A., M. R. Kilburn, and J. S. Cannon (2005). *Early Childhood Interventions: Proven Results, Future Promise*. Santa Monica, CA: RAND. 67
- Karweit, N. L. (1991). Repeating a grade: Time to grow or denial of opportunity? report no. 16. *Center for Research on Effective Schooling for Disadvantaged Students, Baltimore, MD.*. 171, 172, 173
- Kiesling, H. (1984). Assignment practices and the relationship of instructional time to the reading performance of elementary school children. *Economics of Education Review* 3, 341–350. 65, 80
- Koenker, R. and J. Bassett, Gilbert (1978, January). Regression quantiles. *Econometrica* 46(1), 33–50. 77, 176
- Lee, D. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature* 48, 281–355. 114
- Leuven, E., M. Lindahl, H. Oosterbeek, and D. Webbink (2007). The effect of extra funding for disadvantaged pupils on achievement. *The Review of Economics and Statistics* 89, 721–736. 3, 211
- Leuven, E. and H. Oosterbeek (2004, April). Evaluating the effect of tax deductions on training. *Journal of Labor Economics* 2, 461–488. 3, 211
- Lochner, L. (2011). Non-production benefits of education: Crime, health, and good citizenship. Working Paper 16722, National Bureau of Economic Research. 1, 209
- Lochner, L. J. and E. Moretti (2004, March). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review* 94(1), 155–189. 1, 209
- Machin, S. and S. McNally (2008). The literacy hour. *Journal of Public Economics* 92(5-6), 1441–1462. 62, 65, 75, 201, 214
- Manning, A. and J.-S. Pischke (2006). Comprehensive versus selective schooling in England and Wales: What do we know? CEP discussion paper, Centre for Economic Performance, London School of Economics. 107, 155
- Mantzicopoulos, P. and D. Morrison (1984). Kindergarten retention: Academic and behavioral outcomes through the end of second grade. *American Educational Research Journal* 29, 182–198. 26, 171
- McArdle, J., E. Ferrer-Caja, and F. Hamagami (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology* 38, 115–142. 66
- McCoy, A. R. and A. J. Reynolds (1999). Grade retention and school performance: An extended investigation. *Journal of School Psychology* 37, 237–298. 171
- McCrae, R. R. and P. T. Costa (1990). *Personality in Adulthood*. New York: Guilford Press. 66
- McCrae, R. R. and P. T. Costa (1994). The stability of personality: Observation and evaluations. *Current Directions in Psychological Science* 3(6), 173–175. 66
- McCrae, R. R. and P. T. Costa (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives*, pp. 51–87. New York, NY: Guilford Press. 66
- McCrae, R. R. and P. T. Costa (2003). *Personality in adulthood: a five-factor theory perspective* (2nd ed.). New York: Guilford Press. 66
- McEwan, P. J. and J. S. Shapiro (2008). The benefits of delayed primary school enrollment: Discontinuity estimates using exact birth dates. *Journal of Human Resources* 43, 1–29. 33
- Meadows, S. (1993). *The Child as Thinker: The Development and Acquisition of Cognition in Childhood*. Routledge. 9
- Mincer, J. (1974). *Schooling, Experience, and Earnings*. Columbia University Press, New York. 62
- Ministerie van Onderwijs Cultuur en Wetenschap (2011). Kerncijfers 2006-2010. Technical report. 146
- Olds, D. L., H. Kitzman, R. Cole, J. Robinson, K. Sidora, D. W. Luckey, C. R. Henderson, C. Hanks, J. Bondy, and J. Holmberg (2004, December). Effects of nurse home-visiting on maternal life course and child development: Age 6 follow-up results of a randomized trial. *Pediatrics* 114(6), 1550–1559. 67
- Pagani, L., R. E. Tremblay, F. Vitaro, B. Boulerice, and P. McDuff (2001a). Effects of grade retention on academic performance and behavioral development. *Development and Psychopathology* 13(2), 297–315. 24
- Pagani, L., R. E. Tremblay, F. Vitaro, B. Boulerice, and P. McDuff (2001b). Effects of grade retention on academic performance and behavioral development. *Development and Psychopathology* 13, 297–315. 171
- Pearce, M., I. Deary, A. Young, and L. Parker (2005). Growth in early life and childhood IQ at age 11 years: the Newcastle Thousand Families Study. *Int J Epidemiol* 34, 673–677. 66
- Pekkarinen, T., R. Uusitalo, and S. Kerr (2009). Schooltracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics* 93, 965–973. 107, 144
- Peterson, S. E., J. S. DeGracie, and C. R. Ayabe (1987). A longitudinal study of the effects of retention/promotion on academic achievement. *American Educational Research Journal* 24, 107–118. 171, 174
- Powell, M. (1978). Educational implications of current research on teaching. *Journal of Educational Measurement*, 15–21. 64
- Psacharopoulos, G. (1994). Returns to investment in education: a global update. *World Development*, 1325–1343. 62
- Puhani, P. and A. Weber (2005). Does the early bird catch the worm? Instrumental variable estimates of educational effects of age of school entry in Germany. *IZA Discussion Paper* (1827). 33
- Reynolds, A. J. (1992). Grade retention and school adjustment: An explanatory analysis. *Educational Evaluation and Policy Analysis* 14, 171, 180
- Robertson, D. and J. Symons (2003). Do peer groups matter? Peer group versus schooling effects on academic attainment. *Economica* 70, 31–53. 120
- Sammons, P. (1999). *School Effectiveness: Coming of Age in the Twenty First Century*. Swets and Zeitlinger, Lisse, The Netherlands. 62
- Scheerens, J. (2000). *Improving School Effectiveness*. IIEP, Paris. 62
- Schneepf, S. V. (2003). Inequalities in secondary school attendance in Germany. S3RI Applications Working Paper A03/16. 107
- Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics* 34(2), 238–266. 113
- Shepard, L. and M. Smith (1986). Synthesis of research of school readiness and kindergarten retention. *Educational Leadership* 44(3), 26, 31, 34
- Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research* 60(3), 108
- Smith, M. L. and L. A. Shepard (1988). Kindergarten readiness and retention: A qualitative study of teachers' beliefs and practices. *American Educational Research Journal* 25, 171

REFERENCES

- Soar, R. (1978). Setting variables, classroom interaction, and multiple pupil outcomes. Paper presented at the annual meeting of the American Educational Research Association, Toronto. 65, 80
- Staiger, D. and J. H. Stock (1997, May). Instrumental variables regression with weak instruments. *Econometrica* 65(3), 557–586. 73, 149
- Stallings, J. (1980). Allocated academic learning time revisited, or beyond time on task. *Educational Researcher* 9(11), 11–16. 64
- Stallings, J., M. Almy, L. Resnick, and G. Leinhardt (1975). Implementation and child effects of teaching practices in follow through classrooms. *Monographs of the Society for Research in Child Development* 40(7/8), 1–133. 65
- Stallings, J., M. Needels, and N. Stayrook (1979). The teaching of basic reading skills in secondary schools, phase ii and phase iii. Technical report, National Institute of Education, Menlo Park, CA. 64
- Strøm, B. (2004). Student achievement and birthday effects. *Mimeo. Department of Economics Norwegian University of Science and Technology*. 9, 32
- Summers, A. A. and B. L. Wolfe (1977). Do schools make a difference? *American Economic Review* 67(4), 639–652. 120
- Sund, K. (2009). Estimating peer effects in swedish high school using school, teacher, and student fixed effects. *Economics of Education Review* 28, 329–336. 120
- Tomchin, E. M. and J. C. Impara (1992). Unraveling teachers' beliefs about grade retention. *Educational Research Journal* 29, 171, 172
- Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review* 43(4), 1249–1287. 110, 119, 139
- Van Elk, R., M. Van der Steeg, and D. Webbink (2011). Does the timing of tracking affect higher education completion? *Economics of Education Review* 30, 1009–1021. 144, 147, 151, 155
- Waldinger, F. (2007). Does ability tracking exacerbate the role of family background for students' test scores? Unpublished Manuscript, London: London School of Economics. 107
- Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis* 20(3), 137–156. 65
- Westat (2010). Head start impact study: Final report. Technical Report Contract no. 282-00-0022, U.S. Department of Health and Human Services. 74
- Winston, G. and D. Zimmermann (2003). Peer effects in higher education. *Working Paper NBER* (9501). 33
- Yoshikawa, H. (1995). Long-term effects of early childhood programs on social outcomes and delinquency. *The Future of Children* 5, 51–75. 67

Nederlandse samenvatting

9.1 De economie van het school curriculum

Er is een toenemend besef dat het ontwikkelen van menselijk kapitaal een effectief middel is voor de aanpak van verscheidene maatschappelijke problemen. Zowel cognitieve als non-cognitieve vaardigheden zijn belangrijke determinanten van latere uitkomsten zoals het behaalde opleidingsniveau, inkomen, gezondheid, criminaliteit en algehele welzijnsniveaus. Bovendien neemt het rendement op deze vaardigheden toe. Het reële loon voor hoogopgeleiden is de laatste 25 jaar sterk toegenomen. Aan de andere hand stagneert of daalt dit reële loon voor lager opgeleiden. Het verhogen van het niveau van productieve vaardigheden zal daarmee ook de ongelijkheid in de maatschappij kunnen aanpakken, aangezien hiermee de productiviteit van lager opgeleiden verhoogd wordt, terwijl hoger opgeleiden minder schaars worden.

Het ontwikkelen van dergelijke vaardigheden gebeurt in een divers aantal situaties en omgevingen. Academische instituten, zoals scholen en universiteiten, zijn niet de enige plekken waar belangrijke vaardigheden worden geleerd. Deze ontwikkelen zich ook in de familieomgeving en onder klasgenoten en vrienden. Dit ontwikkelingsproces stopt niet wanneer formeel onderwijs eindigt. Ook tijdens het werk gaat deze door. Niettemin vindt een substantieel deel van de ontwikkeling plaats binnen de school. Het belang van onderwijs is zichtbaar in vele meetbare uitkomsten. Eerder onderzoek heeft causale verbanden aangetoond van onderwijs naar lonen, gezondheid en criminaliteit.¹ Bovendien ontwikkelen kinderen ook non-cognitieve vaardigheden tijdens het onderwijsproces. Echter, de exacte processen binnen het onderwijssysteem die leiden tot een effectieve ontwikkeling van dergelijke vaardigheden zijn nog grotendeels

¹Zie o.a. Angrist and Krueger (1991); Lochner (2011); Lochner and Moretti (2004).

onduidelijk. Wanneer we hierin een beter inzicht krijgen, kunnen we daarmee het onderwijssysteem inrichten op een manier die productiever is in het ontwikkelen van cruciale vaardigheden. Daarmee kan het rendement op onderwijs verhoogd worden, en kunnen verscheidene maatschappelijke problemen effectiever aangepakt worden.

Deze dissertatie behandelt hoe verschillende aspecten van het onderwijssysteem van invloed zijn op de ontwikkeling van vaardigheden en uitkomsten op de lange termijn. Het curriculum dat leerlingen doorlopen, kan op verscheidene manieren ontworpen worden. Deze verschillen kunnen liggen in de toegepaste startleeftijd voor formeel onderwijs, of de leeftijd waarop leerlingen in aparte schoolniveaus worden opgedeeld. Het exacte ontwerp van een schoolsysteem heeft grote implicaties voor de ontwikkeling van vaardigheden van leerlingen die dit systeem ondergaan. Dit zal zich vertalen in de prestaties op toetsen, maar ook in latere uitkomsten zoals het behaalde opleidingsniveau of lonen. Wanneer we het curriculum op een manier ontwerpen die de effectiviteit van leren verbetert, zullen deze latere uitkomsten ook verbeteren. Echter, we weten nog steeds weinig over hoe we het leerproces kunnen optimaliseren. Het ontwerp van het schoolsysteem is merendeels gebaseerd op traditie, gezond verstand of ontwikkelingstheorieën en zelden gestaafd op hard bewijs. Er is geen duidelijk beeld aangaande welke leeftijd het beste is voor het leren van welke vaardigheden, welke volgorde van lesgeven het meest efficiënt is, of met wat voor soort klasgenoten studenten het beste leren.

Hoewel economisch onderzoek in dit veld zich uitbreidt, zijn deze vraagstukken historisch gezien het domein geweest van onderwijskundigen. Economische theorie kan de discussie van deze vragen ernstig verrijken. Bijvoorbeeld, discussie over de optimale startleeftijd heeft grotendeels het economische argument genegeerd dat uitstel van leren verloren tijd impliceert. We laten zien dat het opnemen van dit argument de conclusies structureel verandert. Daarnaast herbergt het complexe en dynamische karakter van het leerproces verscheidene economische concepten.

De ontwikkeling van vaardigheden kenmerkt zich door sterke interacties, ook wel complementariteit genoemd. Vaardigheden die in het verleden zijn ontwikkeld, bepalen mede hoe snel iemand nieuwe vaardigheden leert. Deze interacties kunnen vele dimensies hebben. Het is bijvoorbeeld logisch dat het leren van het alfabet een voorwaarde is voor het leren van lezen en schrijven. Maar goed ontwikkelde leesvaardigheid kan ook de productiviteit van leren in vakken zoals rekenen en aardrijkskunde verhogen. Het dynamische karakter van het leerproces benadrukt het belang om effecten op verschillende tijdstippen te meten. Oorspronkelijke effecten kunnen zich op totaal verschillende manieren ontwikkelen, afhankelijk van de kracht

van dergelijke interacties. Dit impliceert ook dat corrigerende investeringen op latere leeftijd vaak geen effect sorteren, aangezien ze op een onderontwikkelde set van vaardigheden bouwen. Investerings hebben het grootste effect op vroege leeftijden, omdat hun rendement bijdraagt voor een groter deel van de levenscyclus.

Buiten het dynamische karakter van het leerproces is het ook belangrijk om de aanwezigheid van endogeniteit te erkennen. Het is naïef om cross-sectionele variatie tussen landen te gebruiken wanneer we het effect van verschillen in de opbouw van het onderwijsproces schatten. De opzet van dit proces is niet willekeurig bepaald. Bovendien zijn verschillen in het curriculum tussen leerlingen binnen een land ook zelden exogeen. Scholen met slechter presterende leerlingen besteden bijvoorbeeld meer tijd aan kernvakken zoals taal en rekenen. Wanneer we simpelweg zouden kijken naar de correlatie tussen lestijd in taal (of rekenen) en prestaties, zullen deze resultaten verstoord zijn, aangezien minder presterende studenten geassocieerd zijn met meer lestijd in taal. Empirische schattingen moeten hiervoor controleren. Eerder onderzoek in onderwijs leunt doorgaans op cross-sectionele variatie in de verklarende variabele. Endogeniteit wordt geadresseerd door te matchen op onderwijsniveau van de ouders, sociaaleconomische status en andere observeerbare variabelen. Maar deze aanpak corrigeert niet voor verschillen in niet-observeerbare variabelen. Onderzoek laat zien dat het corrigeren voor deze endogeniteit cruciaal is wanneer we het effect van onderwijs of arbeidstraining schatten.¹ Voor beleidsanalyse is het cruciaal dat we causale verbanden leggen tussen veranderingen in het onderwijssysteem en uitkomstvariabelen. We ontwikkelen verschillende methodes om te corrigeren voor endogeniteit. Deze methodes gebruiken instrumentele variabelen of regression discontinuity designs. De instrumenten gebruiken exogene variatie die veroorzaakt wordt door de geboortemaand, beleidsveranderingen, of variatie in schoolkenmerken per gemeente. Onze resultaten laten zien dat corrigeren voor endogeniteit leidt tot verschillende schattingen en vaak ook verschillende kwalitatieve conclusies. We kijken in deze thesis ook naar de heterogeniteit van treatment effects. We gebruiken een quantile regression model om effecten over de hele distributie te schatten en kijken ook naar hoe effecten verschillen voor verschillende demografische groepen.

Deze thesis onderzoekt vier specifieke aspecten van het school curriculum: de startleeftijd van formeel onderwijs, de allocatie van lestijd over de verschillende vakken, de implementatie van 'tracking', en het effect van retentie. We ontwikkelen een theoretische aanpak voor elk van deze aspecten, en koppelen deze aan een empirische aanpak. Alle theoretische discussie bouwt

¹Zie, bijvoorbeeld, Abadie et al. (2002); Angrist and Pischke (1999); Leuven et al. (2007); Leuven and Oosterbeek (2004).

verder op The Technology of Skill Formation van Cunha en Heckman (2007). Deze discussies zijn gebundeld in hoofdstuk 2. De vijf opvolgende hoofdstukken focussen op de empirische analyse.

In hoofdstuk 2 specificeren we dat we het effect van verschillen in de startleeftijd kunnen analyseren door middel van variatie in de geboortemaand. Dergelijke variatie leidt tot verschillen in opgebouwde ‘endowments’ bij aanvang van formeel onderwijs, terwijl andere relevante factoren constant worden gehouden. We gebruiken hetzelfde raamwerk om het effect van lestijd in taal te schatten. Lestijd is een vorm van investering, en het langetermijneffect van een dergelijke investering hangt af van het initiële effect en de mate van complementariteit over tijd. We specificeren dat, wanneer men kijkt naar verschillen in lestijd binnen de standaard lesweek, we ook moeten kijken waar deze extra tijd vandaan is gehaald en wat de waarde van deze gereduceerde lessen is. De theoretische aanpak voor de onderwerpen tracking en retentie hebben een sterke overlap. Beiden leiden tot een verschillend investeringspad, dat vooral bestaat uit een ander curriculum en andere klasgenoten. Daarnaast ontwikkelen we een apart raamwerk voor toewijzing aan tracks. Wanneer alle studenten toegewezen zijn aan de track die hun uitkomsten maximaliseert moet dit resulteren in een vloeiend systeem van uitkomsten. Wanneer er discontinuïteiten zijn in dat systeem, dan geeft dit aan dat sommige studenten beter af zijn in een andere track. Wat betreft retentie geeft onze aanpak aan dat deze voordeliger moet zijn voor de onderkant van de distributie, en des te vroeger retentie wordt uitgevoerd.

Hoofdstuk 3 analyseert de optimale timing van beginnen met leren lezen en schrijven. We ontwikkelen een model waarin de leerproductiviteit toeneemt met leeftijd, maar waar de economische kosten toenemen naarmate het leren verder wordt uitgesteld. De Skill Technology Function specificeert hoe leerproductiviteit samenhangt met leeftijd. We schatten deze functie door een combinatie van schattingen binnen landen (gebruik makend van variatie in de geboortemaand van leerlingen) en tussen landen. Hiermee kunnen we de verschillen in het leerproces afleiden van elke verandering in de startleeftijd. We vinden dat oudere studenten het beter doen op toetsen. Dit effect van geboortemaand op prestatie is persistent en stabiliseert zich relatief vroeg. Echter, deze voorsprong compenseert niet genoeg voor het feit dat deze studenten ook ouder zijn wanneer ze deze testen maken, en wanneer ze hun opleiding voltooien. Leerlingen die eerder beginnen, bereiken een bepaald niveau op een eerdere leeftijd, wat significante economische opbrengsten met zich meebrengt. Deze opbrengsten compenseren ruim voor extra gemaakte onderwijskosten. Verdere analyse laat zien dat deze effecten niet worden gedreven door ‘peer effects’, ‘tracking’ of mogelijke endogeniteit van de geboortemaand.

In Hoofdstuk 4 kijken we naar de korte- en langetermijneffecten van veranderingen in de hoeveelheid lestijd besteed aan taal in het primair onderwijs. We analyseren hoe initiële effecten op toetsscores zich ontwikkelen door de tijd heen en of lestijd in taal mogelijk ‘spillover’ effecten heeft naar andere vaardigheden dan taal. Hiervoor gebruiken we exogene variatie in lestijd in taal door de tijd heen, ontstaan door een beleidsverandering in Nederland. We vinden dat meer tijd in taal leidt tot kortetermijnopbrengsten in taalprestaties, maar deze verdwijnen effecten verdwijnen snel. Het effect van tijd in taal op rekenen en zaakvakken is echter wel persistent. Het lijkt erop dat er een bepaald verzadigingsniveau van taalvaardigheid is. Meer lestijd in taal op vroege leeftijd bereikt alleen dat dat niveau eerder wordt bereikt. Deze resultaten worden niet gedreven door verschillen over tijd op het niveau van de school of op individueel niveau.

De Hoofdstukken 5 en 6 gaan over tracking in het onderwijs, wat refereert naar het opsplitsen van leerlingen naar verschillende schoolniveaus. We bekijken twee aparte aspecten van tracking. In hoofdstuk 5 kijken we naar het effect van toewijzing naar een bepaalde track voor leerlingen die rond het gevraagde leerniveau voor die track zitten. Hierbij gebruiken we een discontinuïteit in de kans op de toewijzing naar track rond bepaalde scores van de Nederlandse Cito-toets. Gebruik makend van een fuzzy regression discontinuity design vinden we dat leerlingen een hoger opleidingsniveau behalen wanneer ze toegewezen worden aan de hoogste track. Dit vertaalt zich ook in hogere lonen op latere leeftijd. Deze nemen toe met zo’n 10% per extra jaar onderwijs.

Hoofdstuk 6 kijkt naar het effect van de leeftijd waarop studenten ‘getracked’ worden, en hoe dit afhangt van het niveau van de student. Vroege selectie is gedefinieerd als plaatvindend voor het derde jaar van secundair onderwijs in Nederland. Er is variatie in deze leeftijd binnen Nederland, maar die is niet exogeen. We gebruiken het relatieve aanbod van scholen die vroeg selecteren binnen de gemeente als een instrument voor vroege selectie. We vinden dat vroege selectie leidt tot een lager aantal diploma’s voor hoger onderwijs, wat zich ook vertaalt in lagere lonen voor sommige groepen van studenten. Dit effect is zwakker voor betere studenten, maar wordt alleen positief voor een kleine groep van zeer sterk presterende leerlingen.

Hoofdstuk 7 meet het effect van retentie op schoolprestaties. We gebruiken een IV model om dit te schatten, waarbij we de fractie van studenten met een bepaalde geboortemaand die zijn blijven zitten, gebruiken als instrument voor retentie. We combineren dit met een quantile regression model dat effecten schat voor elk percentiel van de distributie. We maken daarbij een specifiek onderscheid tussen een jaar extra blijven zitten op de kleuterschool, en retentie die daarna plaatsvindt. Retentie heeft een algemeen negatief effect op schoolprestaties, maar kan positief zijn voor kleine groepen studenten onderaan de distributie, wanneer deze plaatsvindt op

de kleuterschool. We vinden dat studenten die zijn blijven zitten na de kleuterschool relatief het slechtst presteren op makkelijke vragen, wat aangeeft dat ze basisvaardigheden missen. Voor kinderen die een jaar extra spenderen op de kleuterschool is er geen negatief effect op deze makkelijke vragen. Retentie in het begin van de middelbare school blijkt ook schadelijker dan retentie in primair onderwijs.

9.2 Beleidsimplicaties

Onze resultaten hebben verscheidene beleidsimplicatie. Ten eerste suggereren ze dat er sterke voordelen zitten aan het vervroegen van de startleeftijd voor formeel onderwijs. De optimale leeftijd voor het beginnen met lezen en schrijven ligt rond de 5 jaar, wat vroeger is dan wat de meeste westerse landen op dit moment aanhouden. De precieze opbrengsten en optima kunnen verschillen per land, maar zijn niet erg gevoelig voor veranderingen in parameters. Bovendien kan onze methode ook op individuele landen toegepast worden. Het feit dat we vinden dat vroegere startleeftijden leiden tot meer ongelijkheid, en het feit dat Hoofdstuk 7 positieve effecten vindt voor verlate starts voor zwak presterende leerlingen, suggereert dat vroege startleeftijden gecombineerd moeten worden met flexibele 'cutoff dates'. Daardoor hebben studenten die nog achter blijven op de kleuterschool de ruimte om een extra jaar door te ontwikkelen.

De resultaten van Hoofdstuk 4 geven aan dat wanneer we de langetermijnprestaties van leerlingen in taal willen verbeteren, kwantitatieve veranderingen in de lestijd niet het juiste middel zijn. Echter, meer tijd in taal heeft wel waarde voor verschillende andere vakken, vooral wanneer deze extra lessen in de lagere klassen van het basisonderwijs plaatsvinden. Spillovers zijn niet aanwezig voor tijd in taal aan het eind van het primaire onderwijs, wat de algehele waarde van taallessen in deze klassen in twijfel stelt. Onze resultaten sluiten niet uit dat *kwalitatieve* veranderingen in taallessen wel een langetermijneffect hebben. Machin and McNally (2008) laten zien dat meer gericht taalonderwijs en betere organisatie van de klas tot langetermijnopbrengsten in taalprestaties kunnen leiden.

De implicaties van onze resultaten voor tracking zijn minder makkelijk zichtbaar. Toewijzing aan de hoogste track beïnvloedt uitkomsten voor studenten die rond het gevraagde niveau van de track zitten. Dit is waardevolle informatie voor studenten en hun ouders die voor de keuze staan om de hogere of lagere track te nemen. Maar het veranderen van toelatingseisen zal meer

dan alleen deze studenten beïnvloeden en zal niet noodzakelijkerwijs de algehele efficiëntie verbeteren. Wanneer we meer studenten naar de hogere track sturen, zal dat het gemiddelde niveau in beide tracks doen dalen. Het verwachte effect op de ‘untreated’ zal dan negatief zijn. Een ruwe analyse geeft aan dat deze ‘peer effects’ geen overheersende rol spelen, vooral wanneer het gaat om toewijzing aan de hoogste track. Het aantal leerlingen in die hoogste track neemt gestaag toe in Nederland en onze analyse suggereert dat dit een goede koers is. We laten ook zien dat het uitstellen van tracking het aantal hoog opgeleiden doet toenemen. Omdat het verhogen van het opleidingsniveau geen doel op zich moet zijn, is het belangrijk om te schatten of dit ook leidt tot hogere lonen op de arbeidsmarkt. We vinden dat dit het geval is voor tracking op het niveau van de twee hoogste tracks, maar onze resultaten voor tracking op het niveau van de twee laagste tracks zijn minder duidelijk. We hebben meer longitudinale data nodig om daar sterke conclusies te trekken. Desalniettemin, vroege selectie leidt in het algemeen tot een lagere kans op hoger onderwijs en vaak ook tot een lager loon, en deze effecten worden alleen positief voor een kleine groep van zeer sterk presterende leerlingen.

Ten slotte, positieve effecten van retentie zijn alleen aanwezig voor leerlingen onderaan de distributie en alleen wanneer vroeg uitgeoefend. Zittenblijven na de kleuterschool is schadelijk voor de hele distributie en deze negatieve effecten kunnen zeer sterk zijn, vooral wanneer retentie op late leeftijd plaatsvindt. In veel landen is de fractie studenten die ooit is blijven zitten groter dan 25%. Onze resultaten geven aan dat dit duidelijk te frequent is. The Technology of Skill Formation geeft aan dat vroege vaardigheden zeer belangrijk zijn voor het algehele leerproces. Retentie op latere leeftijd kan plaatsvinden wanneer de oorsprong van de slechte prestaties van de leerling in een veel eerder stadium liggen. Het feit dat zittenblijvers het vooral slecht doen op makkelijke vragen onderschrijft dit. Dit is ook in lijn met bewijs dat later herstel moeilijk is wanneer vroege investeringen niet ondernomen zijn (Cunha et al., 2010).

9.3 Reservaties en toekomstig onderzoek

De bevindingen in deze thesis zijn niet universeel. De resultaten van de Hoofdstukken 4, 5 en 6 zijn gebaseerd op Nederlandse data. Die van Hoofdstuk 3 en 7 komen voort uit internationale data en representeren daarbij een gemiddeld effect voor een breed scala aan landen. De gevolgen van veranderingen in de startleeftijd kunnen afhankelijk zijn van andere aspecten van het schoolstelsel, zoals bijvoorbeeld de nadruk op lezen en schrijven op de kleuterschool. Deze aspecten kunnen van land tot land verschillen. Desalniettemin geven we aan dat onze resultaten

niet gevoelig zijn voor veranderingen in parameters. Daarnaast zijn de landen die we meenemen in onze analyse relatief homogeen en is Nederland representatief voor vele andere ontwikkelde landen. Bovendien laten we nieuwe methodes en aanpakken zien om deze vraagstukken te benaderen, die ook toegepast kunnen worden op andere landen.

Hoewel we verschillende extensieve analyses uitvoeren die de robuustheid van onze schattingen aantonen, kunnen sommige aspecten onze resultaten potentieel beïnvloeden. We kunnen niet volledig uitsluiten dat de relatieve leeftijdspositie in de klas onze schatting van het Birth Month Effect verstoort. Hoewel we bewijs hiertegen aanvoeren, kan toekomstig onderzoek zich nog verder richten op de invloed van relatieve leeftijd ten opzichte van klasgenoten op schoolprestaties. Daarnaast geeft onze analyse van heterogene treatment effects voor retentie niet aan vanaf exact welk percentiel retentie adviesbaar is. De variabele waar we onze ranking op baseren wordt door de treatment zelf beïnvloed. Verder onderzoek kan waardevol zijn wanneer dit toetsresultaten of indicaties voor leerniveau gebruikt van voordat de treatment plaatsvindt. Daarnaast voeren we verschillende controles uit die de validiteit van onze instrumenten bevestigen, maar er bestaat helaas geen formele test voor ‘instrument validity’ in exact geïdentificeerde modellen.

De hoofdstukken in deze thesis vullen enkele lege of zwak bezette plekken op in de literatuur van de onderwijseconomie. Maar de behandelde onderwerpen brengen nog steeds veel onbeantwoorde vragen met zich mee. We laten zien wat het effect is van taaltijd op een verscheidenheid aan vaardigheden, maar een optimale verdeling van lestijd vraagt dat we de waarde van lestijd in alle schoolvakken weten. Bovendien is de waarde van specifieke vaardigheden voor latere uitkomsten, zoals lonen of opleidingsniveau, nog steeds grotendeels onduidelijk. Het ontwerp van het optimale tracking systeem vraagt ook meer onderzoek. Wij maken een onderscheid tussen tracking voor en na 14 jaar. We weten daarbij nog steeds niet hoe andere leeftijden zich tot elkaar verhouden. Daarnaast kan er ook nog gekeken worden naar het optimale aantal tracks, en de optimale divisie van leerlingen naar tracks vanuit het perspectief van de gehele studentenbevolking. Voor beleidsdoeleinden is het cruciaal dat we daarbij causale verbanden tussen veranderingen in het curriculum en verschillende uitkomsten blootleggen. Aangezien aspecten van het curriculum endogeen zijn, en aangezien experimentele variatie zeldzaam is in dit onderzoeksveld, is het cruciaal dat we gebruik maken van (andere vormen van) exogene variatie wanneer we deze effecten schatten.

Curriculum Vitae

Ron Diris was born on June 22nd, 1986 in Born, The Netherlands. He completed a Bachelor of Science (2004-2007) in Infonomics and a Master of Science (2007-2008) in International Economic Studies at Maastricht University. During his Bachelor studies, Ron spent a semester abroad at the University of California Los Angeles (UCLA).

He joined the Economics departments of Maastricht University in September 2008 as a PhD candidate. His research was supported by the Maastricht Research School of Economics of Technology and Organization (METEOR). Ron's work was or will be presented at various universities and international conferences, among them the joint meeting of the European Association of Labour Economists (EALE) and the Society of Labor Economists (SOLE) in London in 2010, the EALE conferences of 2011 (Cyprus) and 2012 (Bonn), and the European Economic Association conferences of 2010 (Glasgow) and 2011 (Oslo). He conducted part of his research as a visiting scholar at the University of Chicago (January - May 2012). Ron finished his PhD research in August 2012.

In September 2012, Ron started as a (trainee) Post-doctoral Researcher at the Center for Economic Studies at the Katholieke Universiteit Leuven. His research interests are economics of education, labor economics, human capital formation and non-cognitive skills.