Evaluating the Effects of Educational Policies

Roel van Elk

Evaluating the Effects of Educational Policies

Roel van Elk

ISBN/EAN: 978-94-91955-01-3

Printed by Pasmans Drukkerij, Den Haag

© 2014 R.A. van Elk
All rights reserved. No part of this thesis may be reproduced, stored, or transmitted in any form or by any means without prior written permission of the author.

Evaluating the Effects of Educational Policies

De effecten van onderwijsbeleid geëvalueerd

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

vrijdag 2 mei 2014 om 13.30 uur

Roel Adriaan van Elk
geboren te Nijmegen

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie

Promotor:
Prof.dr. H.D. Webbink

Overige leden:	Prof.dr. L. Borghans
	Prof.dr. A.J. Dur
	Prof.dr. P.W.C. Koning

Preface

My interest in writing a PhD thesis started ten years ago. At that time, I was finishing the Master's program in mathematical economics and econometric methods at Tilburg University. I did my Master's internship at a commercial organization specialized in online marketing. My research was about optimal advertising strategies on search engines. During my internship I found out that I enjoyed the process of conducting my own research and writing a thesis. As opposed to many of my friends that worked on their theses, I had no idea what was meant with a 'writers block'. In addition, I found out that it satisfied me that my findings could be applied in practice. I was intrinsically motivated to translate the results of my thesis into useful advices for the organization.

After finishing my Master's thesis I had the opportunity to start as a PhD student at Tilburg University. Although I considered this as an interesting option, I was not fully convinced of working as a full time PhD student. I decided to look further for a job where I could combine 'practical work' with doing a PhD research, and applied at CPB Netherlands Bureau for Economic Policy Analysis. The thing that attracted me most at CPB was the work at the crossroads of the economic sciences and public policy. I have always felt the need to focus my research on topics with clear goals and practical relevance. CPB provided the opportunity to work in a scientific environment on topics that were directly relevant for policymakers and society. I started as a young professional in 2005 with the prospect of potentially writing a PhD thesis later on in my career. During my young professional period I got the opportunity to take M-Phil courses at the Tinbergen Institute to improve my economic and econometric skills. I mostly liked the courses on (applied) micro-econometrics and a course in the economics of education by Hessel Oosterbeek and Erik Plug. After two years, I started working at the education program, headed by Dinand Webbink. We mainly used quasi-experimental evaluation techniques and set up field experiments to investigate the effects of specific educational policy measures. I enjoyed the pleasant collaboration with Dinand and my CPB colleague Marc van der Steeg, with whom I have worked on numerous projects. We often carried out our research at the request of the Ministry of Education and it satisfied me that our empirical findings were directly informative to policymakers. Dinand taught me how to write papers that were suitable to submit to an economic journal and started motivating me for writing a PhD thesis. I have to say that after having worked at CPB for about four years the idea of writing a PhD thesis had been erased a bit, but his enthusiasm made me start thinking about it again. When Dinand left CPB to work as Professor at the

Erasmus University Rotterdam in 2010, this provided a clear opportunity to turn my thoughts about writing a PhD thesis into concrete plans.

With the support of CPB I started as an external PhD student under the supervision of Dinand at the Erasmus University Rotterdam. This enabled me to combine my work at CPB with writing a PhD thesis on policy evaluations in education. I felt privileged with this unique opportunity. The fact that I had already worked on educational policy evaluations for a couple of years helped to get off to a flying start. Although work priorities naturally kept me from spending time on my thesis sometimes, the combination between my work at CPB and my PhD research turned out to be fruitful. My PhD research has benefited from substantial spillovers of my work at CPB. Most of the chapters in this thesis are based on published CPB Discussion Papers. The CPB rather generously enabled me to devote time on my PhD research, which contributed much to the progress I could make. I have enjoyed working on this thesis and I am happy to see now that my alternative route has eventually, after ten years, still resulted in a PhD thesis.

My experience with writing a PhD thesis could not have been so positive without the help and support of several important people surrounding me. First of all, I would like to thank Dinand for our fruitful discussions, his always positive and constructive comments, creative ideas, enthusiasm and sense of humour. He gave me a lot of freedom to carry out my PhD research in the way that suited me best. I feel lucky that he was my supervisor.

I am indebted to my employer, CPB, for enabling me to combine work with writing a PhD thesis. I would like to thank Coen Teulings, Casper van Ewijk, George Gelauff and Ruud Okker for giving me the opportunity to start my PhD research in 2010. I am very grateful to Bas ter Weel for his trust and support in the latter phase of my PhD research. The last two chapters of the thesis have benefited from his valuable comments. Furthermore, I would like to thank Debby Lanser for our great cooperation during the last couple of years and her flexibility when I needed time for my PhD research. My co-authors also deserve special mention. It was a pleasure working with Marc and Suzanne. Special thanks to Marc for being my paranimf. Thanks also go to my other CPB colleagues who have provided a nice and inspiring work environment each day.

Furthermore, I am grateful to the members of my PhD committee, Lex Borghans, Robert Dur, and Pierre Koning, for their time and effort.

Last but not least I would like to thank my family and friends, who have indirectly contributed a lot to my work performance. In particular I would like to thank my mother who has always been a special support during my entire educational career, and my father who always gave me the feeling that I did well. I am sure that he would have been proud. Special thanks also go to Freek, for being a ‘*broeder*’, friend and paranimf. Finally, I would like to express my love and gratitude for my wife Rieneke. Our story also started ten years ago at Tilburg University.....After writing my Master's thesis while living together on a 16m2 student room, I knew that our relationship was solid. She has been on my side ever since. I was finishing this thesis during the first months after the birth of our beautiful son Mauk. I cannot say that he contributed substantially to it. But much more importantly, he contributed most to my life. I am proud of him.

Roel van Elk
The Hague, March 2014.

Contents

1

Introduction

1.1 Motivation and purpose

This thesis focuses on identifying educational policies that work. This is important for several reasons. First, recent experience has shown us that the introduction of new policy measures has not always been a clear success. The parliamentary investigation committee on education reform, chaired by the current Dutch Minister of Finance Jeroen Dijsselbloem, evaluated some of the most significant educational policy changes in the Netherlands in the 1990s.[1] The committee concluded that the Dutch government had neglected its task to ensure the provision of high-quality education (Parliamentary investigation committee on education reform, 2008). Among other things, the committee noted the implementation of several - potentially counteracting - changes at the same time, policy choices that were chiefly based on financial motives, and the lack of a thorough problem analysis and scientific foundation that legitimized the reforms. One of the advices for future improvement concerns the use of policy evaluations to scientifically validate new measures. The committee recommends implementing a new policy instrument on a small scale first to assess its impact in case it lacks the necessary scientific basis.

The increased attention for a scientific foundation of policies is observed in other countries as well. In the United Kingdom, the Blair Government stated that policy choices should be based on scientific evidence rather than on political or ideological arguments (Blair and Cunningham, 1999). In the United States, the No Child Left Behind Act of 2001 placed emphasis on the use of interventions that have been demonstrated to work. Implementing policies that do not work can be costly and may have detrimental consequences for the quality of education.

[1] The policy reforms that were analysed include the introduction of a new type of secondary education (*'vmbo'*) and substantial adjustments in the education programs in secondary education (*'basisvorming'* and *'tweede fase'*).

Second, educational performance is an important determinant of economic and social outcomes. The interest of economists in education dates back to early studies on human capital in the 1950s and 1960s (Becker, 1964; Mincer 1958, 1962). The human capital theory treats education as an investment in the acquisition of skills to improve future outcomes. The costs include expenses from schooling and the earnings foregone by being in school rather than at work. An important benefit is the increase in future earnings. Since the initial human capital contributions a large literature has emerged that documents the importance of schooling for economic and social outcomes. Higher educated people generally earn higher wages (e.g. Card, 1999; Ashenfelter et al., 1999), are healthier (Lleras-Muney, 2005; Oreopoulos, 2007) and get less often involved in criminal activities (Lochner and Moretti, 2004; Machin et al., 2011; Webbink et al., 2013). The private returns to education, i.e. the percent increase in earnings caused by an additional year of schooling, are consistently estimated in a broad number of empirical studies within a range from 5 to 10% (Card, 1999; Ashenfelter et al., 1999; Harmon et al., 2003; Heckman et al., 2006). In addition, an increasing body of studies shows that education promotes economic growth. Most of these studies relate the average educational attainment in countries to economic growth and find social returns to education that slightly exceed the private returns (De la Fuente and Domenench 2006; Cohen and Soto 2007; Coe et al. 2009). More recent studies have also focused on the returns to cognitive skills, as proxied by test scores. The measurement of human capital by cognitive skills rather than school attainment allows taking into account quality differences of a year in education across countries. These studies provide evidence that cognitive skills are strongly related to individual earnings and to the economic growth of nations (Hanushek and Woessmann, 2008; 2012).

The importance of human capital for economic and social well-being explains the worldwide interest in the potential of government policies to improve educational performance. In most countries governments play a predominant role in the provision and financing of education. Policymakers face limited budgets and have to decide how to allocate scarce resources. This requires evidence of what works and what does not. Policy evaluations can provide insights for policymakers into the effects of government interventions that can help guide their decisions. These insights can be used to improve the education system by introducing or expanding policies that have been shown to be effective or by renouncing ineffective policies.

The complex nature of education production poses challenges for a credible policy evaluation. Student performance at any point in time is the cumulative result of the entire history of all inputs that affect student's learning, such as school and teacher characteristics, peers, family inputs, neighbourhood characteristics and the individual's ability (Hanushek, 1986; Hanushek and Rivkin, 2006). The multitude of observable and unobservable factors that affect educational outcomes makes it difficult to isolate the causal impact of a specific policy intervention. The identification of causal effects of policies has received much attention in the empirical economic literature in the past decades. This has led to a shift from control-based research approaches to design-based research methods that provide credible ways to deal with these challenges (e.g. Imbens and Wooldridge, 2009; Angrist and Pischke, 2010).

This thesis aims to contribute to the search for effective policies in education. The thesis presents four empirical evaluations of educational policies. It applies state-of-the art design-based research methods in real policy cases in the Netherlands. In each of the studies particular attention is paid to making the case for a credible identification of the causal effect of the policy. The thesis adds to different strands of the literature of the economics of education by providing new evidence on the effects of different types of policy interventions. The interventions studied concern one institutional policy, one incentive-based policy and two resource-based programs. The first study investigates the effect of the timing of tracking on higher education completion. The second study focuses on the effectiveness of a new type of financial incentive for regional educational authorities in reducing school dropout. The third study evaluates the effects of a special program for multi-problem school dropouts on schooling, labour market and crime outcomes. The fourth study examines the impact of a comprehensive school reform policy for failing schools on student achievement.

1.2 Developments in policy evaluation

The multitude of factors that influence educational outcomes poses serious challenges for researchers aiming to determine the causal impact of policy interventions. A reliable policy evaluation requires that the effects of the intervention can be credibly disentangled from the effects of other, potentially unobserved, factors.

An essential principle underlying policy or treatment evaluations follows from the potential outcome model (Rubin, 1974; 1977). This model assumes that each subject has two potential outcomes: an outcome when exposed to the treatment and an outcome when not exposed to the treatment. The causal effect of the treatment then follows from the difference between the two potential outcomes. In case of heterogeneous effects of the treatment, the causal average treatment effect follows from the population's average difference between the two potential outcomes. In practice, however, we observe at most one of the outcomes. The potential outcome without treatment is not observable for subjects that are exposed to the treatment. We thus have to rely on another group that is not exposed to the treatment for the potential outcome without treatment. The critical assumption needed to determine the causal effect of the treatment is that the observed outcome of this control group equals the potential outcome of the treatment group had they not been exposed to the treatment. Hence, the most important challenge for the identification of causal effects is to find a credible control group that satisfies this assumption.

Self-selection into a treatment or control condition threatens the validity of this assumption. Individuals who choose to be exposed to the treatment are by definition different from individuals who choose not to be exposed to the treatment. These, potentially unobservable, differences can invalidate a comparison of the outcomes between treatment and control groups. For instance, when evaluating the effects of class size on student achievement, one should be aware that choices made by schools or parents can affect the assignment to a small or a large class. Schools may be willing to group weaker students together in a small class. Parents that are more involved with their child's learning may arrange that their child is placed in a small class. In such cases, differences in outcomes between students in small and large classes not only reflect the effect of class size but also the effects of confounding factors such as (unobserved) ability or family inputs.

Selection is the most important problem in policy evaluation studies. If the policy variable of interest is correlated with omitted variables or unobservable factors that affect the outcomes, standard regression methods fail to produce reliable estimates of the impact of the policy. These problems for the identification of causal effects received little attention in the early literature on education production. Hanushek (1986) summarized the result of 147 studies that estimated education production functions. The estimated effects of school inputs such as class size and teacher characteristics varied widely across those studies. For each of the inputs the results differed both in sign and statistical significance. The author concluded from these conflicting results that there appears to be no clear relationship between school inputs and student achievement. However, selection bias is a serious concern in the studies reviewed that may have caused the inconsistency in estimated effects. The increased interest in appropriate ways to handle selection problems originates mainly from studies in the field of labour economics on the impact of government training programs (Ashenfelter, 1978; Ashenfelter and Card, 1985; Lalonde, 1986). These studies found difficulties in establishing robust results and questioned the reliability of econometric evaluation techniques. An influential study by Lalonde (1986) examined the impact of a federal training program for disadvantaged workers on earnings and compared the outcomes of two different approaches. The first approach was based on random assignment to the program, while the second was based on an econometric evaluation. The outcomes of both approaches differed and the author concluded that studies that do not take into account selection issues can yield biased estimates. This conclusion largely contributed to the awareness of the importance of properly addressing selection problems for identifying causal effects. Consequently, the research focus in empirical economics has shifted more and more towards research designs that provide solutions for the selection problem from the early 1990s onwards. This has resulted in an increase of design-based studies that pay explicit attention to the identification of treatment effects. The shift from control strategies and econometric methods to design-based research has been an important development in empirical economics in the past decades.

This development, referred to as 'the credibility revolution in empirical economics' by Angrist and Pischke (2010), has contributed largely to the understanding of credible ways to deal with selection issues in policy evaluations. The most convincing way to establish causal effects are (social) experiments, in which individuals are randomly assigned to a treatment or control condition. Random assignment assures that treatment status is not correlated with omitted variables. In such experiments the difference in mean outcomes by treatment status gives an unbiased estimate of the average treatment effect. In a regression framework, regressing the outcome variable on an intercept and an indicator for the treatment yields an unbiased estimate for the average treatment effect. Adding covariates to the regression specification can increase precision of the estimates. The number of randomized experiments in empirical economics has grown rapidly over the past decades. One of the largest randomized trials in the field of education is the Tennessee STAR experiment (Krueger, 1999). Students in primary school were randomly assigned to one of three groups: small classes, large classes, and large classes with a full-time teacher's aide. The results showed that a small class modestly increases student achievement and that the inclusion of a teacher's aide has little effect on student performance. More recent examples of social experiments include studies on the use of financial incentives in education (e.g. Angrist and Lavy, 2009; Angrist et al., 2009; Kremer et al., 2009).

However, randomized experiments in education are often not feasible. Experiments are expensive, time consuming and may not always be appropriate or desirable. Moreover, it may be difficult to get approval for an experiment since random assignment of students is often considered to be unethical. In case a randomized experiment is not possible, natural or quasi-experiments can provide credible solutions. These quasi-experiments exploit some kind of exogenous variation in treatment induced by institutions, specific government rules or the forces of nature. We distinguish three dominant types of natural experiments: instrumental variables models, regression discontinuity designs and difference-in-differences models.

1.2.1 Instrumental variables

Instrumental variables (IV) models address the selection problem by the use of an exogenous variable, the instrument, which is correlated with the treatment variable but otherwise uncorrelated with the outcome variable. Instrumental variables can overcome the estimation biases that arise from selection by using only the exogenous variation in the treatment status that is induced by the instrument. The IV method allows estimating treatment effects consistently with two-stage-least-squares (2SLS). If treatment effects are homogeneous, this yields the average treatment effect among all subjects. Imbens and Angrist (1994) show that, in case of heterogeneous effects, the IV estimate yields a local average treatment effect (LATE). They introduce the assumption of monotonicity, which implies that the instrument affects treatment status in a monotone way, i.e. that an increase in the value of the instrument does not decrease the value of the treatment status. Under the monotonicity assumption and the standard IV assumptions that the instrument (i) has a clear impact on the treatment variable (first-stage), (ii) is independent of potential outcomes, conditional on covariates, and (iii) only affects the outcome variable through the treatment status (exclusion restriction), IV estimates yield a causal treatment effect for the subpopulation of individuals whose treatment status is affected by the instrument. Hence, estimated treatment effects using IV should be interpreted as a LATE that applies only to the subsample of subjects that comply with the instrument.

The art of applying IV models lies in finding appropriate instruments that satisfy the required assumptions. Special cases are studies that use a randomized assignment to treatment as an instrument for actual treatment (e.g. Angrist, 1990; Bloom et al., 1997). The actual treatment status may differ from the instrument if subjects do not comply with their assignment. In such cases with a binary instrument, the assumption of monotonicity partitions the population into three subgroups: compliers, never-takers and always-takers (Angrist et al., 1996). Compliers are those who always comply with their assignment. Both never-takers and always-takers are the subjects whose treatment status is not affected by the instrument. These subjects never (or always) take the treatment, independent of their assignment. Since the treatment is invariant to the assignment, the IV estimates are uninformative on the treatment effects for these subgroups. In cases with one-sided noncompliance, the LATE can be shown to equal the average treatment-on-the-treated effect (Bloom, 1984). This occurs when participation in the treatment is voluntary among those assigned to the treatment, while those not assigned are not allowed to participate in the treatment. The intuition is that such situations rule out the existence of always-takers, which implies that all treated individuals are compliers.

Overviews of the use of IV in economics are presented in Angrist and Krueger (2001) and Angrist and Pischke (2009). An example where IV has been frequently used is the literature on the returns to schooling. Educational attainment is an endogenous variable that is related to many factors such as (unobserved) ability or motivation. Instruments for schooling can help overcome endogeneity biases in the estimation of the returns to education. These studies have used variation in schooling induced by differences in the regional supply of schools (Card, 1995), tuition costs (Kane and Rouse, 1995) or compulsory schooling laws (Angrist and Krueger, 1991; Acemoglu and Angrist, 2001; Oreopoulos, 2006).

1.2.2 Regression discontinuity designs

Regression discontinuity designs have been used in the field of psychology dating back to the 1960s and have only recently come to the forefront in the economics literature (Cook, 2008). Regression discontinuity designs exploit threshold values that are used in rules for treatment assignment. The treatment status then depends on the value of an underlying variable. Subjects with values above (or below) the cut-off are assigned to the treatment, while subjects with values below (or above) the cut-off are not. Regression discontinuity designs essentially compare the outcomes of subjects just above and just below the threshold value, while controlling for a smooth function of the underlying variable. The assumption needed for identification of a causal treatment effect is that subjects below the threshold value do not differ from subjects above the threshold value, conditional on a smooth function of the underlying variable and covariates. This assumption implies that there are no other confounding discontinuities around the cut-off. To contribute to the credibility of the identifying assumption, regression discontinuity designs are often estimated in a narrow range around the cut-off.

Regression discontinuity designs can be distinguished in sharp and fuzzy designs (Hahn et al., 2001). In a sharp regression discontinuity design, treatment assignment is a deterministic function of the underlying variable, i.e. all subjects above the threshold value are treated, while all subjects below the threshold are not. In fuzzy regression discontinuity designs, the probability of receiving treatment is discontinuous at the threshold value, but does not change from zero to one. Fuzzy regression discontinuity designs can be analyzed using instrumental variables, in which a dummy variable for being on either side of the cut-off can be used as an instrument for treatment. The ratio between the impact of the instrument on the outcome variable (reduced form) and the impact of the instrument on the treatment status (first-stage) is an IV-estimate of the impact of the treatment.

Imbens and Lemieux (2008) and Van der Klaauw (2008) review the use of regression discontinuity designs in empirical economics. An example in the field of education is Jacob and Lefgren (2004), who examine the impact of remedial education on student performance. In addition, a series of studies have investigated the impact of class size on student achievement making use of discontinuities that follow from maximum class size rules. These rules create exogenous variation in class size. The intuition is that cohorts with a total number of students that just exceeds the maximum size is split in two small classes, while cohorts with a total number of students that is just below the maximum size is left in a single large class. A seminal paper by Angrist and Lavy (1999) exploited these rules in Israel in a regression discontinuity framework and found that small classes improve student achievement. More recently, other studies that used a similar regression discontinuity design showed results that support these findings (Urquiola, 2006; Browning and Heinesen, 2007).[2]

1.2.3 Difference-in-differences

Difference-in-differences (DID) models make us of pre-treatment and post-treatment outcomes for subjects that are affected by a policy change and subjects that are not, or less, affected by the same change. This type of model is particularly applicable if a policy intervention is implemented in specific regions. DID models essentially compare the change in outcomes after and before the start of the policy in the treated regions (the treatment group) to the same development in other regions that are not treated (the control group). The identifying assumption is the common trend assumption. This assumption implies that the development of outcomes in the control group would have been the development of outcomes in the experiment group in the absence of the policy change.

Difference-in-differences models have been widely used in empirical economics (see e.g. Ashenfelter and Card, 1985; Card and Krueger, 1994; Blundell et al., 1998). Examples in the field of education include Pischke (2007), who evaluates the impact of the length of the school year on student performance, and several studies that examine of the impact of accountability systems on educational achievement (Hanushek and Raymond, 2004; Jacob, 2005; Dee and Jacob, 2011). These latter studies exploit the fact that accountability reforms have been implemented in specific states in the Unites States and provide evidence that such systems can help to improve educational performance.

[2] Whereas earlier studies on the impact of class sizes showed mixed results (Hanushek, 1986), the recent design-based studies turn out to yield a more consistent pattern of results that point at modest improvements in educational achievement induced by small classes. The reduced variation in outcomes across studies may be well explained by the more appropriate designs used to address selection problems.

1.3 Outline and results

This thesis presents four empirical evaluations of educational policies in the Netherlands. In each of the studies specific attention is paid to the identification of the causal effect of the policy by making use of design-based research methods. The interventions studied concern one institutional policy, one incentive-based policy and two resource-based programs. Each of the Chapters 2-5 presents a separate study on the impact of a particular educational policy that is readable in itself. The outline of the thesis is presented below including a preview of the main results.

Chapter 2 investigates the effect of the timing of tracking on higher education completion. This chapter relates to the literature on institutional features of education systems. The age at which students are tracked into different types of education is one of the most remarkable differences in education systems across countries. Previous studies on school tracking have often relied on differences between countries or states (Schuetz et al., 2005; Hanushek and Woessmann, 2006). The Dutch education system offers the opportunity to investigate the effect of the age of tracking within a single school system. Some schools in the Netherlands directly track 12 year old pupils that leave primary education into categorial classes of a particular education level, while other schools offer comprehensive classes that postpone the time of tracking with one or two years. This causes variation in the timing of tracking, which we exploit in our analysis. The main challenge for the identification of causal effects is the potential self-selection of students into tracked or comprehensive classes. To address this potential problem we control for a large set of covariates including various measures of socioeconomic background and test scores on arithmetic, language and information processing. Moreover, we adopt an instrumental variable approach in which we exploit regional variation in the supply of schools. The estimation results provide evidence that early tracking negatively affects the probability of higher education completion.

Chapter 3 investigates the effects of a new type of financial incentive targeted at regional education authorities in the Netherlands. This chapter relates to the debate on the use of resource-based versus incentive-based policies in education. Many empirical studies fail to find a positive association between increased resources or inputs and an improvement in educational quality (Woessmann, 2003; Hanushek, 2006). The conclusion that additional resources are no guarantee for improved student performance has shifted the interest of policymakers and researchers more towards the use of incentive-based policies in education. A variety of new incentive-based policies have been started around the world. Empirical evaluations of these programs show that financial incentives can improve education quality (e.g. Angrist and Lavy, 2009; Kremer et al., 2009; Lavy, 2002, 2009). At the same time there remains severe opposition from teacher unions and educators against the use of incentives and strategic responses seem to be a serious threat to the effectiveness of the programs (Glewwe et al., 2003; Jacob, 2005; Figlio and Getzler, 2006). Previous studies have mainly focused on incentives for students, teachers and schools. This chapter aims to contribute to the literature by analysing the effectiveness of a financial incentive scheme for regional education authorities. These authorities were selected based on their pre-treatment number of school dropouts and received additional resources if they reduced school dropout compared to a baseline level. The introduction of the policy in 14 out of 39 regions and the use of the specific selection rule for the participating regions allow us to estimate local difference-in-differences models. This approach essentially combines a difference-in-differences model with a regression discontinuity design. It estimates difference-in-differences models around the cut-off level for being selected into treatment. Using administrative data for all Dutch students in the year before and the year after the introduction of the new policy we find no effect of the financial incentive scheme on school dropout. In addition, we find suggestive evidence for strategic behaviour in response to the program.

The next two chapters are dedicated to an investigation of the potential of comprehensive programs in improving outcomes for troubled individuals or schools. Chapter 4 investigates the impact of the Neighbourhood School Program (NSP), a special program designed to increase school enrolment and employment among a problematic target population of young school dropouts in Rotterdam. Young school dropouts are at risk for future unemployment or getting involved in criminal activities (e.g. Lochner and Moretti, 2004; Machin et al., 2011). The high social costs associated with unemployment and crime legitimize publicly subsidized interventions aimed at improving the prospects for these youths. Previous studies, however, have shown that is difficult to design effective interventions that help this problematic group back on track (LaLonde, 2003). The NSP is largely designed in line with the most promising existing programs and provides a broad range of educational, work, and health services, and guidance by personal coaches. This chapter provides evidence on the effects of the program on school enrolment, employment and criminal behaviour from a field experiment. Since random assignment of the target group to the program and control condition was infeasible, we set up a controlled natural experiment. During specific time windows potential candidates for the program were assigned to the regular interventions used by the municipality. We use the youths that were eligible for the program but that were assigned to the regular interventions because of these time windows as our control group. Our design illustrates a potential set-up of field experiments in cases where random assignment is not feasible. We use administrative information about school enrolment, employment, and criminal behaviour three years after the start of the intervention. The impact of the program is identified by comparing the outcomes of youths assigned to the NSP with the outcomes of youths assigned to regular interventions, conditional on the time of application. We use an instrumental variables approach to address noncompliance with the assignment rule. The estimated effects show that assignment to the program did not increase school enrolment or employment. Most important, we find evidence that assignment to the NSP increased criminal activity, especially among the youths who had been criminally active before the time of entry. This chapter relates to the literature on school dropout and crime. The results are consistent with a large body of the literature that shows no impact of training programs on the labour market outcomes of at-risk-youths and with studies that document the adverse effects of group-based interventions on delinquent behaviour (Dishion et al., 1999; Dodge et al., 2007).

Chapter 5 focuses on the impact of a comprehensive school reform (CSR) policy for failing schools. CSR methods have been widely used to improve the educational quality of failing schools, especially in the United States in the 1990s and early 2000s. Such programs involve a set of integrated changes at all levels within schools rather than isolated interventions targeted at single aspects. Yet the evidence on the effectiveness of such programs in increasing student outcomes is limited, and prior research often did not use credible research designs to handle potential selection issues (Borman et al., 2003). We use a difference-in-differences approach to estimate the effects of a CSR policy, the Amsterdam School Improvement Program (ASIP), on pupil achievement. The program implements a systematic and performance-based way of working and integrates measures such as staff coaching, teacher evaluations and teacher schooling, and the use of new instruction methods. As of 2008, all primary schools in Amsterdam that were judged to perform below national quality standards were invited to voluntarily participate in the program. We make use of administrative data on the high-stakes CITO test scores from 2005 to 2012, which allows us to compare the development of pupil achievement in failing schools in Amsterdam to that in failing schools outside Amsterdam in a difference-in-differences framework. We present intention-to-treat estimates and find substantially negative effects on test scores in the first four years after the introduction of the program. The introduction of the ASIP decreases test scores by 0.17 standard deviations. A potential explanation for this finding is the intensive and rigorous approach that caused an unstable work climate with increased teacher replacement. The resulting loss of school specific knowledge, increase in recruitment and hiring costs, and uncertain work atmosphere felt by teachers may have negatively affected pupil achievement.

Chapter 6 presents a summary of the main findings, followed by a brief discussion on its limitations, suggestions for further research and some overall concluding remarks.

2

Does the timing of tracking affect higher education completion?[1]

Abstract

This chapter investigates the effect of the timing of tracking on completion of higher education by exploiting unique variation from the Dutch education system. At the age of 12 Dutch students can enrol in tracked schools or in comprehensive schools. The comprehensive schools postpone enrolment into tracked classes by one or two years. OLS- and IV-estimates, using regional variation in the supply of schools as instruments, show that early tracking has a detrimental effect on completion of higher education for students at the margin of the Dutch high and low tracks. The negative effects of early tracking are larger for students with relatively high ability or students with a higher socioeconomic background. In addition, we find no negative effects of comprehensive classes on higher ability students. These results suggest that increasing participation in comprehensive classes would increase graduation from higher education.

[1] This chapter is based on joint work with Marc van der Steeg and Dinand Webbink: Van Elk, R., M. van der Steeg, D. Webbink, 2011, Does the timing of tracking affect higher education completion?, *Economics of Education Review*, 30, 1009-1021.

2.1 Introduction

One of the most remarkable differences between school systems is the age at which students are tracked into different levels of education. Countries such as Germany, Austria, Hungary and the Slovak Republic track students as early as age 10. Other countries, such as Norway, Sweden, Canada, the US, the UK and Japan, have a comprehensive school system until the end of secondary education. Differences in the age of tracking might have far-reaching consequences for the equity and efficiency of educational outcomes. However, little is known about the effects of this key institutional feature of education systems.

A researcher interested in the effect of the age of tracking would like to know what would have happened with a student if he/she had not been tracked at, for example, the age of 12 but rather at 14. Investigating this question requires variation in the age of tracking. However, in most countries or states tracking takes place at one age or not at all. Hence, variation in the age of tracking is often not available within a single education system. Therefore, researchers often rely on variation between countries or states, and estimate regression models using a large set of covariates (for instance, Schuetz et al. 2005). The multitude of potential confounders due to differences between countries or states poses serious problems for the identification of the effect of the age of tracking in these models. A recent innovative study by Hanushek and Woessmann (2006) addresses this problem by using a difference-in-differences approach exploiting international test scores taken at different ages. However, this study has also been criticized for neglecting incentive effects of tracking (Eisenkopf, 2007). Studies that investigate the effects of tracking within countries compare the outcomes of tracked students with non-tracked students (Figlio and Page, 2002; Duflo et al. 2008) or exploit the gradual introduction of a comprehensive system (Pekkarinen et al., 2009). To our knowledge there are no previous studies that exploit differences in the timing of tracking within the same education system.

The Dutch system offers the opportunity to investigate the effect of the age of tracking within a single school system. At the age of 12 Dutch students can enrol in tracked classes (categorial classes) or in comprehensive classes. The comprehensive classes consist of a combination of school types and take one or two years. Basically, these comprehensive classes delay enrolment into tracked classes until age 13 or 14. This chapter compares higher education completion rates of students that are tracked at the age of 12 with the outcomes of students that are tracked at the age of 13 or 14 within the Dutch education system. The advantage of this approach is that students in the tracked and comprehensive parts of the system encounter exactly the same educational environment after the period of treatment. Hence there are no concerns for confounding factors due to differences between countries. Relative to existing studies that compare outcomes of tracked versus non-tracked pupils, our paper contributes by comparing two alternative tracking ages. We compare a group of pupils that is tracked early with a group of pupils that is tracked later. This implies that there are two differences in treatment between both groups that may affect future educational outcomes. First, there is the difference of being in a tracked versus a comprehensive class for one or two years. Tracked pupils might, for example, benefit from more homogeneous classrooms, while pupils in a comprehensive class might benefit from the interaction with their peers. Second, there is a difference in the age at which pupils enter the tracked system in which they are allocated to different education types. This may also affect outcomes if the ability of teachers and parents to select the most suitable education types for pupils depends on the pupil's age. Tracking pupils at an early age may imply more uncertainty with respect to the pupil's true capabilities and hence a higher risk of sending them to an inappropriate school type. Hence, the probability of misallocation of pupils to tracks may decrease with tracking age. This aspect is not taken into account in studies that compare tracked versus non-tracked pupils. Suppose a researcher compared the effects of being tracked versus non-tracked at age 12 and found a positive result for tracking. Tracking at age 14 might be better still if a positive effect of realizing better matches between students and education types at a later tracking age dominates the negative effect of being in a comprehensive class for one or two years. This illustrates that comparing tracking ages, taking into account both aspects, is of critical importance for policymakers aiming to improve the school system.

The main challenge for our analysis is the potential self-selection of students into categorial or comprehensive classes. If categorial or comprehensive classes attract more able or more motivated students, unobserved differences might bias the estimates. To address this potential problem we control for a large set of covariates including various measures of socioeconomic background and test scores on arithmetic, language and information processing. Moreover, we adopt an instrumental variable approach in which we exploit regional variation in the supply of schools. Previous papers in the economic literature used regional differences in the supply of schools to instrument for educational attainment (see for instance Card, 1995; Currie and Moretti, 2003; Park and Kang, 2008). In this paper we use regional differences in the supply of tracked or comprehensive classes to instrument for the age of tracking.

The main outcome variable in the analysis is the completion of higher education. This outcome plays a prominent role in the Dutch policy discussion on participation in higher education. A recent report suggests that the early tracking regime in the Netherlands constrains the growth of higher education participation significantly (OECD, 2007). We focus the analysis on a sample of students that is most likely to be affected by early tracking: those who were advised to enrol in the lower track of secondary education. In addition, we investigate the effects of early tracking for higher ability pupils that were advised to enrol in the higher track of secondary education.

Our main finding is that tracking at the age of 12 compared to tracking at 13 or 14 has a negative effect on the probability of higher education completion. Both ordinary least-squares (OLS) and instrumental variables (IV) estimations show negative effects of early tracking. We also find that the comprehensive school has no negative effect on pupils with a higher ability. This suggests that postponing tracking to older children would increase the number of people completing higher education.

The structure of this chapter is as follows. Section 2.2 discusses the previous literature. Sections 2.3 and 2.4 describe the Dutch context and data. The empirical strategy is discussed in Section 2.5. Section 2.6 presents the main estimation results. In Section 2.7 several robustness analyses are presented. Section 2.8 presents an additional analysis for higher ability students. Finally, Section 2.9 concludes and discusses potential policy implications.

2.2 Previous studies

The empirical literature has produced conflicting results on the effects of tracking. Several studies use cross-country data and regress international test scores on institutional differences between countries. Ariga and Brunello (2007) study the effect of the number of years spent in a tracked system on the performance of young adults in a standardized cognitive test. Using constraints on educational participation, such as financial constraint or family reasons, as instrumental variables they find a positive effect of tracking on performance. Hanushek and Woessmann (2006) use a difference-in-differences approach to identify the effect of an early tracking regime on standardized test scores. They match international primary school tests to secondary school tests and compare differences in test scores across countries. They find that early tracking increases inequality, while it does not have a clear impact on average achievement. Recently, this study has been criticized. Jakubowski (2007) argues that the differences in design of the international tests may affect the results. While PISA measures pupils aged 15 (independent of their grade), PIRLS/TIMMS measures achievement in specific grades (independent of age). He concludes that early tracking does not increase inequality when comparing pupils of the same age and grade. Eisenkopf (2007) argues that the findings by Hanushek and Woessmann (2006) are biased because they neglect the incentive effects of tracking. Brunello and Checchi (2007) focus on the interaction between family background and tracking. They estimate regression models with many covariates including country-by-cohort dummies and interactions with family background. Their estimates suggest that tracking reinforces the role of family background with respect to educational attainment and labour market outcomes. Schuetz et al. (2005) use a similar approach in which they investigate the effect of the interaction of family background with an indicator of school tracking (for which they take the age of selection into different tracks) on test scores of pupils in over 50 countries. They also find that early tracking increases the effect of family background.

A second group of studies uses variation within countries. Duflo et al. (2008) provide evidence from a randomized experiment in Kenya. In this experiment 121 primary schools, who all had a single grade 1 class, received funding to hire an extra teacher and split this class into two sections. In 60 randomly chosen schools the pupils were randomly assigned to one of the two sections. In the other 61 schools the pupils were assigned based on a ranking of their previous educational performance. After 18 months pupils in the schools that selected based on performance scored on average 0.14 standard deviations higher than pupils in schools that selected randomly. In addition they find that pupils at all levels of the distribution benefited from tracking. The authors suggest that tracking was beneficial because it helped teachers to focus their teaching to a level appropriate to most students in class. Evidence for developed countries is provided by several studies that exploit educational reforms from countries that switch from a tracked to a comprehensive system. Galindo-Rueda and Vignoles (2005) investigate the effect of the change from selective schools towards a comprehensive system in England. However, Manning and Pischke (2006) note that the gradual introduction of the new system started in poorer areas which might bias the estimates. A recent study by Pekkarinen et al. (2009) exploits the replacement of the Finnish two-track school system by a uniform nine-year comprehensive school. Identification comes from the gradual introduction of the reform during a six-year period. They find that the reform had a small positive effect on verbal test scores but no effect on the mean performance in arithmetic or logical reasoning tests. In addition, the reform improved test scores of students from lowly educated families. Figlio and Page (2002), using US data, investigate the effect of tracking on the improvement in maths test scores between the 8th and 10th grade for different ability groups. They divide the students according to the 8th grade test score into top, middle and bottom thirds of the distribution and estimate separate regressions for each of these subsets in which they exploit variation in tracking across schools. They find no significant effect in each of these regressions and interpret this as evidence that tracking does not harm the low-ability students. The results from additional two-stage least squares estimations suggest that low-ability students may actually gain from tracking. Betts and Shkolnik (2000) examine the effect of formal ability grouping in the US, making use of a nationally representative data set which asks principals whether their schools use tracking in their math classes.[2] The authors find no overall effect on math achievement growth and find little or no differential effects for high-achieving, average or low-achieving students.

Our study provides the following contributions to the literature.

First, we examine the effects of the timing of tracking. In the Netherlands, each pupil ends up in a tracked system at age 14 and we investigate whether tracking at age 12 improves on tracking at 13 or 14. Previous studies, as cited above, investigate whether tracking improves educational outcomes or not.

Second, we exploit the coexistence of a tracked and a comprehensive system within the same school system. Previous studies exploit the gradual introduction of a reform from a tracked system to a comprehensive system within countries, which means that these systems have not been in existence for a long time.

Third, we investigate long-term outcomes of tracking. Most of the previous work focuses on test scores at a relatively young age. This might compromise a proper evaluation of the impact of tracking on performance if tracking is not given sufficient time to work out its effect. We look at later outcomes by comparing higher education completion rates across early versus later tracked students.

2.3 Variation in the timing of tracking in Dutch education

In order to analyse the effects of early tracking, we would like to compare the outcomes of pupils that are subject to an early tracking regime with the outcomes of those who are not. Although from an international perspective the Dutch educational system as a whole is considered to be an early tracking regime, there exists some variation in the timing at which pupils are placed in a certain track. At the start of secondary education, schools in the Netherlands offer different types of first-grade classes. In some schools, 12 year old pupils are directly tracked into categorial classes of a certain education level. Other schools offer one- or two-year comprehensive classes, in which pupils are kept together before they are tracked in a particular school type. Hence, pupils starting secondary education in a comprehensive class postpone their choice of school type by at least one year. This difference causes variation in the timing of tracking, which we can exploit to analyse the effects of early tracking by comparing the educational outcomes of pupils who start secondary education in a categorial class (the ‘tracked’ pupils) to those who

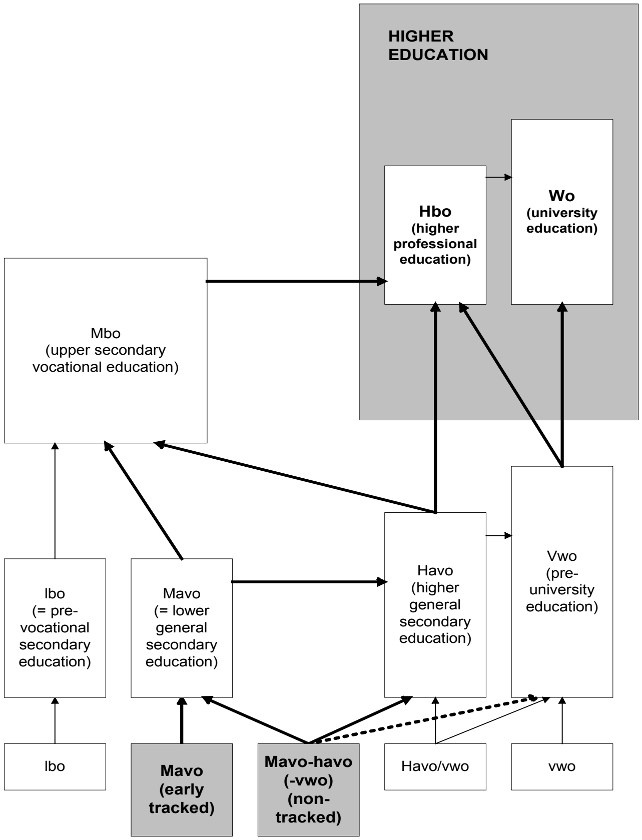
[2] Rees et al. (2000) have criticized the study because of concerns on the reliability of principal-reported measures of tracking.

start in a combined class (the ‘non-tracked’ pupils). Strictly speaking, we investigate the effect of being tracked one or two years earlier.

The structure of Dutch education after leaving primary education at the age of 12 is shown in Figure 2.1. In 1989 secondary education consisted of four tracks: ‘lbo’ (pre-vocational secondary education), ‘mavo’ (lower general secondary education), ‘havo’ (higher general secondary education) and ‘vwo’ (pre-university education). After secondary education students can enrol in ‘mbo’ (upper secondary vocational education) or higher education, depending on the type of school completed. Mbo is oriented towards vocational training and is offered at four levels which prepare students for specific professions in the labour market. Higher education in the Netherlands is offered in two types of institutions: research universities, which offer research oriented programs (‘wo’) and universities of applied sciences, which offer programs of higher professional education (‘hbo’) which prepare students for particular professions. The minimum access requirements for higher education are a havo degree or a level four (the highest level) degree of ‘mbo’. Hence, pupils who have completed havo or vwo are qualified for access to higher education, while pupils with a mavo degree can not directly enrol. We label all pupils who start secondary education in a stream that does not give direct access to higher education (mavo and lower) as being assigned to a low track. Similarly, all pupils who start secondary education in a ‘havo’ or higher are assigned to a high track. Both groups of pupils can be considered to be subject to an ‘early tracking regime’. After all, both are tracked at the age of 12 in either the low track or the high track. The pupils who enter some combined class that consists of at least mavo-havo delay tracking until age 13 or 14. We make no distinction whether this is a mavo-havo, mavo-havo-vwo or another combined class that includes both mavo and havo. As long as the choice of both the low track and the high track is open it is labelled as a comprehensive class. Those pupils can be considered to be subject to a ‘comprehensive system’.

In the literature different definitions of tracking are used. For example, the US has, strictly speaking, a comprehensive school system and employs streaming within schools, which is a milder form of ability grouping. While tracking implies that students are placed in different school types, streaming implies that particular courses are taught at different levels of complexity. Often, this distinction is ignored in the existing literature. We define tracking as the allocation of pupils into high or low track classes. In the Netherlands first grade categorial classes are in general offered by categorial schools and first grade comprehensive classes by comprehensive schools. Parents are equally free to choose for categorial or comprehensive schools and there are typically no differences in the way courses are taught between the two school types.

Figure 2.1 Flow chart of Dutch education system: routes towards higher education for tracked versus non-tracked pupils in first year of secondary education.


HIGHER EDUCATION
Hbo
(higher professional education)
Wo
(university education)
Mbo
(upper secondary vocational education)
Ibo
(= pre-vocational secondary education)
Mavo
(= lower general secondary education)
Havo
(higher general secondary education)
Vwo
(pre-university education)
Ibo
Mavo
(early tracked)
Mavo-havo
(-vwo)
(non-tracked)
Havo/vwo
vwo

Only a few schools offer both comprehensive and categorial classes. The course programs that are offered at these schools are in principle the same as those offered at the single comprehensive or categorial schools

In the Netherlands pupils at the end of primary education are advised about the type of secondary school most appropriate for them. The advice on education level is not binding, but it is a strong recommendation taken seriously by the secondary schools. It can be interpreted as a proxy for perceived ability. In our analysis, we primarily focus on pupils that are advised to follow the mavo track when they leave primary education and compare higher education completion rates for those who are immediately tracked in a categorial mavo to those who postpone their choice by entering a combined mavo-havo or mavo-havo-vwo class. Hence, we select a homogeneous group with respect to ability and compare the subsample of this group that moves into a low track to the subsample that moves into a comprehensive class. From the OECD-perspective (see introduction and OECD, 2007) this group seems most interesting as the former subsample does not qualify for direct access to higher education. If early tracking has a detrimental effect on the completion of higher education, the mavo advice pupils who are tracked early are likely to be those who are most damaged due to their subsequent inability to access higher education directly.

Our main analysis addresses the question whether pupils that are recommended by the teachers to go to the low track benefit from being in a comprehensive class compared to being in the low track. In addition we will analyse the effects of tracking for higher ability pupils, by focusing on pupils that are advised to follow the havo track and compare outcomes of those who enter a comprehensive class to those who are tracked early into the high track.

2.4 Data

In our empirical analysis, we use data from the Secondary Education Pupil Cohort 1989. These are longitudinal data collected by Statistics Netherlands and the University of Groningen (Statistics Netherlands, 1991; Driessen and Van der Werf, 1992). The cohort consists of a representative sample of around 20,000 pupils who enrolled in the first grade of secondary education in 1989. These pupils were followed during their school careers until they left the education system. Hence, for each pupil in each calendar year their corresponding school type and grade is known. The public files of the Secondary Education Pupil Cohorts follow the pupils until the school year 2003-2004. We received access to the private files of Statistics Netherlands with the most recent available data until the school year 2007-2008. This provided us with the most up-to-date information on completion of higher education. A comparison of our sample with the population shows that the shares of pupils who completed higher education (around 30%) in the cohort 1989 are quite similar to the nationwide population figures.[3]

Our data include information on the highest level of education completed from which we derive a dummy variable for completion of higher education. The data also provide information on a pupil's school type

[3] In the Dutch Labour Force Survey 23.6% of the students aged 24 completed higher education in 2001. Furthermore, the share of pupils aged 25-34 with a tertiary education degree is 34% in 2004, while the share of the population aged 25-65 with a tertiary degree is 29% (Minne et al., 2007).

and grade for each calendar year. We observe whether a pupil is enrolled in secondary education in a categorial mavo class or in a comprehensive mavo-havo or mavo-havo-vwo class. The former group is tracked early while the latter delays tracking by one or two years. From this information we construct our main tracking variable. Our data contain a large set of covariates. First, the data provide a large set of socioeconomic background variables. The parents of the pupils were asked to fill in questionnaires which include questions on ethnicity (5 categories), education level (7 levels), profession level (7 categories) and family composition (8 categories). Second, the data contain information on the urbanization of residence (5 categories), gender and age of the pupils. In addition, we condition on scores from 'entrance tests' in arithmetic, language and information processing, which pupils undertook at the start of secondary education. These tests are comparable to the main standardized test at the end of Dutch primary education (the CITO test) and the results serve as an indicator for pupil ability. Scores on each of the tests are between 0 and 20 (maximum score).

The sample of pupils that leave primary education with a mavo-advice includes 4912 pupils, of which 3123 (63.6 %) enter secondary education in a categorial mavo class and 1190 (24.2 %) in a combined mavo-havo or mavo-havo-vwo class. The sample statistics of the variables we use in our analysis are shown in Table 2.1. We exclude pupils who are missing these variables, which leaves us with an estimation sample of 3936 students.[4] Column (1) reports the sample means for the group of pupils who start secondary education in a tracked class, column (2) reports them for the group of pupils who start secondary education in a comprehensive class. Column (3) reports the *p*-value of the difference, calculated using a two-tailed *t*-test or a chi-squared test.

A comparison of the covariates of the two groups of students reveals some differences in socioeconomic background. Parents from students in comprehensive classes are slightly higher educated and their professional level is also slightly higher. Parents of students in tracked classes have Dutch nationality more often.

[4] There does not appear to be any systematic difference between our sample of pupils with non-missing values on all covariates and the excluded pupils on observable characteristics. Inclusion of the additional observations hardly affects the estimation results presented in Section 2.6.

Table 2.1 Sample statistics for sample with mavo-advice, early tracked versus later tracked

	Early tracked: mavo	Non-tracked: mavo-havo (-vwo)	*p*-Value
Ability			
Test score arithmetic	10.6	10.6	0.864
Test score language	11.6	11.7	0.620
Test score information processing	11.7	11.7	0.983
Personal and SES variables			
Female	56	55	0.574
Age	12.6	12.6	0.410
Highest education level parents			0.000
No primary education	1	2	
Primary education	12	13	
Secondary education low	30	24	
Secondary education high	41	39	
Higher education first phase	14	17	
Higher education second phase	2	6	
Higher education third phase	0	0	
Profession level parents			0.000
Worker	29	22	
Self-employed without personnel	5	5	
Self-employed with personnel	5	5	
Lower employee	11	12	
Intermediate employee	22	21	
Higher profession	12	16	
Other	17	19	
Ethnicity			0.000
The Netherlands	89	81	
Other (4 categories)	11	19	
Family composition			0.000
Father and mother	87	77	
Other (7 categories)	13	23	
Urbanization city of residence			0.000
Very high	9	18	
High	16	30	
Median	19	13	
Modest	29	23	
Low	26	17	
Educational outcomes			
Ever participated in higher education	33.4	38.3	0.005
Completed higher education	21.3	26.8	0.001
Number of pupils	2905	1031	

Notes. Test scores are of entrance tests taken in the first year of secondary education. The maximum score for each test is 20. All numbers represent percentages unless stated otherwise.

In addition, students in tracked classes are more likely to live in families with two parents and in regions with a lower degree of urbanization. Despite these differences, the personal characteristics of the students in both groups are remarkably comparable. The age and gender composition is equal. Even more importantly, the scores on all three ability tests taken in the first year of secondary education are also equal. A regression of tracking on all covariates yields no significant effects. As such, the sample statistics suggest that the observable characteristics of the groups are quite similar.

The bottom panel of Table 2.1 gives a first impression of the potential effect of early tracking on educational achievements. We observe that both enrolment and completion of higher education are significantly lower for early tracked pupils. The difference in completion of higher education between the two groups is approximately 5 percentage points.

Our data also allow us to observe the educational career of students in the different tracks. This may indicate to what extent the lower participation in higher education among the tracked pupils can be explained by differences in enrolment opportunities after the first years in secondary education. Pupils who started secondary education in a comprehensive class still have an opportunity to enrol into a higher track (i.e. havo or vwo) in subsequent years, which grants admission to higher education. Pupils who start in a categorial class do not have these possibilities. Only in special cases where pupils turn out to have been placed in the wrong track, are pupils allowed to change tracks, in which case they have to repeat one grade. Table 2.2 shows what happened with the students in each group during the first nine years after enrolment in secondary education. We observe that some of the non-tracked pupils move on directly to higher education types, which results in more non-tracked pupils ending up in education levels that provide direct access to higher education. In the second cohort year, for example, participation in havo or vwo is 14 percentage points higher for the non-tracked students. In the third and fourth cohort years, this difference increases to 18 percentage points. Later on, these differences in educational careers translate into differences in higher education enrolment in favour of non-tracked students.

Table 2.2 Routes through education in first nine years after entering secondary education, tracked (=mavo) versus non-tracked (=mavo/havo or mavo/havo/vwo); mavo advice estimation sample

	< Mavo	Mavo	Comprehensive: mh or mhv	Havo	Havo/vwo	Vwo	Mbo	Higher education	Left
Year 1									
Non-tracked	0	0	100	0	0	0	0	0	0
Tracked	0	100	0	0	0	0	0	0	0
Year 2									
Non-tracked	8	42	32	7	9	0	0	0	1
Tracked	3	94	0	1	1	0	0	0	0
Year 3									
Non-tracked	11	58	9	12	5	3	0	0	2
Tracked	7	90	0	2	0	0	0	0	1
Year 4									
Non-tracked	14	61	1	16	1	3	0	0	4
Tracked	9	87	0	2	0	0	0	0	2
Year 6									
Non-tracked	2	3	0	16	0	4	49	3	22
Tracked	2	2	0	11	0	1	64	0	20
Year 9									
Non-tracked	0	0	0	0	0	0	16	27	57
Tracked	0	0	0	0	0	0	16	22	62

Notes. All numbers represent percentages.

2.5 Empirical strategy

We use two approaches for estimating the effect of early tracking on completion of higher education. In our first approach we use linear probability models that include various controls and estimate the following equation:

$$Y_i = \alpha X_i + \beta T_i + \varepsilon_i , \tag{2.1}$$

where Y_i is a dummy variable for completion of higher education, X_i denotes a vector of background characteristics, T_i is a dummy variable which indicates whether a pupil is tracked early or not and ε_i is the error term. T_i takes the value 1 if a pupil starts secondary education in a categorial mavo and takes value 0 if the pupil starts in a comprehensive class. The parameter of interest is β. Estimation of Equation (2.1) by OLS may provide biased and inconsistent estimates if the error term is correlated with tracking. Hence, in case of unobserved heterogeneity we can no longer interpret β as the causal effect of early tracking. We address this problem in several ways. First, by restricting our estimation sample to the group of pupils that leave primary education with a mavo-advice, we select a homogeneous group of pupils with respect to ability, which reduces potential endogeneity problems. Second, we include a large set of individual control variables like personal and socioeconomic background characteristics. In addition, we are able to control for differences in ability by including pupils' test scores.

Nevertheless, it is conceivable that unobservables exist that are correlated both with early tracking and the outcome variable. The differences in socioeconomic background characteristics we observe between tracked and non-tracked pupils (see Table 2.1) may give rise to concerns about unobserved heterogeneity which threatens the unconfoundedness assumption underlying standard OLS regressions. Motivated parents, for example, may rather place their offspring in a comprehensive class, which gives better opportunities to move into a more advanced type of secondary education (havo or vwo) later on.

Our second approach addresses this potential endogeneity problem. We use an instrumental variables (IV) approach that exploits regional variation in the supply of schools. This approach is similar to previous papers in the economic literature that used regional differences in the supply of schools to instrument for educational attainment (Card, 1993; Currie & Moretti, 2003; Park & Kang, 2008).
We use the relative supply of categorial schools in particular municipality types as an instrument for early tracking and estimate Equation (2.1) by two-stage least squares (2SLS). The first stage, in which early tracking is regressed on the supply-ratio and all covariates, is

$$T_i = \gamma X_i + \delta S_i + u_i \, , \qquad (2.2)$$

where S_i denotes the relative supply-ratio of categorial schools in the municipality of residence type of pupil *i* and u_i is the error term. In our data the municipalities of residence of the pupils are classified in 12 categories based on a number of characteristics including the total number of residents and the percentage of the population active in agriculture. For each of these 12 categories the total number of schools offering first grade categorial mavo classes (categorial schools) and the total number of schools offering first grade comprehensive classes (comprehensive schools) is known. From this we calculate a supply-ratio of tracked schools which is defined as the total number of categorial secondary schools divided by the total number of schools in that type of municipality. Four schools offer both comprehensive and categorial classes. They have been counted both as a categorial and a comprehensive school. Table 2.3 provides an overview of the number of categorial and comprehensive schools in our estimation sample for each of the municipality types. The first column reports the twelve different municipality types.

Table 2.3 Supply-ratio of categorial schools

Municipality type	No. of Categorial schools	No. of comprehensive schools	Supply-ratio
Countryside A1	0	0	
Countryside A2	0	0	
Countryside A3	4	1	0.80
Countryside A4	12	0	1.00
Urbanized countryside B1	11	1	0.92
Urbanized countryside B2	29	5	0.85
Specific commuter municipality B3	13	9	0.59
Rural cities C1	7	1	0.88
Small cities C2	17	3	0.85
Medium-sized cities C3	5	4	0.56
Medium-sized cities C4	12	11	0.52
Big cities C5	25	9	0.74

For each of these types, the numbers of categorial and comprehensive schools in our sample is reported in the second and third column, respectively. The last column shows the corresponding supply-ratios. The relative supply-ratios are substantially larger in the countryside, rural and small cities compared to larger cities and specific commuter municipalities. Hence, for pupils living in these first types the choice-set includes relatively fewer comprehensive schools.

The supply-ratio is a legitimate instrument if the exclusion restriction holds. Hence, the effect of early tracking is identified on the assumption that the supply-ratio only affects the outcome variable of interest through early tracking. To further reduce potential endogeneity bias from unobserved regional characteristics that are potentially correlated to both the supply-ratio and outcomes, we include an urbanization indicator in the analysis to control for possible effects of the municipality of residence on the outcome variable. This urbanization indicator is an additional measure for the type of municipality of residence. In contrast to the classification in 12 categories, this indicator is only based on the number of residents.

A concern with using the supply of schools as an instrument is that differences in supply may reflect differences in the demand for schools. In our case differences in supply-ratios may reflect differences in demand for categorial or comprehensive classes. We pursue this issue by investigating the robustness of the results to inclusion of variables that might affect the demand for certain school types (see Section 2.7.2). An additional issue regarding the exclusion restriction concerns potential education quality differences related to the supply of school types. In case of correlation between education quality and our instrument, the identifying assumption underlying instrumental variable analyses would be violated. We discuss this point in section 2.7.2 and empirically address it by investigating the relationship between the supply-ratio and an observable measure of school quality.

Another concern with our instrument is that we measure the supply of categorial and comprehensive schools at an aggregated level. We only have information on the pupil's municipality type, and not specifically of the pupil's own municipality. If there is a lot of variation within municipality types, then our instrument may not reflect the actual options that are truly facing the pupils, resulting in a weak first stage. Furthermore, information at the individual level, that would also include some distance measure (for example the distance between the nearest categorial and the nearest comprehensive class), would be even more preferable, and would probably generate an even stronger first stage. Unfortunately, these individual data are not available. Nevertheless, as the next section will show, our first stage using aggregated information on the supply of schools remains sufficiently strong.

2.6 Main estimation results

This section shows the main estimation results of the two approaches used in this study. The top panel of Table 2.4 shows the estimation results of a linear probability model that regresses completion of higher education on a dummy variable for early tracking for three specifications. Column (1) includes no further controls; column (2) controls for age, age squared, gender, ethnicity, educational and professional level of the parents, family composition and urbanization of residence; and column (3) also controls for test scores at the start of secondary education. The middle and bottom panel of Table 2.4 show the estimation results of the first and second stage of the instrumental variable approach, respectively.

Table 2.4 The impact of early tracking on completion of higher education (OLS and IV estimates)

	(1)	(2)	(3)
OLS			
Early tracking	- 0.054***	- 0.051***	- 0.045**
	(0.019)	(0.018)	(0.018)
IV			
First stage			
Relative supply-ratio	0.776***	0.866***	0.878***
	(0.224)	(0.254)	(0.255)
F-value excluded instrument	12.0	11.6	11.9
Second stage			
Early tracking	0.026	- 0.107	- 0.129*
	(0.070)	(0.067)	(0.068)
Socio-economic status variables (SES)	no	yes	yes
Test scores	no	no	yes
Observations	3936	3936	3936

Notes. Robust standard errors are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level.

Each estimate is based on a separate regression. Robust standard errors corrected for clustering at the school level are in parentheses. The top panel of Table 2.4 shows that early tracking is associated with a reduction in the probability of completing higher education of approximately 5 percentage points. All three OLS-regressions yield a statistically significant negative effect of early tracking on completion of higher education. The estimated coefficients slightly decrease (in absolute value) when additional control variables are included in the model.

A concern with the OLS-estimates is that they might be biased by unobserved factors. Therefore, we also use an instrumental variable approach. The middle panel of Table 2.4 shows the estimation results of the first stage equation (2.2). As expected, we find that an increase in the supply of comprehensive schools increases the probability of enrolment in comprehensive schools. A well-known concern with the IV-approach is the problem of weak instruments (Bound et al., 1995). Staiger and Stock (1997) proposed using a cut-off value of 10 for the F-value of the excluded instrument. In the previous section we noted that our instrument is based on aggregated information and that an individual measure of the supply of schools would be preferable. Nevertheless, the F-values of our excluded instrument are well above the cut-off level, suggesting that the issue of weak instruments is not really a concern for our analysis. Including some smooth polynomial of the supply-ratios or particular step-functions of the supply-ratios yields similar results for the first stage. The bottom panel of Table 2.4 shows the second stage results. The models in column (2) and (3), which include controls for socioeconomic background and ability, yield negative point estimates. In the full model the estimated effect is -0.13, which is statistically significant at the 10% level. The results from these two IV-regressions are larger than the OLS-estimates. This difference in the size of the estimates might be related to the difference in the treated populations used for the estimation. The OLS estimates give the average treatment effect (ATE) for the whole population. The IV-estimates give the local average treatment effect (LATE). This is the effect of the treatment on the subpopulation of compliers (Imbens and Angrist, 1994): in our case the subpopulation that is more likely to choose the comprehensive class if the supply of comprehensive classes increases. The local average treatment effect (LATE) can differ substantially from the average treatment effect (ATE) if the fraction of the population that is affected is small (Oreopoulous, 2006).

Next, we investigate whether the effects of early tracking are different for specific groups, such as girls and boys, high and low ability students and students with different socioeconomic backgrounds. Table 2.5 presents both OLS and IV estimation results of the full model (including all covariates) for various subsamples of pupils. Columns (1) and (2) show the results for the samples of girls and boys, respectively. Columns (3) and (4) show the results for pupils with relatively high and low ability. For this analysis we divided the sample in two based on the total test score ('low ability' is defined as having a score below the median total test score of 34). Finally, columns (5) and (6) show the estimation results for the pupils having parents with a relatively high and a relatively low education level. All pupils with parents having a highest education level above lower secondary education are in the top group; all other pupils are in the bottom group.

Table 2.5 The impact of early tracking on completion of higher education for subsamples

	(1) Girls	(2) Boys	(3) High ability	(4) Low ability	(5) High education level parents	(6) Low education level parents
OLS						
Early tracking	-0.049^{**}	-0.036	-0.076^{***}	-0.010	-0.067^{***}	-0.008
	(0.024)	(0.024)	(0.024)	(0.022)	(0.025)	(0.021)
IV						
Early tracking	-0.115	-0.131	-0.159^{*}	-0.098	-0.163^{*}	-0.102
	(0.094)	(0.090)	(0.086)	(0.097)	(0.095)	(0.093)
SES	yes	yes	yes	yes	yes	yes
Test scores	yes	yes	yes	yes	yes	yes
Observations	2175	1761	2103	1833	2295	1641

Notes. Robust standard errors are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level.

The estimation results in Table 2.5 show no clear difference between boys and girls. While the OLS point estimate is somewhat larger (in absolute value) for girls, the IV estimate is larger for boys. However, for the other subsamples we observe differences in the effect of early tracking. The effects of early tracking seem to be more negative for students with a higher ability and students with higher educated parents. The size of the estimates is larger for both the OLS and IV approach. Additional analyses, in which pupils are divided into two groups based on the professional level of the parents, show a similar pattern (not shown in Table 2.5). These findings suggest that the detrimental effects of early tracking are larger for pupils who are more likely to enrol in and complete higher education.

Summarizing, this section shows that early tracking reduces the probability of graduation from higher education for pupils at the margin of the high and low tracks (the pupils with a mavo advice). Both the findings from the OLS- and IV-regressions suggest that early tracking has a negative effect on the completion of higher education. The detrimental effects are larger for students with a relatively high ability and a higher socioeconomic background.

2.7 Robustness analyses

This section presents estimation results from several alternative model specifications and analyses designed to probe the robustness of our main results. We concentrate on potential bias of the OLS estimations due to unobserved heterogeneity (Section 2.7.1) and concerns regarding the validity of our instrument in the IV approach (Section 2.7.2).

2.7.1 Robustness of OLS analysis

We present robustness analyses in order to explore the magnitude of potential selection bias, which helps us to establish bounds on the true causal effect of early tracking. First we address selection concerns by focusing the analysis on a range of test scores. Excluding pupils with the highest and lowest ability from the analysis may result in a more homogeneous subsample, thereby reducing the risk of unobserved heterogeneity and selection bias in the estimations. Table 2.6 presents the OLS estimation results of the full model (including all covariates) in which the estimation sample is narrowed based on test scores. The first model presents the results for the pupils with a total test score between 29 and 39 (the second and third quartile of the total test score distribution), the second model for the pupils with test scores between 32 and 37 (quantiles 37.5 and 62.5). Narrowing the subsamples with respect to test scores does not decrease the estimated effect (in absolute value) for early tracking. Narrowing to a subsample including only the middle 25 percent of the total test score distribution even increases the estimated coefficient (in absolute value) to -0.070.

Table 2.6 The impact of early tracking on completion of higher education (narrowing subsamples)

	(1)	(2)
	Test scores between 29 and 39	Test scores between 32 and 37
OLS		
Early tracking	-0.046^{*}	-0.070^{**}
	(0.024)	(0.033)
SES	yes	yes
Test score	yes	yes
Observations	2027	1178

Notes. Robust standard errors are in parentheses. * / ** denotes significance at a 10 / 5 % significance level.

The major concern regarding potential selection bias is that pupils with ‘better’ unobservable characteristics might self-select into the comprehensive classes, in which case the main OLS estimates would be biased upwards (in absolute value). As such, we are particularly interested in establishing a lower bound of the effect of early tracking. Our second analysis aims to provide a lower bound by focusing on a set of municipality types for which observable characteristics are ‘better’ for pupils in tracked classes. We use parental education level as the main criterion in the selection of the municipality types, as tracked and non-tracked pupils differ significantly on this observable variable (see Table 2.1). Our data show that this condition, better observable characteristics for tracked pupils, is met for pupils living in municipality types A4, C2, C3 and C5. Table 2.7 presents the covariates on parental education level and total test scores of the pupils. Early tracked pupils have significantly higher educated parents and higher test scores. Other combinations of 4 or more municipality types yield relatively better characteristics for non-tracked pupils. The OLS estimation results of the full model for this subsample are presented in Table 2.8

Table 2.7 Sample statistics for sample with mavo-advice, early tracked versus later tracked; subsample of pupils living in regions A4, C2, C3, C5

	Early tracked: mavo	Non-tracked: mavo-havo (-vwo)	*p*-Value
Ability			
Total test score	33.83	32.92	0.071
Parental education level			
Highest education level parents			0.001
No primary education	2	3	
Primary education	14	22	
Secondary education low	30	25	
Secondary education high	39	34	
Higher education first phase	14	13	
Higher education second phase	2	3	
Higher education third phase	0	0	
Number of pupils	296	1228	

Table 2.8 The impact of early tracking on completion of higher education (subset of regions)

	(1)	(2)	(3)
OLS			
Early tracking	– 0.027	– 0.040	– 0.036
	(0.028)	(0.028)	(0.029)
Socio-economic status variables (SES)	no	yes	yes
Test scores	no	no	yes
Observations	1524	1524	1524

Notes. Robust standard errors are in parentheses.

The estimates show that early tracking decreases the probability of completion of higher education with 3.6 percentage points in the full model. In this analysis tracked pupils have a significantly higher parental education level and higher test scores. Assuming that these 'better' observable characteristics also imply 'better' unobserved characteristics for tracked pupils, the OLS estimation results may be expected to be biased downwards (in absolute value). Hence, the estimation result may be interpreted as the lower bound on the true effect of early tracking.

2.7.2 Validity of the instrument

To address potential endogeneity bias in the OLS estimation, we use an instrumental variable analysis using the relative supply of tracked schools as instrument for early tracking. Identification depends crucially on the assumption that the supply-ratio only affects the outcome through early tracking. If the only impact of the supply-ratio on higher education completion is through early tracking, then the supply-ratio should be statistically insignificant in a regression on higher education completion that also includes a dummy for tracking. As a first test on the validity of the instrument, we include the supply-ratio in the OLS model in column (3) of Table 2.4. This yields an insignificant effect of the supply-ratio.

A concern with using the supply of schools as instrument is that it may be driven by the demand for specific school types. We address it by investigating the robustness of our estimation results to inclusion of variables that may reflect the demand for schools in various municipality types.

The first variable we use is a measure of the change in supply-ratios between 1989 and 1993. Changes over time may reflect differences in demand for specific school types in 1989. We construct a measure of the change in ratios by dividing the supply-ratio in 1993 by the supply-ratio in 1989 for each municipality type. A value larger than 1 implies an increase in the relative supply of categorial schools between 1989 and 1993. A second indicator is the fraction of highly educated parents in a pupil's municipality type. If higher educated parents place more value on educational attainment and rather send their children to a comprehensive class with better opportunities to end up in a higher track later on, the fraction of highly educated parents may reflect the demand for comprehensive schools. For robustness, we use two different definitions for highly educated parents: the fraction of pupils with parental education levels 4, 5 and 6 and the fraction of pupils with parental education levels 5 and 6.

Table 2.9 shows the estimation results of both the OLS and IV models including these variables. The first model is the full model in column 3 of Table 2.4, including a variable indicating the change in the supply ratio between 1989 and 1993. The second and third models include the two variables for the fraction of highly educated parents in the full model.

Table 2.9 The impact of early tracking on completion of higher education (OLS and IV estimates)

	(1)	(2)	(3)
OLS			
Early tracking	- 0.046**	– 0.045**	– 0.046**
	(0.018)	(0.018)	(0.018)
IV			
First stage			
Relative supply-ratio	0.788**	0.812***	0.782***
	(0.317)	(0.251)	(0.273)
F-value excluded instrument	6.2	10.5	8.2
Second stage			
Early tracking	-0.191*	-0.153*	-0.213*
	(0.106)	(0.091)	(0.120)
Change in supply-ratio over time	yes	no	no
Fraction of highly educated 1	no	yes	no
Fraction of highly educated 2	no	no	yes
Socio-economic status variables (SES)	yes	yes	yes
Test scores	yes	yes	yes
Observations	3936	3936	3936

Notes. Robust standard errors are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level.

The OLS estimates are hardly affected by the inclusion of these variables. The IV results are comparable to those in the main estimations. Estimated coefficients are even larger now (in absolute value), with point estimates between -0.15 and -0.21. Models in which we include combinations of the three variables do not affect the results. We conclude that we do not find evidence that the demand for education might bias our findings.

An additional concern regarding the exclusion restriction lies in potential education quality differences across schools or regions. Differences in the supply of school types might be related to the quality of education, and hence to the outcome. We address the issue of potential confounding regional differences in general by including an urbanization indicator in our main regressions. This indicator is statistically insignificant in all of our models. Specifically with respect to school quality, structural variation originating from differences in resources does not seem plausible since the Dutch system provides equal funding to schools. Nevertheless, schools may not be of equal quality because of differences in for example teacher effort or peer quality. In order to empirically address this issue we use average total test score at the school level as an observable measure of school quality and investigate its correlation with our instrument. In a regression of school average total test score on the supply-ratio and all other covariates, we find a statistically insignificant positive effect of the supply-ratio. This suggests that, conditional on all covariates, our instrument is not significantly related to school quality. Moreover, the positive correlation between a larger relative supply of tracked schools and school quality implies that, if anything, unobserved quality differences would bias our estimates downwards (in absolute value). Hence, we do not find evidence that our finding of a negative effect of early tracking is biased by unobserved school quality.

2.8 Effects of early tracking for higher ability pupils

The previous results suggest that pupils with a mavo advice can gain substantially from being in a comprehensive class. It is important to note that this is the result of a partial analysis, which focuses on a single advice group and compares only assignment in a low track to assignment in a comprehensive class. These results are not sufficient to judge the efficiency of an early tracking regime as a whole. After all, higher ability pupils may be negatively affected by the lower ability ones in a comprehensive class. This section investigates the effects of tracking for the pupils with a havo advice, the level just above a mavo advice. Focusing on the effects of tracking for pupils with different abilities contributes to obtaining a more general view on the efficiency effects of early tracking.

We concentrate on the pupils who leave primary education with a havo advice. We use this group to analyse the effect of being in a high track compared to being in a comprehensive class. The non-tracked pupils are again defined as those who enter secondary education in a mavo-havo or mavo-havo-vwo class. The tracked pupils are defined as those who enter secondary education in a havo or havo-vwo class.[5] This is the mirror image of the previous analysis on the group of mavo-advice pupils, which compared assignment in a low track to assignment in a comprehensive class. We estimate similar models as in Table 2.4. Unfortunately, we cannot estimate IV-regressions because there are no severe restrictions in the supply of high track classes. The possible bias due to self-selection into the high track classes will be discussed below.

Table 2.10 presents the OLS estimates of early tracking on completion of higher education. The estimation sample consists of 1162 pupils that entered secondary education in a havo-vwo class and 608 pupils that entered in a mavo-havo or mavo-havo-vwo class.

Table 2.10 The impact of early tracking on completion of higher education (havo advice group)

	(1)	(2)	(3)
OLS			
Early tracking	0.040	0.004	– 0.008
	(0.029)	(0.027)	(0.026)
Socio-economic status variables (SES)	no	yes	yes
Test scores	no	no	yes
Observations	1770	1770	1770

Notes. Robust standard errors are in parentheses.

The results indicate that there is no significant effect of early tracking on completion of higher education. This suggests that pupils with a havo advice do not experience negative effects of being in a comprehensive class together with lower ability pupils. As mentioned before, we cannot use an IV approach to address the potential endogeneity issue. However, it is plausible to assume that potential self-selection would positively affect educational outcomes for the high track, since pupils with better unobservable characteristics are likely to self-select into the high track and to complete higher education

[5] We include havo-vwo in the high track because there are hardly any schools which offer categorial havo classes in the Netherlands. Actually, there turns out to be no school that offers such a categorial havo class in our estimation sample.

more often. Hence, we expect that the estimated coefficients can be interpreted as the upper bounds of the effects of early tracking. As such, endogeneity does not seem to be a problem because all of our estimates, that might be upward biased, are statistically insignificant.

Summarizing, these analyses suggest that pupils with a havo-advice experience no significant difference whether they start secondary education in a high track or in a comprehensive class. Hence, pupils that are advised to follow the high track do not gain from enrolment in a high track compared to enrolment in a comprehensive class. Together with the previous result that pupils advised to enrol in the low track do gain from being in a comprehensive class, this suggests that higher education completion can be improved by tracking all pupils later.

2.9 Conclusions and discussion

This chapter investigates the effect of early tracking on the completion of higher education. In our analysis, we use data from the Secondary Education Pupil Cohorts of 1989, and exploit differences in the timing of tracking between schools in the Netherlands. To deal with potential endogeneity problems we restrict our estimation sample to a particular school advice group which is homogeneous with respect to ability, use a large set of covariates and adopt an instrumental variables approach. Our main analysis focuses on pupils with a mavo advice. If early tracking would have a detrimental effect on participation in higher education, the mavo advice pupils who are tracked early are likely to be the group which is most negatively affected because the pupils in the mavo track have no direct access to higher education.

We find negative effects of early tracking on completion of higher education. The OLS estimates are supported by the IV estimates. The OLS estimates show that early tracking decreases the probability of completion of higher education by approximately 5 percentage points for pupils with a mavo advice. In the 1989 cohort, average completion of higher education for the tracked pupils is around 21%. Hence, pupils with a mavo advice that are in a categorial mavo can increase their probability of completing higher education by around 25% by entering a comprehensive class. The IV analyses yield even larger effects in absolute value. Hence, pupils with a mavo advice perform better in a comprehensive class than in a low track.

This finding for mavo-advice pupils is not sufficient to judge the efficiency of an early tracking regime as a whole. After all, higher ability pupils may be negatively affected by the lower ability ones in a comprehensive class. To obtain a more complete view on the impact of early tracking, we have also analysed effects of early tracking for pupils with a havo advice. Based on OLS estimations we conclude that pupils with a havo advice experience no significant difference whether they are in a high track or in a comprehensive class. Hence, while pupils with a mavo-advice gain from being in a comprehensive class (compared to being in a low track), pupils with a havo-advice experience no significant difference whether they are in a high track or in a comprehensive class. These results suggest an inefficiency of the early tracking regime with respect to completion of higher education.

Our analysis is not informative on the underlying mechanisms that translate early tracking to outcomes. Potential mechanisms might be, for example, peer effects in the classroom or teacher quality effects. Low ability pupils may benefit from the interaction with their high-achieving peers in comprehensive classes. If peer effects are nonlinear, this gives rise to efficiency gains. In the empirical literature, however, there is no consensus yet on the size and functional form of peer group effects and the role of peers in the discussion on early tracking remains disputed.[6] If better teachers prefer to teach relatively high ability classes, teacher sorting may result in a higher quality of education in these classes.[7]

In addition, our empirical results may reflect differences between tracking ages in the ability of educators to allocate students to appropriate school types. The role of these mechanisms remains a valuable topic for further research which may contribute to a more profound understanding of the effects of early tracking.

Our study supports the OECD conclusion regarding the Dutch education system that the early tracking regime has a negative effect on participation in higher education. Pupils with a mavo advice are more likely to complete higher education when they start secondary education in a comprehensive class. As such, we expect that the recent decrease in the number of early tracked pupils in the Netherlands will increase future graduation from higher education. Completion of higher education can be further increased by stimulating participation of pupils with a mavo or havo advice in combined first-grade classes that postpone the point in the educational career at which these children enter different tracks.

[6] Sund (2009) finds empirical evidence for the existence of nonlinear peer effects. Other recent work on peer effects includes Ammermüller and Pischke (2009) and Eisenkopf (2010).

[7] Betts and Shkolnik (2000) find some evidence that schools would be more likely to allocate highly educated teachers to higher ability classes. Rivkin et al. (2005) and, more recently, Hanushek (2011) point out the central importance of teacher quality for student achievement.

3

Can financial incentives for regional education authorities reduce school dropout?[1]

Abstract
This chapter investigates the effect of a new type of financial incentive in education targeted at regional authorities. Previous studies have focused on financial incentives for students, teachers or schools. We identify the effect by exploiting the gradual introduction of a new policy aimed at reducing school dropout in the Netherlands. The introduction of the policy in 14 out of 39 regions and the use of a specific selection rule for the participating regions allow us to estimate local difference-in-differences models. Using administrative data for all Dutch students in the year before and the year after the introduction of the new policy we find no effect of the financial incentive scheme on school dropout. In addition, we find suggestive evidence for manipulation of outcomes in response to the program.

[11] This chapter is based on joint work with Marc van der Steeg and Dinand Webbink: Van Elk, R., M. van der Steeg, D. Webbink, 2013, Can Financial Incentives for Regional Education Authorities Reduce School Dropout?, *De Economist*, 161 (4), 367-398.

3.1 Introduction

The use and effectiveness of financial incentives in the education system is a controversial issue in many countries. Incentive-based policies may provide a cost-effective way to improve educational achievement compared to traditional resource policies. As ‘throwing money at schools’ appears to be no guarantee for improvements in educational quality (Hanushek, 2006; Woessmann, 2003), policymakers and researchers around the world have become increasingly interested in the use of incentives.

In the U.S., the No Child Left Behind Act of 2001 implied a movement towards school accountability in order to improve student performance. The U.S. Ministry of Education recently made available 285 million dollars in the Teacher Incentive Fund grant applications for school districts to support local projects with performance-based pay for teachers. Various countries have introduced policy programs that contain incentive structures for students. An example from the U.S. is the Quantum Opportunities Program (Maxfield et al., 2003), which contains financial incentives based on student’s effort and performance. Other projects include the Education Maintenance Allowance in the UK (Dearden et al., 2009) which offers compensation to low-income families for school attendance and cash payments for particular educational achievements, and the PACES program in Colombia (Angrist et al., 2002) that provided private school vouchers to students which were only extended if pupils performed well.

At the same time, teachers and their unions often strongly oppose the introduction of financial incentive schemes. Skeptics argue that incentives can have unintended and detrimental consequences in the long run as they might weaken intrinsic motivation (e.g. Frey and Oberholzer-Gee, 1997) or encourage educators to focus on the specific rewarded measure at the expense of other relevant tasks or learning objectives (Holmstrom and Milgrom, 1991). In addition, individual teacher incentive programs may negatively affect mutual collaboration (Lazear and Rosen, 1981), while incentive schemes for group performance may lead to free rider behaviour (Lazear, 1999).

The empirical evidence on the effectiveness of incentives in education is ambiguous. Various studies have found that financial incentives for students, teachers or schools noteworthy improved the educational performance of students (e.g. Angrist and Lavy, 2009; Kremer et al., 2009; Lavy, 2002, 2009; Hanushek and Raymond, 2005; Dee and Jacob, 2011). Other studies, also based on credible research designs, did not find an effect of financial incentives on student achievement (e.g. Fryer, 2011; Springer et al., 2010). Moreover, there is evidence that financial incentives induce strategic behaviour of schools by keeping low-performing students from test attendance (Ladd and Walsh, 2002; Jacob, 2005; Figlio and Getzler, 2006) or by teaching-to-the-test (Jacob, 2005; Glewwe et al., 2003). These findings suggest that financial incentives can improve student performance and therefore should be considered as a promising tool for educational policy. In addition, they point out the importance of properly constructed and implemented programs to avoid unintended consequences.

A potentially important feature of incentive programs is the aggregation level at which incentives are targeted. Previous studies have focused on financial incentives for students, teachers or schools. To our knowledge a more centralized incentive scheme in education targeted towards a regional level has not been investigated yet.

This study aims to contribute to existing literature by investigating the effects of a new type of financial incentive targeted at regional education authorities in the Netherlands. These authorities received additional resources if they reduced school dropout compared to a baseline level. The advantage of this type of incentive is that it is allows for an improved coordination and less fragmentation of activities. Whereas students, teachers or schools may have their own – potentially conflicting or short run– objectives, local education authorities have a bird’s eye view on the education system within their district. This enables them to coordinate activities of educators or exploit economies of scale, which may result in a more efficient use of resources, an increased collaboration between schools or combined efforts of schools and other relevant public agencies. An inherent drawback of this type of incentive scheme seems to be that it is only indirectly related to teacher and student behaviour within schools.

For identification of the effect we exploit the gradual introduction of the new policy in 2006 in 14 out of 39 regions based on the number of school dropouts in a baseline year. This introduction strategy and the use of a specific selection rule for the eligibility of regions allow us to estimate local difference-in-differences models. We are able to use administrative data for all Dutch students in the year before and the year after the introduction of the new policy.

Our main finding is that the financial incentive scheme did not have an effect on school dropout. The estimated effects on school dropout are statistically insignificant. These results are robust for a variety of specifications and sensitivity checks. In addition, we find suggestive evidence for strategic behaviour of the treatment regions in response to the incentive program.

The remainder of this chapter is organized as follows. Section 3.2 describes the financial incentive scheme, including details on school dropout in the Dutch education context and the program characteristics. Section 3.3 presents the econometric framework. Section 3.4 describes the data and

Section 3.5 presents the main estimation results. Sensitivity checks are presented in Section 3.6. Section 3.7 concludes.

3.2 The financial incentive scheme

3.2.1 School dropout in the Netherlands

In Dutch education a student is classified as a school dropout in a particular school year (t - (t+1)) when he/she is (i) aged 12 to 22, (ii) registered in education at October first of year t (the start of the school year), (iii) not registered in education at October first of year t+1, (iv) without having completed a certain level of education, called the 'start-qualification'. This start-qualification is considered to be the minimum level of education needed to participate well in the labour market and corresponds to a degree in higher secondary education or intermediate vocational education.[2]

After leaving primary education at the age of 12, pupils can enter three levels of secondary education: pre-vocational secondary education ('*vmbo*'), higher general secondary education ('*havo*') and pre-university education ('*vwo*').[3] In secondary education grade repetition is permitted and pupils are expected to participate at least until age 16 because of compulsory education.[4] Pupils older than 16 are free to leave education.[5] After secondary education students can enrol in intermediate vocational education ('*mbo*') or higher education, depending on completed school type.[6] Pupils who have completed *havo* or *vwo* are qualified for access to higher education, while pupils with a *vmbo* degree can enrol in *mbo*. *Mbo* is oriented towards vocational training and is offered at 4 levels which prepare students for specific professions in the labour market. A start-qualification is defined as being graduated for *havo*, *vwo* or at least the second level of intermediate vocational education.

This definition based on enrolment data thus concerns the 'new dropouts' in a particular school year and does not take into account the stock of previous dropouts that are still aged below 23, without a start-qualification and not in education ('old dropouts'). This is currently considered to be the most important

[2] Unemployment among youths without a start-qualification is more than twice as high as unemployment among youths with a start-qualification (Statistics Netherlands, 2011). In the dropout definition it makes no difference whether a dropout actually has a job or not.

[3] The *'vmbo'* track was introduced in 1999 as a combination of the previous '*lbo*' and '*mavo*' tracks.

[4] Despite this minimum school-leaving age, around 2 percent of the pupils aged 16 or younger turned out to leave education in the school year 2005-2006.

[5] After the intervention year 2006-2007, compulsory education in the Netherlands was extended to youth aged between 16 and 18 without a start-qualification.

[6] Intermediate vocational education is also called upper secondary vocational education. This thesis uses both terms as synonyms to indicate '*mbo*'.

measure of school dropout in the Netherlands.[7] The total stock of dropouts consists of both the 'new dropouts' and the 'old dropouts'.

Concerns on the substantial number of pupils that left education without attaining a start-qualification have led to intensified policy efforts in recent years. Various new policy measures have been introduced and budgets for existing measures have been raised over the last couple of years. Examples of new measures included in the recently introduced national dropout reduction policy agenda are certification courses for 18-23 year olds in which competences gained through work experience are acknowledged, and investments in a better registration system of dropouts for policy evaluation and adjustment (Ministry of Education, 2008a).

3.2.2 Program details

A key component of the recent Dutch dropout reduction policy is the introduction of a financial incentive program for regional education authorities in 2006. In 1994, the government specified 39 school districts and raised local education authorities, called 'RMC-regions', to monitor, register and fight early school-leaving regionally.[8] Each district has its own regional education authority, which is part of a municipality within the district. The RMC regions have two main legal tasks. First, they have to monitor and register school dropout. Second, they have to support pupils at the risk of early school-leaving in order to prevent them from dropping out and help dropouts by guidance or counselling them back to school or work.

The regional education authorities coordinate anti-dropout activities within their districts and are responsible for the creation of local networks in which schools, school attendance officers and various regional youth assistance bodies work together. The authorities are autonomous in their choice of specific policy measures to prevent early school-leaving. They can for example choose to intensify the care or individual guidance for potential dropouts, improve registration systems of school dropout or school absenteeism, or enhance regional cooperation between schools.

The local education authorities receive funding from the central government to carry out their main tasks. In the school year 2006-2007, the direct central government funding amounted to 17.5 million Euros. On

[7] This differs from the measure of school dropout in the targets of the European Union (EU), which is the share of students aged between 18 and 24 with only lower secondary education at best and not in education or training. Hence, the most important difference compared to the EU measure is the age criterium. The required education level is well comparable in both definitions. Furthermore, the EU measure is in terms of a total stock of school dropouts rather than the number of new dropouts in a particular year.

[8] In Dutch 'RMC' is an abbreviation of 'regional report and coordination functions for early school-leaving'.

average, this corresponds to a little less than half a million Euros per RMC region. In addition to (direct) central government funding, RMC regions also use other municipal funds to carry out their tasks. These ‘own’ municipal financial means amounted to 21.5 million Euros in 2006-2007 (Research voor Beleid, 2008). Together, the average yearly amount of money available to RMC’s per new school dropout is a little less than 750 Euros.

The Dutch Ministry of Education gradually introduced a financial incentive program for the local education authorities in the summer of 2006. In the first year, covenants that included a financial incentive to reduce school dropout were offered to 14 out of 39 school districts with the highest number of dropouts.[9] For each reduced early school-leaver the RMC regions could earn 2,000 Euros. The total budget of the covenant program amounted to 16 million Euros. This is roughly twice the amount of the direct contribution from the central government to the covenant regions for the school year 2006-2007. On a per-dropout basis the offered 2,000 Euros is more than 2.5 times the amount of money from direct government funding and municipal means. Hence the program offers a substantial financial reward compared to regular funding. The financial reward is meant for an improvement of school dropout policies. Within this general purpose, the RMC districts are free to decide on the specific allocation of the additional resources.

Eligibility

Eligibility for the program was based on the absolute number of dropouts in the year 2004-2005. More specifically, selection of the regions was based on presence in either the top 10 ranking of the number of ‘new dropouts’ (pupils that drop out during the school year) or the top 10 ranking of the total number of dropouts (which include both new dropouts and the stock of previous dropouts that are still aged below 23, without a start-qualification and not in education) in the reference year 2004-2005. This reference year was chosen because the figures for the pre-treatment year 2005-2006 were not available at the time the covenants were signed. This resulted in a list of 12 RMC regions. Two other regions, which wanted to join the covenant program and were just outside the top ten lists, were added later on own request. These are ‘Centraal en Westelijk Groningen’ and ‘Zuidoost-Brabant’. The other 25 regions have not been

[9] Although in most districts there were schools that also signed the covenant, the financial incentive is exclusively meant for the regional education authority. Hence, only the authority receives a payment from the government in case a reduction in school dropout is realized.

selected and are not treated with a financial incentive program.[10] The complete rankings and selected RMC districts can be found in Appendix 3.A.

Financial incentive

The covenants provide financial incentives and are based on a ‘no-cure no-pay’-principle. For each reduced dropout in 2006-2007, relative to 2004-2005, the local education authority receives 2,000 Euros.[11] It does not matter how the reduction is achieved. This payment scheme stops if a reduction of 10% in the number of school dropouts is realized. In this way the Ministry has limited the maximum financial reward for a region.

Formally, the financial reward scheme can be defined as follows:

$$P_j = \min\{(2{,}000 * \max\{(D2004_j - D2006_j), 0\}) , D2004_j * 0.10 * 2{,}000 \},$$

with:

P_j = reward from the Ministry to RMC region j in euros,

$D2004_j$ = total number of dropouts in region j in 2004-2005,

$D2006_j$ = total number of dropouts in region j in 2006-2007.

Types of measures

RMC regions were free to choose which instruments to use to reduce the total number of dropouts. The Ministry presented a menu of promising options, but it did not impose any specific actions the regions should undertake.

The projects which RMC’s were planning to carry out are explicitly mentioned in the individual covenants. The majority of the measures agreed upon have a preventive rather than a curative character.[12] Examples of frequently undertaken preventive measures are projects to promote a good transition from vmbo to mbo, the use of so-called ‘care-advise-teams’ (teams of different actors offering care and guidance at school to students at risk of dropping out), and projects promoting the number and choice

[10] After the first year the government scaled up the program and offered new covenants to all 39 regions in late 2007 and early 2008. These new covenants are four-year instead of one-year arrangements and differ with respect to the design from the 2006 covenants which we evaluate in this study.

[11] An inherent drawback of the choice to use 2004-2005 as the reference year is that fluctuations in the dropout figures in 2005-2006 relative to 2004-2005 already affected outcomes of the covenants. We come back to this issue in Section 3.6.

[12] The focus on preventive actions is in line with the economic literature which suggests that anti-dropout interventions targeted at students-at-risk who are still in education are much more effective than curative interventions targeted at students who already have dropped out (Heckman, 2000).

process of apprenticeships. Some regions have also devoted more attention to an active approach of pupils, e.g. by providing information on school choice, checking enrolment in the beginning of the new school year and visiting pupils that are not enrolled.[13]

3.3 Empirical strategy

This section discusses the empirical strategy to identify the effectiveness of the covenant policy. Since assignment of regions to the program was not random, identification of the treatment effect is the central issue.

3.3.1 Difference-in-differences

In order to assess the impact of the covenant policy on school dropout, we start with a difference-in-differences (DID) estimation approach on the full sample of all 39 regions. This approach exploits the availability of data for the full population in the pre-treatment (2005-2006) as well as the post-treatment year (2006-07) and basically compares changes in dropout rates over time between covenant and non-covenant regions.

We implement this strategy by estimating the following linear probability model[14]:

$$Y_{ijt} = \beta_0 + \beta_1 C_{ij} + \beta_2 T_t + \beta_3 C_{ij}{*}T + \beta_4 X_{ijt} + \alpha_j + \varepsilon_{ijt} \ , \tag{3.1}$$

with Y_{ijt} being a dummy variable indicating whether pupil i in region j in year t is a dropout; C_{ij} being a dummy variable which takes value 1 if RMC region j belongs to the 14 covenant regions, and 0 if it belongs to one of the remaining 25 regions. T is a time dummy variable taking value 1 in 2006-2007 (post-treatment year) and value 0 in 2005-2006 (pre-treatment year). X_{ijt} is a vector of background characteristics of pupil i in region j in year t, α_j is a region fixed effect and ε_{ijt} is the error term representing all unobservables of pupil i in region j in year t.

For each region j and year t, we include all relevant pupils aged 12 to 22 that are registered in education and do not have a start-qualification at the beginning of the year. The coefficient β_3 then gives the treatment effect of interest.

[13] We refer to Ministry of Education (2008b) for a complete list of measures in the different covenant regions.

[14] We find similar results when using a probit model.

The difference-in-differences framework allows us to overcome many of the threats to identification. We control for all time-invariant (unobserved) heterogeneity across RMC districts and are able to control for observable changes in student composition by including student individual background characteristics. The identifying assumption underlying the difference-in-differences estimation approach is that treatment and non-treatment regions have a common trend, which implies that the paths of outcomes for both groups would not be systematically different in the absence of the program.

The common-trend assumption, which rules out group specific trends and composition effects, is not testable. A well-known informal technique to provide insight to its validity is to test for equality of pre-treatment trends between experimental and control groups. However, since the BRON data we use for our analysis are only available from the year 2005-2006 onwards (see Section 3.4), we are not able to investigate this further.

Composition effects, which might threaten the common trend assumption, do not seem plausible in our application. It does not seem realistic that the policy introduction affected the composition of students in the treatment and control districts. Moving to another region or choosing a school outside the region because of the introduction of the program seems unlikely, especially since treatment regions may not be the preferred regions for self-selection as they typically face high levels of school dropout. This conjecture is supported by an investigation of the development of the absolute number of registered pupils over time for all treatment and control regions. The cohort of relevant students in the post-intervention year is 1% larger than the pre-intervention year, and both the treatment and non-treatment regions faced an equal increase of 1%. Furthermore we do not find any significant differences in the development of observable student composition characteristics over time between treatment and control regions.

3.3.2 Local difference-in-differences

Although the use of a difference-in-differences framework allows to control for fixed unobservables, there might still be concerns regarding potential unobservable differences between treatment and control regions that violate the common trend assumption, in which case estimation results would be biased and inconsistent.

To further contribute to the credibility of the identifying assumption we exploit the assignment rule for the treatment. Selection into the program depended in a deterministic way on the number of total dropouts and the number of new dropouts in the school year 2004-2005. All districts that were ranked in the top 10 of either the number of total dropouts or the number of new dropouts in the school year 2004-2005 received treatment, while all other districts did not receive treatment. Hence, eligibility for the program can be described as a discontinuous function of these observed underlying variables. This allows us to identify regions that were close to being selected and regions that were close to not being selected. This type of information is usually exploited in a regression-discontinuity framework.[15] In our approach we exploit the information about the eligibility for the program to improve our estimates in a local difference-in-differences framework. Districts around the threshold values are expected to be more similar in outcomes and characteristics. We construct samples of regions around the cut-off value for eligibility in the program and estimate difference-in-differences models for these samples. We expect that the common trend assumption is more likely to hold in these local difference-in-differences models. [16]

Construction of the discontinuity samples is done by exploiting the selection rule for treatment. After ranking all regions according to the number of total dropouts and new dropouts in the reference year, we can determine the 'just-selected' (treatment) regions by looking at the covenant regions that would not have been selected if e.g. the top 5 (instead of the top-10) of both lists would have been selected for treatment. Similarly, we can determine the 'just-not-selected' (non-treatment) regions, by looking which non-covenant regions would have been selected if e.g. the top 20 of both lists would have been selected for treatment. Adding these regions just below and just above the threshold together yields the discontinuity sample.

The procedure can be formalized as follows. Let $x(k, l)$ denote the region x that is in k-th position in the first ranking and in l-th position in the second ranking (k=1,...,39 and l=1,...,39). The set of treatment regions, denoted by T, is then given by:

T = { $x(k,l)$: $k \leq 10$ or $l \leq 10$ }.

Let DS (T) denote the set of treatment regions that is 'just selected' and let DS (NT) denote the set of non-treatment regions that is 'just not' selected. Then:

DS (T) = { $x \in$ T: $k > 5$ and $l > 5$ } and
DS (NT) = { $x \notin$ T: $k < 20$ or $l < 20$ }.

[15] For a recent review of the use of regression discontinuity designs in empirical economics see Van der Klauw (2008). Angrist and Lavy (1999) and Jacob and Lefgren (2004) are examples in the field of education.

[16] A similar approach is used in Leuven et al. (2007), who evaluate the effect of extra funding for disadvantaged pupils on achievement.

The set of regions in our discontinuity sample, consisting of both treatment and non-treatment regions, is then defined by:

$DS = DS\ (T) \cup DS\ (NT)$.

The above formalization of the construction of the discontinuity sample refers to hypothetical top-5 and top-20 selection criteria for treatment. The choice of bandwidth around the cut-off is arbitrary and related to a trade-off: increasing the sample increases the number of observations which gives more power to our estimations, while it increases the risk of violation of the identifying assumption. We also construct a second, smaller discontinuity sample of just-selected treatment and just-not selected non-treatment regions making use of hypothetical top-7 and top-17 criteria.

In addition to these discontinuity samples we construct two subsamples based on the (ranking of) average dropout probabilities in the pre-treatment year 2005-2006. This is done by including only those regions remaining after having removed the 10 (or 13) regions with the highest and the 10 (or 13) regions with the lowest pre-treatment dropout probabilities. We present estimations on these ‘matched samples’ as an alternative for the ‘discontinuity samples’ to improve on the treatment-control balance.

Standard applications of regression discontinuity designs typically include polynomials of the variables determining eligibility to control for continuous effects of the underlying variables. In our difference-in-differences framework we will both use specifications including fixed region effects and specifications that include polynomials of the total number of dropouts in the school year 2004-2005.

Two school districts, which were initially not eligible for the program, received treatment later in the year on own request. This makes the regression discontinuity design ‘fuzzy’ and may compromise estimation results. To address self-selection of these noncompliant regions we also present 2SLS estimates in which actual treatment status is instrumented by eligibility.

3.3.3 Standard errors and inference

Since our treatment variable does not vary within school districts, we have to be careful with inference. As Moulton (1986) shows, neglecting the presence of common group errors may substantially bias standard errors downwards. In all our estimation results we will therefore present (heteroskedasticity robust) standard errors corrected for clustering at the school district level. Still, this might not be a complete response, given the limited number of school districts in our sample. Donald and Lang (2007), who investigate the inference properties of difference-in-differences estimators, point out that *t*-statistics are asymptotically normally distributed only if the number of clusters goes to infinity.[17] Neglecting that standard asymptotics cannot be applied when the number of clusters is small, may largely overstate significance of the statistics.[18] Although we appear to have a reasonably large number of 39 clusters, we will view standard errors with caution. We address this issue in the interpretation of our estimation results, especially in case of the local difference-in-differences estimations in which only a subset of regions is taken into account.

3.4 Data

Until recently, the main sources for monitoring student drop-out in Dutch education were the data provided by the RMC regions. However, those data turned out to suffer from serious shortcomings with respect to consistency and uniformity (Deloitte, 2006; Sardes, 2006). As a response to the concerns regarding the reliability of the RMC-data the government invested in a better registration system of dropouts. This has resulted in new administrative data based on individual education numbers that have become available from the school year 2005-2006 onwards. These so-called 'BRON data' are currently used by the Ministry of Education to monitor school dropout and analyse policy interventions. In addition to these BRON data, RMC figures are still produced by the RMC districts.

Since these improved data on school dropout have not yet become available at the time of the introduction of the covenant policies in the summer of 2006, the Ministry still had to rely on the RMC data for determining the selection of regions based on the number of dropouts in 2004-2005. The use of these figures for the selection of eligible regions also implied that RMC data from 2006-2007 have been used by the Ministry for measuring performance in terms of reduced school dropout between the reference year 2004-2005 and 2006-2007.

Because of the problems with the RMC data, we only make use of the BRON data to evaluate the effectiveness of the incentive program. These administrative data are based on yearly enrolment figures of

[17] Other recent work on inference properties in difference-in-differences is Bertrand et al. (2004) who study implications of serial correlation.

[18] Donald and Lang (2007) suggest estimation using group means in case of a small number of clusters and propose to use *t*-distributions for inference rather than the standard normal distribution.

schools which are checked by accountants.[19] The accountants' checks minimize the risk of administrative errors which contributes to the accuracy of the data we use.

The dataset covers the whole student population and contains information on the year-to-year progress of each individual student throughout education. Dropout is measured consistently and uniformly in BRON, based on administrative enrolment figures at the first of October in each year. For instance, if a student was present in education at October first of 2006, but not one year later, while not having attained a start qualification in the meantime, this student is marked as a school dropout for the school year 2006-2007. Since registration occurs at one moment in a year, we cannot distinguish between pupils who are dropped out during the school year (t - (t+1)) and returned in education before the first of October of year t+1, and pupils who did not drop out at all during the school year (t - (t+1)). Stated differently, the results of preventive actions by the local education authorities that keep pupils from dropping out cannot be distinguished from the results of curative actions that bring dropouts during the school year back in education. Hence, using BRON data on dropout as outcome variable in our analyses measures the effect of the incentives on both prevention of new dropouts and curative actions focused on those students that dropped out in the school year under consideration. As such, the evaluation using BRON sheds light on the total performance of the local education authorities in reducing school dropout.

In addition to yearly administrative information on school enrolment, the BRON data contain a rich set of individual background characteristics for the whole relevant student population. First, it contains information on personal characteristics like age, gender and ethnicity (7 categories). We use information on age to construct a dummy variable that indicates whether a student is subject to compulsory schooling (which was the case if students were aged 16 or younger in 2006-2007). Second, it provides information on education level and grade. In addition it contains neighbourhood characteristics such as the size of the municipality (in three categories: four largest cities, medium-sized and small-sized), and a dummy variable indicating whether a student is inhabitant of a so-called poverty accumulation area. Poverty accumulation areas are postal code areas which are characterized by an accumulation of social problems. Underlying indicators are the percentage of low incomes, the share of welfare recipients and the share of non-western foreigners. We have completed our data set by merging these data with figures from the Ministry of Education on the central government contributions (per student) to the RMC regions in pre- and post-treatment year.

[19] This is because the central government yearly contributions to schools depend on the number of enrolled students to a large extent.

We use individual data on dropout for both the pre-treatment year 2005-2006 and the post-treatment year 2006-2007. The relevant education levels for dropout registration are secondary education and intermediate vocational education.[20] Therefore, our total estimation sample consists of students aged 12 to 22 that were enrolled in secondary education or in intermediate vocational education at the start of the school years 2005-2006 and 2006-2007. This concerns 1,286,173 students in 2005-2006 (of which 523,762 are in a non-covenant region and 762,411 are in a covenant region) and 1,301,423 students in 2006-2007 (of which 530,015 are in a non-covenant region and 771,408 are in a covenant region). The total sample contains 2,587,596 students.

Table 3.1 reports descriptive statistics for all 14 covenant and all 25 non-covenant regions in the pre-treatment year 2005-2006. Both the outcome variable (dropout probability) and the covariates used in our analyses are presented. In the evaluation these variables can be used to control for compositional changes in the student population. The first column reports the sample means for the group of students in the non-covenant regions and the second column reports them for the group of students in the covenant regions. The third column reports the *p*-value of the difference, calculated using a two-tailed *t*-test or a chi-squared test. As expected, a comparison of the covariates of the two groups of students reveals some differences. The selection criterion based on total number of dropouts implies that regions which contain the largest cities have been selected for the treatment while smaller cities were not selected.

Differences between treatment and control regions are therefore likely to reflect differences in characteristics between inhabitants of more and less urbanized municipalities. Larger cities in the Netherlands for example typically face a higher proportion of disadvantaged students. Table 3.1 confirms this. The average dropout probability in treatment regions was significantly larger (4.6 versus 3.9 percent). In addition, differences in ethnicity, urbanization degree and poverty accumulation area are observed. Both groups seem reasonably comparable with respect to age, gender and education level.

In order to obtain more comparable treatment and control groups we construct four smaller estimation samples (see Section 3.3). The first two are discontinuity samples which exploit the selection rule of the covenants, whereas the latter two are matched samples that consist of covenant and non-covenant regions within smaller intervals of pre-treatment year dropout probabilities.

[20] Within these levels we distinguish between 10 categories.

Table 3.1 Sample means for covenant and non-covenant regions, school year 2005-2006 (pre-treatment year)

	Non-covenant regions	Covenant regions	*p*-Value
Dropout probability	3.88	4.64	0.000
Personal characteristics			
Gender	0.51	0.51	0.030
Age	15.31	15.33	0.000
Age: compulsory education	0.70	0.70	0.305
Ethnicity			0.000
Dutch	0.86	0.73	
Surinam	0.01	0.04	
Aruba/The Antilles	0.01	0.01	
Turkey	0.02	0.04	
Morocco	0.02	0.04	
Other foreign (non-western)	0.04	0.05	
Other foreign (western)	0.05	0.07	
Education level			
Level			0.000
first grade secondary education	0.30	0.31	
vmbo (level 1)	0.05	0.05	
vmbo (level 2)	0.05	0.05	
vmbo (level 3/4)	0.08	0.08	
havo	0.11	0.10	
vwo	0.11	0.12	
mbo-1	0.01	0.01	
mbo-2	0.07	0.08	
mbo-3	0.07	0.07	
mbo-4	0.14	0.13	
Exam class	0.14	0.14	0.089
Environment of the pupil			
Degree of urbanization			0.000
Inhabitant of G4 (4 largest cities)	0.00	0.18	
Medium-sized municipality	0.21	0.26	
Small municipality	0.79	0.57	
Inhabitant poverty accumulation area	0.05	0.20	0.000
RMC budget per student (€)	14.91	9.36	0.000
Total number of pupils	523,762	762,411	

Tables of descriptive statistics for each of the four subsamples can be found in Appendix 3.B. Although still statistically significant in most cases, differences in socio-economic characteristics (e.g. ethnic distribution, share of inhabitants of poverty accumulation area) between the covenant and non-covenant regions become much smaller. Clearly experimental and control groups are more comparable on observable covariates in the four subsamples. In addition, the pre-treatment dropout probabilities are closer to each other in the subsamples.[21]

In the discontinuity samples, we observe that pre-treatment dropout probabilities are smaller in the covenant regions. This indicates that some large regions have been selected for treatment, while smaller regions with larger dropout probabilities have not.[22]

A regression of dropout probability in the pre-treatment year 2005-2006 (based on BRON) on a constant and the absolute number of dropouts in the year 2004-2005 (based on RMC) at the region level yields an insignificant coefficient for the absolute number of dropouts (with a *t*-value of 0.65) in the first discontinuity sample. The correlation between the two variables is 0.17. A similar regression for the second discontinuity sample yields a *t*-value of -0.51 and a correlation of -0.18. This implies that within the discontinuity samples there seems to be no systematic association between the selection rule based on absolute number of dropouts according to the RMC-data and the pre-treatment dropout percentage in the regions. This supports the notion that within these subsamples regions are more or less 'randomly' assigned to the incentive program. We conclude that the discontinuity samples improve the treatment-control balance.

Table 3.2 shows the development of the average dropout probabilities between the pre- and post-covenant school year. Average dropout probabilities in the pre-treatment year equal 4.6 percent in the covenant regions and 3.9 percent in the non-covenant regions. Note that these probabilities concern the total sample including all pupils aged 12 to 22.[23] We observe a decline in the average dropout probability of 0.16 percentage point in the 14 covenant regions. The 25 non-covenant regions witnessed a decline in the dropout probability as well, though somewhat smaller, of 0.13 percentage point. The average fall in the dropout probability is therefore 0.03 percentage points larger in the covenant regions. In the next section we further investigate the effects of the covenants in a difference-in-differences approach.

[21] In the matched samples comparability of dropout probabilities is a consequence of the construction method.

[22] Hence, the eligibility rule selects larger regions into treatment rather than the 12 worst-performing regions. We address potential implications of this eligibility rule on the outcomes in Section 3.6.

[23] Among the pupils aged above 16, for example, average dropout percentages are 9.9 and 8.4 percent in the covenant and non-covenant regions, respectively.

Table 3.2 Development of school dropout percentages between pre-treatment (2005-2006) and post-treatment year (2006-2007), covenant versus non-covenant regions, complete sample

	2005-2006	2006-2007	Δ 2006/07 - 2005/06 (%-point)
Covenant regions (14)	4.64	4.48	– 0.16
Non-covenant regions (25)	3.88	3.75	– 0.13
Difference			-0.03

Notes. Source: BRON data.

3.5 Estimation results

3.5.1 Main findings

This section presents and discusses our main estimation results. Table 3.3 shows the difference-in-differences estimates of the effect of the covenants on the probability of school dropout using six different specifications. The first specification does not include control variables; the second specification only includes the school district fixed effects. In the third model personal characteristics, such as age, age squared, a dummy variable indicating whether a student is subject to compulsory schooling, gender, and ethnicity are included. The fourth model additionally controls for education level and a dummy variable for being in an exam class. The fifth model also includes environmental variables which are municipality size, a dummy variable indicating whether a student is inhabitant of a poverty accumulation area, and the regular central government contribution per student to its RMC district. In the sixth model the region fixed effects are replaced by a third order polynomial of the absolute number of dropouts according to the RMC-data in the base year 2004-2005. All models are estimated by OLS and estimated coefficients are reported in terms of percentage points. Standard errors corrected for clustering at the region level are in parentheses.

In all specifications, the estimates of the effect of the covenants do not significantly differ from zero. Adding additional covariates increases the estimated coefficients (in absolute value), but all estimates remain statistically insignificant. Note that the large estimation sample including all students contributes to the precision of our estimates and allows us to identify even small effects. The point estimate of the full model including all covariates and fixed district effects is -0.146, which suggests that the financial incentive decreases dropout probability with 0.146 of one percentage point. Models including second or higher order polynomials yield similar results.[24]

Table 3.3 OLS Difference-in-differences estimates of the effect of the covenants on the probability of dropping out, estimates in percentage points

	(1)	(2)	(3)	(4)	(5)	(6)
Covenant	-0.029 (0.097)	-0.028 (0.097)	-0.048 (0.091)	-0.058 (0.094)	-0.146 (0.103)	-0.102 (0.096)
Control Variables						
Personal characteristics	no	no	yes	yes	yes	yes
Education level	no	no	no	yes	yes	yes
Environment	no	no	no	no	yes	yes
Fixed effects	no	yes	yes	yes	yes	no
Polynomial of dropouts	no	no	no	no	no	yes
Observations	2,587,596	2,587,596	2,587,596	2,587,596	2,587,596	2,587,596

Notes. Robust standard errors are in parentheses.

Since two RMC districts which did not initially satisfy the eligibility criteria eventually received voluntary treatment, we proceed with an IV approach to address potential self-selection problems. These two regions contain 157,438 pupils, which is 6.1% of the total population. We estimate models with actual treatment status instrumented by initial eligibility for the program. In these analyses we cannot include region fixed effects (or a smooth function of dropouts), because of collinearity in the first stage regression. Table 3.4 presents the 2SLS estimates of the effect of the covenant. These estimates are

[24] Since we find insignificant effects, the proposed analyses by Donald and Lang (2007) using group means do not seem necessary. For completeness, Table 3.C.1 in Appendix 3.C presents the estimation results of an analysis at the school district level using group means, including 78 observations. We find statistically insignificant effects. The positive point estimates may be explained by the better performance of larger districts (see Table 3.7), which are less heavily weighted now.

similar to the OLS estimates. We find statistically insignificant effects of the covenant on school dropout. [25]

Table 3.4 2SLS estimates of the effect of the covenants on the probability of dropping out, estimates in percentage points

	(1)	(2)	(3)	(4)
Covenant	-0.032	-0.058	-0.060	-0.088
	(0.114)	(0.101)	(0.110)	(0.112)
Control Variables				
Personal characteristics	no	yes	yes	yes
Education level	no	no	yes	yes
Environment	no	no	no	yes
Observations	2,587,596	2,587,596	2,587,596	2,587,596

Notes. Robust standard errors are in parentheses.

A potential concern with the difference-in-differences estimations on the full sample is that covenant and non-covenant regions may differ on (unobservable) characteristics that break down the identifying common trend assumption. We therefore perform additional analyses in which we restrict our estimation sample to covenant and non-covenant regions which are more similar to each other. We have constructed four subsamples. The first two are discontinuity samples identified on the basis of the selection rule of the treatment regions; the latter two are matched samples which are constructed by selecting regions within a similar interval of dropout probabilities in the pre-treatment year (see Section 3.3).

Since the self-selection of the two regions breaks down the sharp regression discontinuity, we present IV models in which the treatment is instrumented by initial eligibility.

[25] As an alternative way to address the potential self-selection issue, we also performed analyses on a restricted estimation sample, in which the two voluntary treated districts are left out. Table 3.C.2 in Appendix 3.C reports the estimated effects on the limited sample of 37 school districts. The results are very similar to our main impact findings in Table 3.3.

Table 3.5 reports the full model estimation results including all covariates (except the fixed district effects or smooth function of dropouts) for the four subsamples.[26] For each subsample the number of included covenant and non-covenant regions is reported.

Table 3.5 2SLS estimates of the effect of the covenants on the probability of dropping out, discontinuity samples (DS) and matched samples (MS), estimates in percentage points

	(DS 1)	(DS 2)	(MS 1)	(MS 2)
Covenant	0.021	-0.102	0.022	0.078
	(0.140)	(0.159)	(0.133)	(0.184)
Control Variables				
Personal characteristics	yes	yes	yes	yes
Education level	yes	yes	yes	yes
Environment	yes	yes	yes	yes
Number of covenant districts	8	5	8	5
Number of non-covenant districts	8	5	11	8
Observations	1,143,050	690,111	1,256,966	816,546

Notes. Robust standard errors are in parentheses.

The 2SLS estimates yield larger point estimates compared to the full sample estimates, except for the estimate in the second discontinuity sample. In the other three subsamples points estimates are even positive. All estimated coefficients remain statistically insignificant. These results confirm previous findings.

Especially in our local difference-in-differences estimations including a relatively small number of districts, we have to be cautious with inference. As Donald and Lang (2007) point out, precision may be overestimated in case of few clusters. We do not find a statistically significant effect in our specifications,

[26] Table 3.C.3 in Appendix 3.C additionally presents the full model OLS estimates on the four subsamples for both models including fixed district effects and models including a third order polynomial of the absolute number of dropouts in the reference year. All OLS estimates are statistically insignificant. Point estimates are negative and smaller compared to the corresponding estimates on the full sample (except for the full model estimate with fixed effects in the second discontinuity sample).

which makes that we can ‘safely’ conclude that we find no evidence that the program has been effective.[27]

As an alternative robustness check, we have also estimated our main model specifications on a subsample which excludes the four largest cities in the Netherlands. These large urban cities, which are only included in the treatment group, might face a different nature of dropout problems. The estimation results are presented in Table 3.C.4 in Appendix 3.C. We find very similar results, which imply that the inclusion of the large cities in the treatment group does not importantly affect our results.

3.5.2 Heterogeneous treatment effects

It is conceivable that the covenants had no overall effect on school dropout for the whole student population, but still had a partial effect on a specific subpopulation of students. This might for example be due to a particular focus on certain education levels in the covenant projects or because of a different response of subgroups to the undertaken activities. In order to test for the occurrence of heterogeneous treatment effects, we have estimated the difference-in-differences effects on various subsamples of the student population. We estimated separate regressions for boys and girls, for students in secondary education and in intermediate vocational education, for students aged above 16 and students aged 16 or younger, for students in an exam class and students which are not in an exam class, and for Dutch students and foreign students. Table 3.6 presents the OLS estimation results for the full model specifications with fixed effects in the first row and the full model instrumental variables estimates in the second row.

While OLS estimation points to a larger effect for boys (in absolute value), the 2SLS estimations yields a larger point estimate for girls.

With respect to the education level we find a larger point estimate (in absolute value) in the OLS analysis for mbo students, while 2SLS estimates yield similar effects for both groups. In all of these specifications, the estimated effects are statistically insignificant. In addition, a separate regression only including mbo level 1 and level 2 students (those with the highest dropout probabilities) yields no significant results (not in the table).

[27] Since we find insignificant effects, we do not proceed with the proposed alternative analyses by Donald and Lang (2007).

Table 3.6 Heterogeneous effects: difference-in-differences estimates of the effect of the covenants on the probability of dropping out, estimates in percentage points

	Gender		Education level		Age		Exam class		Ethnicity	
	(1) boys	(2) girls	(1) secondary education	(2) mbo	(1) age>16	(2) age<=16	(1) exam year	(2) no exam year	(1) Dutch	(2) foreign
OLS	-0.191 (0.130)	-0.095 (0.121)	-0.101 (0.066)	-0.238 (0.232)	-0.281 (0.252)	-0.076 (0.070)	-0.250 (0.215)	-0.131 (0.098)	-0.142* (0.078)	-0.043 (0.258)
2SLS	-0.049 (0.154)	-0.128 (0.114)	-0.089 (0.068)	-0.087 (0.261)	-0.043 (0.260)	-0.104 (0.076)	-0.463* (0.257)	-0.026 (0.101)	-0.065 (0.075)	-0.040 (0.303)
Control Variables										
Personal characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Education level	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Environment	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	1,316,687	1,270,909	1,818,136	769,460	790,942	1,796,654	377,008	2,210,588	2,029,238	558,358

Notes. Robust standard errors are in parentheses. * denotes significance at a 10 % significance level.

The OLS analyses yield a larger estimated coefficient (in absolute value) for pupils aged above the compulsory schooling age of 16, while IV analyses yield a larger estimated coefficient for younger pupils. None of these estimated coefficients is statistically significant.

Point estimates are larger for pupils in exam years, with a 2SLS estimate that is statistically significant at a 10% level. This might provide some evidence that there has been a beneficial effect of the covenant on dropout behaviour of pupils in exam classes of secondary education. Nevertheless, we have to be careful being decisive on this regarding the risk of oversignificance of the estimates. Further inspection of the results for exam classes by estimating separate regressions on exam classes in vmbo -where dropout rates are inherently higher compared to havo or vwo exam classes- yield insignificant results (not in the table). In addition, a full sample OLS analysis including an interaction term between a dummy variable for being in an exam class and the variable of interest yields a highly insignificant estimated coefficient for the interaction term. In sum, these results provide little evidence that the covenants have been successful in reducing school dropout in exam classes. The negative OLS estimate for Dutch students is statistically significant at the 10% level, which suggests that the intervention may have been effective for this group. The IV-estimate, however, does not confirm this. Hence, these results do not provide clear evidence of a beneficial effect of the policy for any one of these subgroups.

3.6 Sensitivity analyses

The design of incentive programs is essential to motivate agents and avoid undesired behaviour. We investigate to what extent specific characteristics of the covenant program might have influenced the effectiveness. We focus on consequences for incentive strength for RMC districts to put in additional effort and on consequences for potential adverse responses to the program. Incentive strength may be mitigated by the choice of the reference year and by district size. We explore both issues by performing analyses for specific subsamples of districts. We analyse the potential of strategic behaviour of the districts by comparing outcomes based on BRON and RMC data.

3.6.1 Choice of reference year

Since the rewards are based on the differences in the absolute number of dropouts between 2004-2005 and 2006-2007, pre-treatment dropout development between 2004-2005 and 2005-2006 already affects (expected) outcomes. Some regions have been ‘lucky’ (i.e. the ones that witnessed a decline in 2005-2006), while others were ‘unlucky’ (i.e. the regions with an increase in the pre-covenant year). This may have mitigated the incentive to put in additional effort aimed at decreasing dropout in the year 2006-2007. Low dropout rates in the pre-treatment year 2005-2006 might either discourage the districts or provide an additional incentive to work hard to compensate the negative point of departure. Likewise, an increase in pre-treatment dropout development might be interpreted as an additional stimulant or as a license to passivity.

To analyse the effect of pre-treatment dropout development on the impact of the program, we construct specific subsamples of regions. We distinguish between RMC districts that faced a decrease in the total number of dropouts between 2004-2005 and 2005-2006 according to the RMC data, and RMC districts that faced an increase.

Table 3.7 presents separate estimation results for various subsets of regions. The first column presents the full model results for the subset of all regions that faced a decline. The second column includes all regions that faced an increase. In the third and fourth columns, we include all control regions (either facing a decline or increase) and only the treatment regions with a decline in model (3) and treatment regions with an increase in model (4). All difference-in-differences models are estimated with OLS including school district fixed effects. Estimations with 2SLS are only presented for the models with regions that face an increase (since the two deviant regions both faced an increase).

The results yield larger point estimates (in absolute value) for the subsets of regions that witnessed a decline, but none of them are statistically significant. Hence, there is no evidence that the choice of reference year has systematically mitigated the effectiveness of the financial incentive. An additional estimation (not in the table) of a model which includes the increase (or decrease) in the absolute number of dropouts between 2004-2005 and 2005-2006 in terms of percentages as covariate (instead of the fixed effects) and an interaction term with the treatment effect of interest yields an insignificant effect for the interaction term. This indicates that the treatment effect is not significantly different between the 'lucky ones' and 'unlucky ones'. Therefore, we conclude that the choice of reference year and the subsequent confrontation of the districts with pre-treatment dropout development did not structurally alter the conclusions on the effectiveness of the program.

3.6.2 District size

A second difference between RMC regions that might be important for the effectiveness of the covenants is size. A potential disadvantage of centralizing programs is that incentives for the agents involved become diluted. In the context of the covenant policy one might expect the program to be less effective in larger districts because of weakened incentives. Empirical evidence along these lines is provided by Burgess et al. (2004), who evaluated the impact of a performance pay program in a major government agency, and found a substantial positive effect in small teams and a negative effect in large teams. Since the eligibility rule implies that larger districts are more likely to be selected into treatment, this may have affected the results of the covenant policy.

The last two columns of Table 3.7 present estimation results for subsets of the smallest 7 treatment regions and the largest 7 treatment regions, measured with the number of relevant students at the start of the school year 2006-2007. In both models all non-treatment regions are taken into account.[28] Average district sizes in both treatment groups are 40,580 and 69,622. Both OLS and 2SLS estimations are presented.

[28] Note that since eligibility was based on the absolute number of dropouts, it does not work to divide the total sample of districts in a subset of 'larger' and 'smaller' regions as almost all treatment regions would be in the former subset.

Table 3.7 Sensitivity analyses: difference-in-differences estimates of the effect of the covenants on the probability of dropping out, estimates in percentage points

	Pre-treatment dropout development				District size	
	(1) All regions with decline	(2) All regions with increase	(3) Treatment regions with decline; all control regions	(4) Treatment regions with increase; all control regions	(5) small	(6) large
OLS	-0.173 (0.228)	-0.136 (0.112)	-0.220 (0.211)	-0.141 (0.087)	-0.081 (0.117)	-0.227* (0.124)
2SLS		-0.020 (0.104)		-0.039 (0.080)	-0.014 (0.112)	-0.118 (0.144)
Control Variables						
Personal characteristics	yes	yes	yes	yes	yes	yes
Education level	yes	yes	yes	yes	yes	yes
Environment	yes	yes	yes	yes	yes	yes
Observations	890,343	1,697,253	1,573,269	2,068,104	1,618,761	2,022,612

Notes. Robust standard errors are in parentheses. * denotes significance at a 10 % significance level. The OLS models also include school district fixed effects.

Estimated coefficients are larger (in absolute value) in the larger districts. The OLS estimation in the subset of large treatment regions is statistically significant at a 10% level. This finding is contrary to expected. A suggestive explanation might be that in larger regions, there is a higher potential benefit from an increased coordination of activities of educators within the district. Nevertheless, we have to be cautious in the interpretation of this finding. First of all, the estimated coefficient is only significant at a 10% level. Following Donald and Lang (2007) we have to be cautious drawing conclusions from inference too soon. Secondly, when estimating the model with 2SLS the point estimate decreases and the effect becomes insignificant. Moreover, an additional OLS estimation (not in the table) of a full sample model in which we include the district size (instead of the fixed effects) and an interaction term between size and the treatment variable of interest, yields a highly insignificant interaction term. This indicates that the treatment effect does not significantly differ with district size. Altogether, we conclude that there is too little evidence to state that the incentive program has been more effective in larger regions. Nevertheless, the estimates suggest that the selection of larger regions into treatment has not negatively affected the impact of the covenant policy.

3.6.3 Strategic behaviour

There is much empirical evidence for strategic behaviour in response to financial incentives (see e.g. Jacob, 2005; Jacob and Levitt, 2003). In our evaluation we use administrative BRON data which are checked by accountants. Hence we can be sure that our data, which played no role in the determination of rewards, are not biased by manipulation.

Since the policy was introduced in the transition period between using RMC figures and BRON data, the government had to rely on the RMC figures yet for measuring performance in terms of reduced number of dropouts between the reference year 2004-2005 and 2006-2007. Hence, the monetary reward was based on the RMC figures provided by the districts themselves. This may have provided an opportunity for strategic behaviour in the treatment districts. If regions have put effort in manipulation of the data rather than in reducing dropout, this may have negatively affected the impact of the policy.

The coexistence of RMC and BRON data for the years 2005-2006 and 2006-2007 allows us to look further into the issue of strategic behaviour. We exploit availability of both data sets by comparing dropout development in the RMC data with dropout development in the BRON data for both treatment and control groups. Despite the serious reliability problems with RMC data, this may provide some insight in the occurrence of strategic behaviour. Using BRON we did not find evidence that the covenants have been effective in reducing dropout. In case of manipulation, we would expect to observe a more favourable dropout development in the RMC data of the treatment districts.

Table 3.8 shows the difference in the number of dropouts between 2005-2006 and 2006-2007 in terms of percentages for the treatment and the non-treatment regions, based on both RMC data and BRON data.[29] We have excluded one non-covenant region which reported an unrealistically large increase in dropouts.[30] In the treatment regions, dropout decreased with 3.2 percent according to the BRON data and

[29] The RMC data we use concern the new dropouts, i.e. all early school-leavers within the relevant school year. Hence, the stock of previous dropouts that still satisfies the definition of a school dropout in the relevant school year is not included in both RMC and BRON data.

[30] This region is 'Midden-Brabant' which reported an increase of 910% in dropouts between 2005-2006 en 2006-2007. Inclusion of this outlier would increase average dropout development in the non-treatment regions according to RMC from 5.8% to 12.2%.

decreased with 7.5 percent according to the RMC data. The larger reduction in school dropout according to RMC is in line with expectations in case of strategic behaviour. However, it may also reflect other changes over time in the RMC data, like for instance an improvement in the accuracy of the RMC registrations.

If these other changes are general developments in the RMC data and not specifically related to the treatment regions, we can control for them by comparing the relative development of the RMC figures in the treatment regions with the relative development in the non-treatment regions. In the non-treatment regions we observe an increase in school dropouts of 5.8 percent according to RMC relative to a 3.3 percent decrease according to BRON. Hence, the treatment regions report a 4.3 percentage points more favourable dropout development compared to BRON, while the non-treatment regions report a 9.1 percentage points less favourable dropout development. This suggests that the introduction of the covenant is associated with a 13.4 percentage points more favourable reported dropout development in the RMC figures relative to BRON. Strategic self-reporting behaviour in response to the program might explain this difference.

Table 3.8 Difference in the number of dropouts between pre-treatment year (2005-2006) and post-treatment year (2006-2007) in terms of percentages, based on both BRON and RMC data, covenant versus non-covenant regions, complete sample

Data Source	Covenant	Non-covenant
RMC	-7.5	5.8
BRON	-3.2	-3.3
Difference	-4.3	9.1

In addition, we have compared the BRON and RMC figures for a specific subset of treatment regions with a weak performance in dropout reduction. We would expect regions that actually do not realize a reduction in school dropouts to be most likely to behave opportunistically. It turns out that 3 treatment regions did not realize a reduction in dropouts according to BRON. These regions on average face an increase in school dropout of 4.6 percent. The corresponding reported average dropout development in the RMC data is, however, -4.6 percent (not in the table). Hence, we indeed observe a larger difference between the self-reported RMC figures and the administrative BRON data (of 9.2 percentage points) in the regions that seem most suspicious of strategic behaviour. This adds to the conjecture that manipulation of the data underlies the reported figures in Table 3.8.[31]

Nevertheless, we cannot completely exclude that the observed pattern reflects changes in the RMC data that are different between treatment and non-treatment regions. Acknowledging that the observed pattern might still be driven by a lack of uniformity in (the development of) the RMC data, we interpret our findings as suggestive evidence for data manipulation in response to the introduction of the financial incentive scheme.

3.7 Conclusions

Incentive-based policies may provide a cost-effective way to improve educational achievement compared to traditional expenditure-based policies. Policymakers and researchers have moved their interest more and more towards the role of incentives in the last decade. A variety of new incentive-based policies have been started, especially in the US and the UK. Empirical evaluations of these programs show that financial incentives can improve education quality. At the same time there remains severe opposition from teacher unions and educators against the use of incentives and strategic responses seem to be a serious threat to the effectiveness of the programs.

Knowledge on optimal policy design is important to successfully implement incentive-based schemes. This study contributes to the literature on the role of incentives in education. So far, the main interest has been on incentives targeted at the level of students, teachers or schools. We analyse the effectiveness of a program targeted at the level of school districts.

We exploit the gradual introduction of a substantial financial incentive scheme for local education authorities to reduce school dropout in the Netherlands in 2006. In the first year, covenants with financial incentives were offered to 14 out of 39 school districts with the highest number of dropouts. This selection rule, and a unique dataset containing individual information on the whole student population in both the pre- and post-treatment year, offers a good opportunity to evaluate the effects of the treatment. We use a (local) difference-in-differences approach and find statistically insignificant effects on dropout probability. These findings are robust to a variety of specifications.

[31] When dividing the remaining 11 treatment regions in different subsets of high performing, middle performing and low performing regions, we observe in all of these subsets a larger reported dropout reduction in RMC compared to BRON. We do not find a clear pattern of increasing differences between RMC and BRON figures when actual performance (in terms of BRON) decreases.

Specific program details may have importantly affected the empirical findings on effectiveness. In order to add to the interpretation of our results, we have investigated sensitivity to some of these program-specific characteristics.

First, the choice of the reference year 2004-2005 implies that the pre-treatment dropout development between 2004-2005 and 2005-2006 may already have affected outcomes. Second, the choice to select and reward regions based on the absolute number of dropouts implied that incentives were targeted at larger districts which might have negatively influenced the results because of dilution of incentives. Sensitivity analyses, however, suggest that these aspects did not importantly affect the results.

A third and probably most important issue is that the government used the region-level RMC data to measure performance, which may have provided an opportunity for strategic behaviour of treatment regions. From a comparison between BRON and RMC figures we tentatively conclude that there appears to be a strategic response of the local education authorities in terms of manipulation of the data.

In sum, we find no evidence that an incentive scheme targeted at regional education authorities has been effective. The ineffectiveness of the scheme might be due to strategic behaviour of treatment districts at the cost of purposive actions. This once more points out the importance of a proper design with well-thought prior conditions for a successful implementation of incentive programs in education. In addition, we should be aware that the effectiveness of centralized incentives can be dependent on specific characteristics of the education system, like the large school autonomy in the Netherlands. As school autonomy might limit the power of education authorities, this type of incentive scheme may be more successful in other education systems. The promise of incentive-based policies to increase educational quality at a relatively low cost asks for additional research efforts to improve knowledge on optimal program designs.

Appendix 3.A: Selection of covenant regions

Table 3.A.1 Selection of covenant regions: rankings of regions according to the number of new dropouts and total number of dropouts in 2004-2005

Ranking	Region	Number of new dropouts	Region	Total number of dropouts
1	**agglomeratie amsterdam**	7402	**agglomeratie amsterdam**	18047
2	**rijnmond**	6276	**rijnmond**	15379
3	**west-brabant**	4643	**west-brabant**	7803
4	**haaglanden**	4374	**haaglanden**	6360
5	**utrecht**	2662	**gewest limburg-zuid**	5631
6	**arnhem/nijmegen**	2260	**flevoland**	4845
7	**twente**	2072	**utrecht**	4185
8	**gewest limburg-zuid**	1841	**noordoost-brabant**	3310
9	**noordoost-brabant**	1829	**zuid-holland-zuid**	3111
10	**gewest noord-limburg**	1793	**arnhem/nijmegen**	2616
11	*zuidoost-brabant*	1773	zuid-holland-noord	2395
12	west-kennnemerland	1292	*centraal en westelijk groningen*	2254
13	*centraal en westelijk groningen*[b]	1279	**twente**	2162
14	zuid-holland-noord	1219	**gewest noord-limburg**	2059
15	**flevoland**	1218	*zuidoost-brabant* [b]	1773
16	westfriesland	1216	noord-kennemerland	1507
17	ijssel-vecht	1075	west-kennnemerland	1292
18	zuidoost-drenthe	1052	eem en vallei	1283
19	eem en vallei	1038	westfriesland	1216
20	oost-gelderland	980	zuidoost-drenthe	1171
21	rivierenland	977	oost-gelderland	1148
22	**zuid-holland-zuid**	935	ijssel-vecht	1138
23	stedendriehoek	748	friesland-oost	1060
24	friesland-oost	697	noord- en midden-drenthe	1046
25	noord-kennemerland	665	rivierenland	1023
26	noord-groningen en eemsmond	646	oost-groningen	984
27	noord- en midden-drenthe	581	kop van noord-holland	842
28	midden-brabant	554	stedendriehoek	809
29	gooi en vechtstreek	520	zuidwest-drenthe	740
30	zuid-holland-oost	467	friesland-noord	693
31	kop van noord-holland	417	gooi en vechtstreek	665
32	oost-groningen	391	noord-groningen en eemsmond	659
33	zuidwest-drenthe	385	midden-brabant	564
34	friesland-noord	383	noordwest-veluwe	530
35	zuidwest-friesland	380	zuidwest-friesland	517
36	walcheren	273	zuid-holland-oost	505
37	oosterschelde regio	255	walcheren	427
38	noordwest-veluwe	230	oosterschelde regio	367
39	zeeuwsch-vlaanderen	165	zeeuwsch-vlaanderen	356

Notes. Regions in bold are the selected covenant regions. The selection principle is presence in the top 10 of at least one of the two lists of new or total number of dropouts in 2004-05. Regions in italic are the two self-selected covenant regions that were interested in signing a covenant.

Source: RMC registration figures.

Appendix 3.B: Descriptive statistics of subsamples

Table 3.B.1 Sample means for covenant and non-covenant regions, school year 2005-2006 (pre-treatment year), DS 1

	Non-covenant regions	Covenant regions	*p*-Value
Dropout probability	3.93	3.82	0.043
Personal characteristics			
Gender	0.51	0.51	0.245
Age	15.28	15.34	0.000
Age: compulsory education	0.70	0.69	0.000
Ethnicity			0.000
Dutch	0.85	0.82	
Surinam	0.01	0.02	
Aruba/The Antilles	0.01	0.01	
Turkey	0.02	0.03	
Morocco	0.02	0.02	
Other foreign (non-western)	0.04	0.04	
Other foreign (western)	0.06	0.06	
Education level			
Level			0.000
first grade secondary education	0.31	0.31	
vmbo (level 1)	0.05	0.05	
vmbo (level 2)	0.05	0.04	
vmbo (level 3/4)	0.08	0.08	
havo	0.11	0.10	
vwo	0.12	0.12	
mbo-1	0.01	0.01	
mbo-2	0.07	0.08	
mbo-3	0.08	0.07	
mbo-4	0.13	0.14	
Exam class	0.15	0.14	0.268
Environment of the pupil			
Degree of urbanization			0.000
Inhabitant of G4 (4 largest cities)	0.00	0.00	
Medium-sized municipality	0.28	0.32	
Small municipality	0.72	0.68	
Inhabitant poverty accumulation area	0.05	0.14	0.000
RMC budget per student (€)	12.22	10.00	0.000
Total number of pupils	223,324	344,091	

Table 3.B.2 Sample means for covenant and non-covenant regions, school year 2005-2006 (pre-treatment year), DS 2

	Non-covenant regions	Covenant regions	*p*-Value
Dropout probability	4.04	3.79	0.000
Personal characteristics			
Gender	0.51	0.51	0.311
Age	15.26	15.32	0.000
Age: compulsory education	0.71	0.70	0.000
Ethnicity			0.000
Dutch	0.84	0.84	
Surinam	0.01	0.01	
Aruba/The Antilles	0.01	0.01	
Turkey	0.02	0.03	
Morocco	0.01	0.02	
Other foreign (non-western)	0.04	0.04	
Other foreign (western)	0.06	0.06	
Education level			
Level			0.000
first grade secondary education	0.31	0.31	
vmbo (level 1)	0.05	0.05	
vmbo (level 2)	0.05	0.04	
vmbo (level 3/4)	0.08	0.08	
havo	0.11	0.11	
vwo	0.13	0.12	
mbo-1	0.01	0.01	
mbo-2	0.07	0.07	
mbo-3	0.08	0.07	
mbo-4	0.12	0.14	
Exam class	0.15	0.14	0.000
Environment of the pupil			
Degree of urbanization			0.000
Inhabitant of G4 (4 largest cities)	0.00	0.00	
Medium-sized municipality	0.25	0.27	
Small municipality	0.75	0.73	
Inhabitant poverty accumulation area	0.06	0.11	0.000
RMC budget per student (€)	12.35	10.32	0.000
Total number of pupils	133,613	209,250	

Table 3.B.3 Sample means for covenant and non-covenant regions, school year 2005-2006 (pre-treatment year), MS 1

	Non-covenant regions	Covenant regions	*p*-Value
Dropout probability	3.99	4.01	0.731
Personal characteristics			
Gender	0.51	0.51	0.523
Age	15.29	15.30	0.357
Age: compulsory education	0.70	0.70	0.308
Ethnicity			0.000
Dutch	0.85	0.82	
Surinam	0.01	0.01	
Aruba/The Antilles	0.01	0.01	
Turkey	0.02	0.03	
Morocco	0.02	0.03	
Other foreign (non-western)	0.04	0.04	
Other foreign (western)	0.06	0.06	
Education level			
Level			0.000
first grade secondary education	0.30	0.31	
vmbo (level 1)	0.05	0.05	
vmbo (level 2)	0.05	0.04	
vmbo (level 3/4)	0.08	0.08	
havo	0.11	0.11	
vwo	0.12	0.12	
mbo-1	0.01	0.01	
mbo-2	0.07	0.07	
mbo-3	0.07	0.07	
mbo-4	0.14	0.14	
Exam class	0.15	0.14	0.246
Environment of the pupil			
Degree of urbanization			0.000
Inhabitant of G4 (4 largest cities)	0.00	0.04	
Medium-sized municipality	0.27	0.24	
Small municipality	0.73	0.72	
Inhabitant poverty accumulation area	0.04	0.12	0.000
RMC budget per student (€)	14.33	9.93	0.000
Total number of pupils	251,112	373,409	

Table 3.B.4 Sample means for covenant and non-covenant regions, school year 2005-2006 (pre-treatment year), MS 2

	Non-covenant regions	Covenant regions	*p*-Value
Dropout probability	3.96	4.03	0.317
Personal characteristics			
Gender	0.51	0.51	0.682
Age	15.32	15.30	0.015
Age: compulsory education	0.70	0.70	0.000
Ethnicity			0.000
Dutch	0.85	0.82	
Surinam	0.01	0.01	
Aruba/The Antilles	0.01	0.01	
Turkey	0.02	0.03	
Morocco	0.02	0.02	
Other foreign (non-western)	0.04	0.04	
Other foreign (western)	0.06	0.06	
Education level			
Level			0.000
first grade secondary education	0.30	0.31	
vmbo (level 1)	0.05	0.05	
vmbo (level 2)	0.05	0.04	
vmbo (level 3/4)	0.08	0.08	
havo	0.11	0.11	
vwo	0.11	0.12	
mbo-1	0.01	0.01	
mbo-2	0.07	0.07	
mbo-3	0.07	0.07	
mbo-4	0.14	0.14	
Exam class	0.14	0.14	0.639
Environment of the pupil			
Degree of urbanization			0.000
Inhabitant of G4 (4 largest cities)	0.00	0.00	
Medium-sized municipality	0.34	0.32	
Small municipality	0.66	0.68	
Inhabitant poverty accumulation area	0.05	0.13	0.000
RMC budget per student (€)	14.04	9.92	0.000
Total number of pupils	183,362	222,299	

Appendix 3.C: Additional analyses

Table 3.C.1 OLS Difference-in-differences estimates of the effect of the covenants on the probability of dropping out, analysis at the school district level, estimates in percentage points

	(1)	(2)	(3)	(4)	(5)
Covenant	0.051	0.102	0.052	0.043	0.030
	(0.337)	(0.171)	(0.162)	(0.161)	(0.149)
Control Variables					
Personal characteristics	no	yes	yes	yes	yes
Education level	no	no	yes	yes	yes
Environment	no	no	no	yes	yes
Polynomial of dropouts	no	no	no	no	yes
Observations	78	78	78	78	78

Notes. Robust standard errors are in parentheses.

Table 3.C.2 OLS Difference-in-differences estimates of the effect of the covenants on the probability of dropping out, sample without two self-selected regions, estimates in percentage points

	(1)	(2)	(3)	(4)	(5)	(6)
Covenant	-0.030	-0.029	-0.052	-0.059	-0.159	-0.098
	(0.104)	(0.100)	(0.098)	(0.101)	(0.111)	(0.103)
Control Variables						
Personal characteristics	no	no	yes	yes	yes	yes
Education level	no	no	no	yes	yes	yes
Environment	no	no	no	no	yes	yes
Fixed effects	no	yes	yes	yes	yes	no
Polynomial of dropouts	no	no	no	no	no	yes
Observations	2,430,158	2,430,158	2,430,158	2,430,158	2,430,158	2,430,158

Notes. Robust standard errors are in parentheses.

Table 3.C.3 OLS Difference-in-differences estimates of the effect of the covenants on the probability of dropping out, for discontinuity samples (DS) and matched samples (MS), estimates in percentage points

	DS 1		DS 2		MS 1		MS 2	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Covenant	-0.099 (0.110)	-0.040 (0.102)	-0.147 (0.131)	-0.019 (0.113)	-0.099 (0.124)	-0.016 (0.103)	-0.053 (0.155)	-0.017 (0.130)
Control Variables								
Personal characteristics	yes	yes	yes	yes	yes	yes	yes	yes
Education level	yes	yes	yes	yes	yes	yes	yes	yes
Environment	yes	yes	yes	yes	yes	yes	yes	yes
Fixed effects	yes	no	yes	no	yes	no	yes	no
Polynomial of dropouts	no	yes	no	yes	no	yes	no	yes
Number of covenant districts	8	8	5	5	8	8	5	5
Number of non-covenant districts	8	8	5	5	11	11	8	8
Observations	1,143,050	1,143,050	690,111	690,111	1,256,966	1,256,966	816,546	816,546

Notes. Robust standard errors are in parentheses.

Table 3.C.4 OLS Difference-in-differences estimates of the effect of the covenants on the probability of dropping out, sample without four largest cities, estimates in percentage points

	(1)	(2)	(3)	(4)	(5)	(6)
Covenant	-0.013	-0.014	-0.034	-0.054	-0.143	-0.094
	(0.073)	(0.073)	(0.070)	(0.071)	(0.096)	(0.075)
Control Variables						
Personal characteristics	no	no	yes	yes	yes	yes
Education level	no	no	no	yes	yes	yes
Environment	no	no	no	no	yes	yes
Fixed effects	no	yes	yes	yes	yes	no
Polynomial of dropouts	no	no	no	no	no	yes
Observations	2,316,453	2,316,453	2,316,453	2,316,453	2,316,453	2,316,453

Notes. Robust standard errors are in parentheses.

4

The effects of a comprehensive program for youths-at-risk: Evidence from a controlled natural experiment[1]

Abstract

This chapter evaluates the effects of a comprehensive program designed to increase school enrolment and employment among youths-at-risk. The treatment consisted of a broad range of educational, work, and health services, and guidance by personal coaches. For evaluating the impact of the program we exploit variation in assignment to the program induced by capacity restrictions of the two program sites. During specific time windows potential candidates for the program were assigned to the control group that received treatment as usual. We find that the program did not succeed in increasing enrolment in education or employment three years after the start of the program. Most important, we find that assignment to the program increased criminal activity, especially among the subpopulation of youths who had been suspected of a crime before the start of the program. Peer effects caused by grouping youths-at-risk together may explain the adverse impact on criminal behaviour.

[1] This chapter is based on joint work with Marc van der Steeg and Dinand Webbink: Van Elk, R., M. van der Steeg, D. Webbink, 2013, The effects of a special program for multi-problem school dropouts on educational enrolment, employment, and criminal behaviour, CPB Discussion Paper 241.

4.1 Introduction

Young individuals that do not finish their schooling and do not have a job are at risk for future unemployment and for getting involved in criminal activities (e.g., Lochner and Moretti, 2004; Machin et al., 2011). The high social costs associated with unemployment and crime legitimize publicly subsidized interventions aimed at improving the prospects for these youths. Many of these interventions focus on reducing school dropout. In general, interventions targeted at youths still enrolled in school are more effective than interventions targeted at out-of-school youths (Carneiro and Heckman, 2003).[2] The results of programs focused on out-of-school youths, which is the target group of the intervention studied in this chapter, are ambiguous. Many publicly subsidized programs appear not to be successful in improving the lives of these youths (Lalonde, 2003).[3] The Job Corps program might be considered as one of the most promising interventions. Job Corps is a relatively intensive and expensive program that provides a comprehensive set of training services – including vocational education, counselling, work experience, social skills training, and health education – for disadvantaged youths aged 16 to 24 in a residential setting. The program has been found to increase educational attainment and earnings and to reduce arrest rates during the treatment period (Schochet et al., 2001). However, the earnings gains and the reduction in criminal activity do not persist in the longer term (Schochet et al., 2008). The National Guard Youth ChalleNGe Program, an intensive military structured treatment that offers residential-based education and mentoring to young dropouts, increased college enrolment, employment, and earnings, but did not decrease criminal behaviour among participants, and negatively affected health and lifestyle outcomes (Millenky et al., 2011). The Year-Up program, which offers six months of full-time classes followed by six months of internships with U.S. companies to minority youths, aged 18 to 24, increased earnings one year after the end of the program but did not affect employment (Roder and Elliot, 2011). Carneiro and

[2] Examples of programs for youth still enrolled in school in the United States that improved educational outcomes are Big Brothers Big Sisters (Tierney et al., 1995) and the Philadelphia Futures' Sponsor-A-Scholar (Johnson, 1999). Two programs that provided financial incentives to teenage parents – Ohio's Learning, Earning, and Parenting Program and the Teenage Parent Demonstration – increased future earnings and employment among students who were still in school, but not among participants who had already dropped out (Granger and Cytron, 1998). The Quantum Opportunities Program, which combines counselling minority students with financial incentives, has increased educational attainment in the short run (Taggart, 1995; Maxfield et al., 2003). A 10-year follow-up evaluation, however, suggests that most of the beneficial effects disappear in the long run and that the program increased the likelihood of committing a crime for male participants (Rodriguez-Planas, 2012). Recently, a novel intervention for disadvantaged youths in Chicago high schools, that includes mentoring and cognitive behavioural therapy to reduce intuitive judgement and decision-making problems, has been found to improve schooling and crime outcomes (Heller et al., 2013).

[3] Experimental evaluations of federal programs such as JobStart, the Job Training Partnership Act, and the Work Experience Programs do not find beneficial effects on employment or earnings (Cave and Doolittle, 1991; Couch, 1992; Orr et al., 1994).

Heckman (2003) have summarized the lessons from previous studies into guidelines for designing effective interventions for youths-at-risk. The most promising programs consist of an integrated mix of education, occupational skills training, work-based learning, and supporting services, and also include adult mentoring to handle inappropriate attitudes. In this chapter we investigate the effects of a program which was designed according to these guidelines and focused on youths who were considered as a really problematic group.

In 2009 the Dutch government introduced the so-called Neighbourhood School Program (NSP) in Rotterdam, the second largest city of the country with a relatively high unemployment and crime rate. The NSP is a comprehensive and expensive program for youths-at-risk aged 16 to 23 living in disadvantaged areas. The aim of the program was to help youths-at-risk return to school or enter the labour market. Individuals in the NSP received a comprehensive set of educational, work and health services and were assigned to a personal coach. The design of the program seems quite similar to the Job Corps program. Compared to other programs discussed above the NSP is targeted at a more specific subpopulation of youths-at-risk who face complex behavioural, financial, or health related problems. Around 50% of the target population had been suspected of a crime before they entered the program and around 70% had only completed primary education. The program's content is relatively intensive and adds assistance by professional health specialists to work-based and educational elements to help youths solve their problems. To evaluate the impact of the program we exploit exogenous variation in treatment status induced by a specific assignment rule to the program. This approach was chosen because random assignment of the target group to the program and to a control condition was considered not feasible and appropriate. After a series of discussions with the team responsible for the implementation of the program it was agreed to create a controlled natural experiment. The agreement was that during specific time windows potential candidates for the program would be assigned to the regular interventions used by the municipality. The time windows were chosen based on the expected number of participants in the program and the capacity restrictions of the two program sites. We use the youths that were eligible for the program but that were assigned to the regular interventions because of these time windows as our control group. Our data include administrative information about school enrolment, employment, and criminal behaviour three years after the start of the intervention. We investigate the impact of the program by comparing the outcomes of youths assigned to the NSP with the outcomes of youths assigned to regular interventions, conditional on the time of application. For assessing the impact of the program two additional issues are important; non-compliance with the assignment rules and non-starting in the program. It turned out that the project team did not fully comply with the assignment rules that were agreed upon. To address non-compliance with the assignment rule we use an instrumental variables approach in which the intended assignment is used as an instrument (Angrist et al., 1996). The second issue is about youths that did not start in the program they were assigned to. In the empirical analysis we cannot estimate standard treatment-on-the-treated effects because the exclusion restriction that 'never-takers' and 'always-takers' in treatment and control groups are treated in the same way is not likely to hold (see Section 4.4). This implies that the treatment that we evaluate in this study is the assignment to the NSP and the treatment effect is the combined effect of participants and non-starters in the NSP. In many policy evaluations the effect of the assignment to treatment is considered to be the most relevant effect.

We find that assignment to the program did not increase school enrolment or employment three years after the start of the program. Most important, we find evidence that assignment to the NSP increased criminal activity, especially among the youths who had been suspected of a crime before the time of entry. Additional analyses suggest that peer effects caused by grouping at-risk-youths together can explain this effect. Our study contributes to various strands of the economic literature on school dropout and crime. First, it investigates the effects of an intervention for youths-at-risk designed according to the main lessons from previous studies. Second, our findings with respect to school enrolment and employment are in line with a large body of the literature that shows no impact of training programs on the labour market outcomes of youths-at-risk (e.g., Carneiro and Heckman, 2003; LaLonde, 2003). Third, our results regarding crime are consistent with studies that document the adverse effects of group-based interventions on delinquent behaviour (Dishion et al., 1999) and studies that find evidence that peer effects due to grouping at-risk adolescents together can explain the reinforcement of criminal activity (Dodge et al., 2007). Fourth, our study might be important for designing field experiments in situations in which random assignment appears not to be feasible. We provide an example of a non-standard field experiment in which the assignment to the intervention was based on specific time windows.

The rest of this chapter is organized as follows. Section 4.2 describes the institutional background, the content of the intervention, and assignment to the program. Section 4.3 presents the data and Section 4.4 discusses the empirical strategy. Section 4.5 reports the main findings and Section 4.6 presents additional analyses that add to the interpretation of our main findings. Section 4.7 concludes.

4.2 The NSP

4.2.1 Background

In the Dutch education system a school dropout is defined as someone aged under 23 who leaves school without having completed a particular level of education called the start-qualification. This start-qualification is considered the minimum level of education needed to participate well in the labour market and corresponds to a degree in higher secondary education or intermediate vocational education.[4] Rotterdam, the second largest city of the Netherlands, typically has a large share of school dropouts and unemployed youth (Ministry of Education, 2010; Municipality of Rotterdam, 2011). The fraction of school dropouts in Rotterdam (6.1 %) is almost twice the national average (3.2 %) and larger than that of the three other large cities (5.4 %). Out-of-school youths aged 18 to 27 without a job can apply for help and income services at their municipality of residence. Municipalities are obliged to offer a suitable reintegration program to applicants. Acceptance of this offer is compulsory for eligibility for social benefits. Youths aged under 18 without a start-qualification are not served because of the compulsory education requirement. In Rotterdam, the Young People's Office (YPO) is responsible for the assignment of youths to reintegration programs. Each applicant is invited for a consultation, after which he or she is referred to the best-suited program. If youths refuse to participate in the proposed program, they do not receive a second offer and lose their rights to social benefits.

4.2.2 The intervention

The NSP was launched at two sites in Rotterdam in 2009 as an alternative to regular reintegration programs. The purpose of the NSP is to help multi-problem school dropouts return to formal education or work. Initially, the Ministries of Education, Social Affairs and Employment, Justice, and Health contributed 5.6 million Euros to finance a two-year pilot. After two years, the pilot was extended.

The program's target population consists of school dropouts aged 16 to 23 who are unemployed and face complex problems in several areas of their lives. More specifically, youths must satisfy all of the following criteria for NSP eligibility:

- Is aged 16 to 23,

[4] Unemployment among youths without a start-qualification was 13 % in 2009 and is more than twice as high as unemployment among youths with a start-qualification (Statistics Netherlands, 2011).

- Dropped out of school without having obtained a degree in intermediate vocational education or higher secondary education,
- Does not work in a structured job for more than 12 hours a week,
- Faces problems in at least two areas of their lives, including work, finances, health, housing, justice, and social environment, and
- Has an IQ above 70.

Participants in the program are formally registered as being enrolled in education and receive student grants. The program provides an integrated treatment of educational, work and care services. A special team of adult coaches and health specialists support the youths to solve their problems, increase their skills, and help them return to formal education or work. Each participant is counselled by a personal coach and is expected to be present five days a week, from 9:00 to 15:30.

The program's educational component consists of courses in specific subjects in which youths can obtain certificates. The educational level of these courses is comparable to the lowest level of intermediate vocational education. Besides increasing learning skills and knowledge, successful participation in these courses can contribute to self-esteem. With respect to work, neighbourhood schools use their regional networks to arrange small jobs or internships with local firms. These internships teach youths to participate in the labour market under the supervision of their coaches. In addition, group trainings are organized to improve social skills.[5] Next to the educational and work services, youths are professionally supported by health specialists, including social workers, psychiatric nurses, a behavioural specialist, and a (part-time) psychiatrist to help solve their personal problems. The NSP is characterized by its small-scale and personal approach. Neighbourhood schools offer customized treatment with respect to both program content and duration. The precise content is adjusted to the specific needs of the individuals and the program length is flexible, with an average duration of around 10 months. Each of the two neighbourhood schools, one located in the north and the other in the south of Rotterdam, has a capacity of 100 places. Youths accepted into the program are usually sent to one of these locations, based on their place of residence. The total costs of the program equal around 14,000 Euros per place per year.

Compared to the regular programs, the NSP differs in four ways. First, it is a small-scale program that groups youths together and treats them in a personal, customized way with flexible time duration. Second, it offers a more comprehensive mix of educational, job training, and health services. Third, it is a more

[5] A growing body of literature addresses the role of non-cognitive skills in explaining criminal behaviour. Hill et al. (2011) review interventions that focus on non-cognitive factors and argue that promoting social skills is one of the elements shared by a number of effective interventions.

intensive and expensive program. Fourth, participants are formally enrolled in education and receive student grants rather than social benefits.

4.2.3 Assignment to the NSP

In a series of discussions with the team that was responsible for the implementation of the NSP and the YPO it was agreed to create a controlled natural experiment. The agreement was that during several time windows potential candidates for the program would be assigned to the regular interventions used by the municipality. The time windows were chosen based on the expected number of participants in the program and the capacity restrictions of the two program sites. In the first period, between 1 August 2009 and 1 March 2010, all youths who satisfied the eligibility criteria were assigned to the NSP.[6] As of 1 March 2010, the neighbourhood schools had reached their maximum capacity and hence could not accept more participants. Subsequently, all youths who satisfied the eligibility criteria and had their consultation at the YPO in the second period, between 1 March 2010 and 15 April 2010, were assigned to a regular reintegration program.[7] These youths received 'the treatment as usual' by the YPO and will be used as the control group in our analysis. They would have been assigned to the NSP in other periods. This procedure was repeated in subsequent periods. In the third period, between 15 April 2010 and 1 July 2010 all youths who satisfied the eligibility criteria were sent to the NSP and in the fourth period, between 1 July 2010 and 15 September 2010, all youths who satisfied the eligibility criteria were assigned to a regular program.

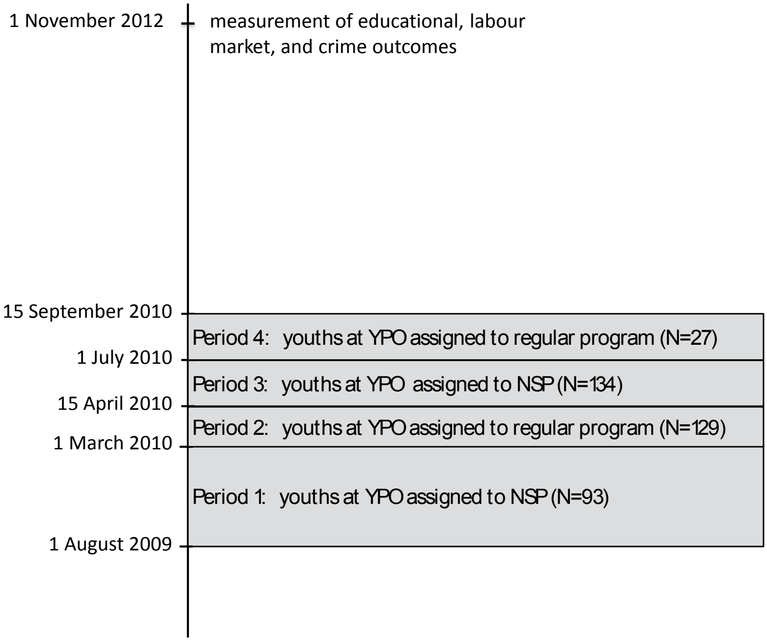
Figure 4.1 presents our research design. The assignment rule ensures that treatment status depends deterministically on an individual's consultation date at the YPO. Youths who enter the YPO in the first or third period are assigned to the NSP, while similarly eligible pupils who enter the YPO in the second or fourth period are assigned to one of the regular programs. This approach allows us to identify the effects of assignment to the NSP by comparing relevant schooling, labour market, and crime outcomes for both groups, conditional on the time of consultation. We measure these outcome variables on 1 November 2012, two to three years after assignment to a program (see Section 4.3).

[6] When youths apply, the YPO ensures that they meet the eligibility criteria. We chose to start collecting data from 1 August 2009 and thereby excluded youths that entered the YPO in the beginning of the NSP, between March and August 2009. This reduces the probability that the results are affected by potential start-up problems at the NSP.

[7] It was agreed that all youths would be sent to a regular program during this entire predetermined period. Hence, even when new places became available in the neighbourhood schools during this period (because of the outflow of participants), none of the applicants were sent to the NSP.

Our analysis focuses on youths aged 18 and older, since the regular reintegration programs do not accept 16- and 17-year-olds because of compulsory education. A total of 383 individuals who satisfied the profile had a consultation at the YPO between 1 August 2009 and 15 September 2010, 93 in the first period, 129 in the second period, 134 in the third period, and 27 in the fourth. Variation in the number of consultations by period is mainly caused by typical school dropout patterns during the school year. Most dropouts leave school in the last months of the school year. This explains the peak in consultations during March till July (periods 2 and 3) and the much lower number of consultations in the first half of the school year and during the summer holidays (periods 1 and 4).

Figure 4.1 Research design.


1 November 2012
measurement of educational, labour market, and crime outcomes
15 September 2010
Period 4: youths at YPO assigned to regular program (N=27)
1 July 2010
Period 3: youths at YPO assigned to NSP (N=134)
15 April 2010
Period 2: youths at YPO assigned to regular program (N=129)
1 March 2010
Period 1: youths at YPO assigned to NSP (N=93)
1 August 2009

Non-compliance

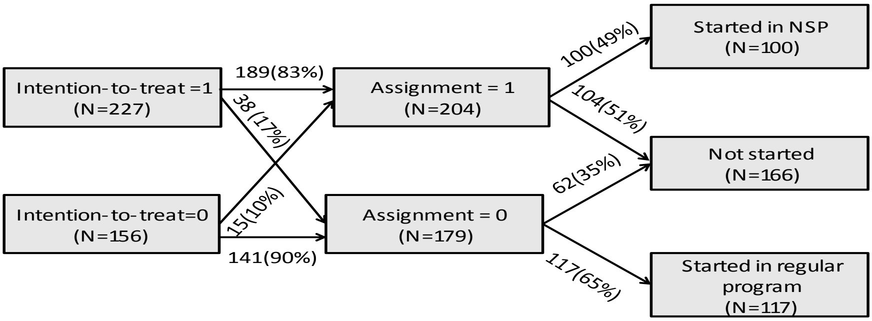
According to the assignment rule, 227 youths should be assigned to the NSP and 156 youths should be assigned to a regular program. We received information on each individual's consultation date and corresponding assignment from the YPO administration. These youths were assigned to one of the two neighbourhood schools or one of the 27 reintegration programs.

These data reveal that in practice not all youths were assigned according to their time of entry. The YPO employees did not always act in line with the assignment rule and sometimes over-ruled our research design by referring individuals who were not intended to be treated to the NSP, and vice versa. It turns out that 39 of the 227 individuals who should have been assigned to the NSP were sent to a regular program. Similarly, 12 of the 156 individuals who should have been assigned to a regular program were assigned to the NSP. Because of this non-compliance, we distinguish between an 'intention-to-treat' status and an 'assignment' status. Youths who entered the YPO in the first or third period were intended to be treated and youths who entered the YPO in the second or fourth period were not intended to be treated. The intention-to-treat variable is thus only determined by the consultation date. The assignment variable depends on the actual assignment and takes the value of one if an individual is assigned to the NSP and zero if not. The assignment differs from the intention-to-treat when the actual assignment is not in accordance with the assignment rule. A second – and less important – reason why assignment does not always coincide with the intended treatment is that four individuals had a second consultation at the YPO after 15 September 2010. These four individuals had initially been assigned to a regular program and were reassigned to the NSP after their second consultation. [8] Of the 227 individuals who were intended to be treated, 38 (17%) were assigned to a regular program. Similarly, 15 of the 156 individuals (10%) not intended to be treated were assigned to the NSP.

Furthermore, not all youths who were assigned to the NSP decided to participate. Of the 204 individuals assigned to the NSP, 104 refused to participate in the program. Hence, 51% of the youths that were assigned to NSP decided not to start at the neighbourhood school after the consultation at the YPO. These youths sometimes communicated that they did not have the intention to start, but in most cases they simply did not show up and could not be contacted. This large fraction of non-starters indicates that it is difficult to get the target group into treatment. For a comparison, the fraction of non-starters among the remaining 179 youths assigned to one of the regular reintegration programs was also relatively high, at 35%. There are two potential explanations for the larger fraction of non-starters among youths assigned to the NSP. First, they may be less inclined to participate in a more intensive program that requires more effort and time. Second, youths in the NSP receive student grants instead of social benefits, which is financially less attractive. Figure 4.2 summarizes the assignment flows by intention-to-treat status and starting behaviour by assignment status for all 383 youths in our estimation sample.

[8] One of the individuals with a later assignment to the NSP, was intended to be treated, but initially referred to the control group. The other three individuals all had an intention-to-treat status equal to zero.

Figure 4.2 Intention-to-treat, assignment, and starting behaviour.



4.3 Data

Our data come from three sources. First, we use data about the educational and labour market position from the YPO's administrative records. This information is collected in the status forms completed by the YPO on four different reference dates: 1 April 2011, 1 October 2011, 1 May 2012, and 1 November 2012. In addition to individual information on consultation dates and assigned programs, the status forms contain four dummy variables for each youth, indicating (i) whether the individual has ever participated in the assigned program, (ii) whether the individual has moved out of the program on the corresponding date, (iii) whether the individual is enrolled in regular education on the corresponding date and (iv) whether the individual had a job on the corresponding reference date. Second, individual background characteristics are collected from registration forms, which are completed by the YPO during the consultations. Third, we use the administrative data from the police of Rotterdam–Rijnmond. These data contain information on the number of crimes an individual is suspected of having committed during the period between1 August 2001 and 1 November 2012.

As a first outcome, we use a dummy variable that indicates return to school or to the labour market. This variable has value one if an individual is in education and/or has a job and zero otherwise. We use this outcome variable to investigate the effects of the NSP on school enrolment and/or employment. Our main analyses use the data from the status form on reference date 1 November 2012, since this provides the most recent information on educational enrolment and employment and is most informative of the longer-term effects of the NSP. At this date all participating youths had moved out of the NSP. In addition, we have constructed a variable that indicates the yearly average number of crimes an individual is suspected of during the period between the consultation date and 1 November 2012. Although suspected individuals might not have committed a crime, it seems plausible that this variable has a positive correlation with criminal activity. We label this variable as the crime rate after the consultation date and use it to assess the effects of the NSP on criminal behaviour. Although a reduction in criminal activity is not an explicit goal of the program, it seems to be a relevant outcome regarding the multi-problem target population and the program's health services focused on solving personal and behavioural problems.

As covariates we can use a set of individual background characteristics, including gender, age, country of birth and educational attainment. In addition, we construct three variables indicating the average yearly number of crimes an individual is suspected of during the eight-, four- or two-year periods before the consultation date at the YPO. We denote these as the eight-, four- and two-year pre-treatment crime rates and add these objective measures of problem intensity as covariates to our analyses.

Table 4.1 shows the descriptive statistics for all individuals in our estimation sample by intention-to-treat status. The first column reports the sample means of the 227 youths who were intended to be treated. The second column reports the sample means of the 156 youths who were not intended to be treated. The third column reports the *p*-value of the difference, calculated using a two-tailed *t*-test or a chi-squared test. The latest educational position (10 categories) indicates the most recent education type an individual was enrolled in at the time of consultation. The highest completed education level (eight categories) refers to the highest educational degree an individual has obtained. A large fraction, 69%, of the individuals in our estimation sample only completed primary education. The pre-treatment crime rates show the yearly average number of crimes an individual is suspected of in the period two, four or eight years before the consultation date. Individuals in our estimation sample are, on average, suspected of around 0.3 – 0.4 crimes a year, depending on the period. A total of 52% of all youths in our sample were suspected of a crime at least once before the time of consultation, 53% in the intention-to-treat group and 50% in the other group (not shown in the table). The sample statistics reveal that both groups are quite similar on observed characteristics. We observe no statistically significant differences with respect to gender, age, educational attainment, or number of suspected crimes. Hence, the time windows we exploit for identifying the effects of the NSP are not (largely) correlated with observed characteristics of individuals.

Table 4.1 Descriptive statistics by intention-to-treat status

	Intention-to-treat = 1	Intention-to-treat = 0	*p*-Value
Covariates			
Gender (male = 1)	0.58	0.65	0.16
Age	20.59	20.63	0.80
Country of birth			0.73
The Netherlands	0.77	0.71	
Morocco	0.05	0.04	
The Antilles	0.08	0.10	
Surinam	0.04	0.06	
Other	0.06	0.08	
Latest educational position			0.31
Primary education	0.01	0.00	
Practical education	0.00	0.01	
Special education	0.04	0.05	
Pre-vocational secondary education	0.04	0.07	
Intermediate vocational education (level 1)	0.26	0.19	
Intermediate vocational education (level 2)	0.42	0.44	
Intermediate vocational education (level 3)	0.08	0.11	
Intermediate vocational education (level 4)	0.08	0.08	
Higher secondary education	0.00	0.01	
Unknown	0.03	0.05	
Highest completed education level			0.25
Primary education	0.74	0.63	
Pre-vocational secondary education	0.12	0.20	
Intermediate vocational education (level 1)	0.07	0.06	
Intermediate vocational education (level 2)	0.04	0.04	
Intermediate vocational education (level 3)	0.00	0.00	
Intermediate vocational education (level 4)	0.00	0.01	
Higher secondary education	0.00	0.01	
Unknown	0.03	0.05	
Pre-treatment crime rate (2-year period)*	0.39	0.44	0.55
Pre-treatment crime rate (4-year period)*	0.44	0.41	0.78
Pre-treatment crime rate (8-year period)*	0.35	0.31	0.49
Outcome variables			
Educational and/or labour market position	0.30	0.28	0.71
Crime rate (after consultation)*	0.33	0.29	0.55
Total number of observations	227	156	

*Notes.** Crime rates are the average yearly number of crimes an individual is suspected of during the relevant period.

The bottom panel of Table 4.1 gives a first impression of the potential effects on outcomes. We observe no significant differences between the groups with respect to either educational and/or labour market position and crime rates after the consultation at the YPO. The fraction of individuals in our sample who are either enrolled in school or employed on 1 November 2012 equals 29%. The average crime rate after the consultation date equals 0.31, which is slightly below the average pre-treatment crime rate.

As discussed above, the selective assignment of youths to either the NSP or a regular program is a concern. To provide insight into this issue, we also compare observable characteristics by assignment status. Table 4.2 shows the sample statistics for both the group of youths actually assigned to the NSP and those assigned to the treatment as usual (a regular reintegration program).

The observable characteristics of the two groups are quite similar. Both groups do not differ significantly with respect to gender, age, country of birth, or educational attainment. With respect to criminal activity, however, we observe that youths with a higher eight-year period pre-treatment crime rate are more frequently assigned to the NSP. This may suggest selective assignment.

In our analyses (see Section 4.5) we control for all covariates. With respect to criminal behaviour, our main analyses include the eight-year pre-treatment crime rate. Although including a shorter period might yield a better indicator of an individual's behaviour at the time of entry at the YPO, we believe that including a longer period is most informative. In addition, this is the only covariate that differs significantly between those assigned to the NSP and those assigned to a regular program.

Table 4.A.1 in the Appendix presents the correlations between the most important variables in our analyses. Being male is positively correlated with both pre- and post-treatment crime rates. The pre-treatment crime rate is strongly correlated with the post-treatment crime rate and negatively correlated with an educational or labour market position. The other variables are less strongly correlated.

Table 4.2 Descriptive statistics by actual assignment

	Assigned to the NSP	Assigned to regular program	*p*-Value
Covariates			
Gender (male = 1)	0.61	0.60	0.77
Age	20.53	20.68	0.36
Country of birth			0.40
The Netherlands	0.77	0.72	
Morocco	0.05	0.04	
The Antilles	0.08	0.10	
Surinam	0.04	0.05	
Other	0.05	0.10	
Latest educational position			0.38
Primary education	0.01	0.01	
Practical education	0.02	0.00	
Special education	0.05	0.04	
Pre-vocational secondary education	0.05	0.06	
Intermediate vocational education (level 1)	0.27	0.19	
Intermediate vocational education (level 2)	0.42	0.44	
Intermediate vocational education (level 3)	0.08	0.11	
Intermediate vocational education (level 4)	0.07	0.10	
Higher secondary education	0.00	0.01	
Unknown	0.03	0.05	
Highest completed education level			0.25
Primary education	0.75	0.63	
Pre-vocational secondary education	0.12	0.19	
Intermediate vocational education (level 1)	0.06	0.07	
Intermediate vocational education (level 2)	0.04	0.05	
Intermediate vocational education (level 3)	0.00	0.00	
Intermediate vocational education (level 4)	0.00	0.01	
Higher secondary education	0.00	0.01	
Unknown	0.03	0.05	
Pre-treatment crime rate (2-year period)*	0.44	0.38	0.51
Pre-treatment crime rate (4-year period)*	0.48	0.36	0.17
Pre-treatment crime rate (8-year period)*	0.39	0.27	0.04
Outcome variables			
Educational and/or labour market position	0.27	0.32	0.30
Crime rate (after consultation)*	0.36	0.26	0.13
Total number of observations	204	179	

*Notes.** Crime rates are the average yearly number of crimes an individual is suspected of during the relevant period.

4.4 Empirical strategy

For estimating the effect of the NSP we exploit variation in participation in the program induced by an assignment rule based on time windows. We start by estimating the impact of assignment to the NSP with the following equation:

$$Y_i = \beta_o + \beta_1 D_i + \beta_2 X_i + \mathrm{S}\ (T_i\ )+ \varepsilon_\mathrm{i}\ , \tag{4.1}$$

where Y_i is the outcome variable of interest of individual i; X_i is a vector of individual background characteristics, including gender, age, age squared, country of birth, most recent education level, highest completed education level, and eight-year pre-treatment crime rate; D_i is a dummy variable that indicates whether individual i was assigned to the NSP; S (T_i) is a smooth polynomial function of the month of entry, and ε_i is the error term. The coefficient of interest is β_1. Since our counterfactual is assignment to a regular reintegration program, the estimated coefficient β_1 should be interpreted as the effect of assignment to the NSP compared to assignment to a regularly available program. We use two main outcome variables: a dummy variable indicating whether the individual is enrolled in education and/or has a job on 1 November 2012 and the crime rate in the period between the consultation date and 1 November 2012.

A concern with estimating Equation (4.1) is non-compliance with the intended treatment by the YPO employees responsible for the assignment of individuals. Since estimation with ordinary-least squares (OLS) can yield inconsistent results because of selective assignment, we use an instrumental variables (IV) approach. To address the endogeneity problem, we instrument assignment by intended treatment and estimate equation (1) with two-stage-least squares (2SLS). The first stage equation, in which the assignment is regressed on the intended treatment and covariates, is

$$D_i = \gamma_o + \gamma_1 Z_i + \gamma_2 X_i + \mathrm{S}\ (T_i\ )+ \mathrm{u_i}, \tag{4.2}$$

where Z_i is a dummy variable indicating the intention-to-treat status. This variable takes the value one if individual i had a consultation at the YPO between 1 August 2009 and 1 March 2010 or between 15 April 2010 and 1 July 2010 and zero if individual i had a consultation in the remaining periods, between 1 March 2010 and 15 April 2010 or between 1 July 2010 and 15 September 2010.

We identify the treatment effect on the assumption that the intended treatment (Z_i) is not correlated with the outcome variable (Y_i), conditional on covariates. Our model controls for smooth polynomials of the month of entry, which pick up potential continuous time effects on outcomes. Since variation in our instrument is due only to differences in the time of entry, it seems plausible that the instrument is not correlated with unobservables in the error term, conditional on the continuous function of the month of entry. Hence, the identifying assumption is that there are no other discontinuous effects of the time of entry on the outcomes.

A potential threat to the validity of our instrument might be the manipulation of consultation dates. Since we are using administrative data of the YPO, an independent agency with no particular interest in the relative performance of the NSP compared to other reintegration programs, manipulation of consultation dates seems not very plausible. Moreover, an inspection of the number of intakes around the cut-off dates does not reveal any suspicious patterns. We observe no increase or decrease in the number of consultations during these periods. It should also be noted that we observe non-compliance with the assignment rule. If the YPO would manipulate the dates this seems less likely.

If the standard IV-assumptions hold, estimation with 2SLS yields a local average treatment effect that can be interpreted as the average treatment effect of assignment to the NSP for the subpopulation of compliers (Imbens and Angrist, 1994). Compliers are individuals whose treatment status is affected by the instrument. In our case, these are youths whose assignment to the NSP is completely determined by the intended treatment.

It should be noted that our analysis focuses on the causal impact of assignment to the NSP, which from a policy perspective also seems the most relevant treatment variable. We instrument assignment to the NSP with the intended assignment to the NSP. This differs from the standard approach dealing with non-compliance to treatment in which actual treatment is instrumented with assignment to treatment. Due to the policy implementation of the NSP we cannot estimate the causal effect of participation in the NSP by using intended assignment to the NSP as an instrumental variable. We are concerned that the exclusion restriction will not hold as it seems not likely that never-takers and always-takers (with respect to participation) in the treatment and control group will receive equal treatment. The YPO offers only one intervention to individuals in the treatment and in the control group. Individuals that are assigned to the NSP but who refuse to participate are not allowed to start in other programs. Similar youths in the control group, that would have refused to participate in case of assignment to the NSP, can either start in a regular program or not start in any program at all. Hence, the instrument not only affects participation in the NSP but also affects the subsample of individuals who refuse to participate when assigned to NSP. We observe that 51 % of the youths assigned to the NSP did not start, whereas 35 % of the youths in the control group did not start in a program. This indicates that part of the youths in the control group that would have refused to start in case of assignment to the NSP, accept to start in a regular program. It follows that never-takers in the treatment and control group do not receive equal treatment and the exclusion restriction will not hold. This implies that we can only estimate the effect of assignment to the NSP, which is driven by both the effect of participation in the NSP and the effect of a different probability of starting in a program.

4.5 Estimated effects on schooling, labour market and crime outcomes

4.5.1 Main results

This section shows the main estimation results of the effect of assignment to the NSP on the educational and/or labour market position and on the criminal activity. In Table 4.3 we present the OLS and IV estimates using six model specifications. The first model regresses the outcome variable on a constant, a dummy variable for assignment to the NSP, and a linear function of the month of entry at the YPO. Model (2) additionally controls for gender, age, age squared, and country of birth. Model (3) adds the most recent educational position and the highest completed education level, while model (4) also includes the crime rate during the eight-year period before the consultation date at the YPO.[9] In models (5) and (6) the linear function of the month of entry is extended to quadratic and cubic polynomials, respectively.[10] The top panel of Table 4.3 presents the estimated effects on a dummy variable for school enrolment and/or employment on 1 November 2012, while the middle panel reports the estimated effects on the number of crimes an individual is suspected of during the period between the consultation date and 1 November 2012. In addition, the bottom panel of Table 4.3 shows the estimates of the first-stage regression, in which assignment to the NSP is regressed on the intended treatment. Each estimate is based

[9] Including the number of crimes during the shorter two- or four-year period instead of the eight-year period before the consultation date yields similar results.

[10] Alternative specifications in which we include continuous functions of the day or week of entry rather than the month of entry yield similar estimation results.

on a separate regression and robust standard errors corrected for clustering at the program/location level are shown in parentheses.[11]

The OLS regressions show that assignment to the NSP is not associated with the educational or labour market position. All estimates are statistically insignificant and point estimates are slightly below zero. Since the OLS estimates may suffer from endogeneity, we proceed with IV estimates. The IV estimates are consistent with the OLS findings. In all specifications the effect of assignment to the NSP does not significantly differ from zero. Hence, we find no evidence that assignment to the NSP affects school enrolment or employment probabilities.

With respect to criminal activity, the OLS regressions yield positive but statistically insignificant effects. The IV regressions also show positive point estimates. In the full model, including higher- order polynomials of the time of entry, we find marginally significant positive effects of assignment to the NSP. This suggests that assignment to the NSP leads to an increase in criminal activity. To test the sensitivity of this result, we estimate the full-model IV specification on a sample from which we excluded individuals with the largest number of pre-treatment crimes (all individuals suspected of 15 crimes or more in the eight-year period before the consultation date, 5 % of the sample). This yields an estimate of 0.14 that is statistically significant at the 10 %-level (not shown in Table 4.3), which suggests that the result is not driven by a small subgroup. The inclusion of a higher-order term of the month of entry substantially affects the estimated effect on the crime rate. This indicates that there are unobserved differences in inflow over time, that are picked up by the smooth function of the month of entry.[12] The first-stage regression determines the effect of the intention-to-treat on assignment. The estimates are highly significant, with *F*-values above 93 in all our models. This implies that our analyses do not suffer from a weak instrument problem (Staiger and Stock, 1997).

[11] The total number of clusters equals 29, which are the programs youths are referred to after their consultations. We distinguish 27 different regular reintegration programs and two locations within the NSP, in the north and south of Rotterdam.

[12] Table 4.A.2 in the Appendix presents additional estimation results for the specifications (1)-(4), including a third order polynomial of the month of entry. Adding covariates does not affect the estimated effects.

Table 4.3 Estimates of the effect of assignment to the NSP on schooling, labour market and crime

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: Educational/labour market position					
OLS						
Assignment to the NSP	-0.040 (0.067)	-0.051 (0.049)	-0.039 (0.053)	-0.029 (0.051)	-0.023 (0.041)	-0.023 (0.041)
IV (second stage)						
Assignment to the NSP	0.032 (0.082)	0.010 (0.073)	0.024 (0.088)	0.032 (0.083)	0.052 (0.076)	0.052 (0.073)
	Dependent variable: Crime rate					
OLS						
Assignment to the NSP	0.069 (0.077)	0.083 (0.060)	0.081 (0.055)	0.041 (0.044)	0.101 (0.060)	0.101 (0.061)
IV (second stage)						
Assignment to the NSP	0.031 (0.108)	0.083 (0.088)	0.099 (0.075)	0.068 (0.053)	0.165* (0.084)	0.163* (0.090)
	Dependent variable: Assignment to the NSP					
First-stage						
Intended assignment	0.723*** (0.067)	0.721*** (0.064)	0.721*** (0.065)	0.719*** (0.066)	0.686*** (0.071)	0.692*** (0.070)
F-test for instrument	116.45	126.91	123.04	118.68	93.35	97.73
Controls						
SES	no	yes	yes	yes	yes	yes
Educational level	no	no	yes	yes	yes	yes
Pre-treatment crime	no	no	no	yes	yes	yes
Polynomial control for month of entry	linear	linear	linear	linear	quadratic	cubic
Observations	383	383	383	383	383	383

Notes. Robust standard errors are in parentheses. Asterisks indicate that the estimates are statistically significant at the *** 1% level and *10% level. The SES control variables include gender, age, age squared, and country of birth. The educational level controls include the latest educational position and the highest completed education level. Pre-treatment crime is the yearly average number of crimes an individual is suspected of during the eight-year period before the consultation date.

4.5.2 Heterogeneity

We now investigate the impact of assignment to the NSP for subgroups. We examine whether the effects differ between suspected and non-suspected youths, men and women, older and younger individuals, and relatively lower- and higher-educated youths. A suspected individual is defined as an individual who has been suspected of a crime at least once during the eight-year period before entry at the YPO. A non-suspected individual was never suspected of a crime during this period. We define individuals aged 20.6 years (the average age in the estimation sample) or older as old and individuals aged under 20.6 years as young. Lower-educated youths are those who have only completed primary education and higher-educated youths have obtained a secondary or higher educational degree.[13]

Table 4.4 presents the full-model IV estimation results of the effects of assignment to the NSP for these subpopulations (compare model (6) in Table 4.3). The second-stage results show that assignment to the NSP has no significant effects on educational or labour market position for all subpopulations. With respect to the number of crimes, we find a statistically significant positive effect, 0.36, for suspected individuals, while the point estimate for non-suspected individuals is close to zero.

This finding implies that assignment to the NSP increases crime rates during the period between the consultation date and 1 November 2012, with 0.36 for the (compliant subpopulation of the) sample of youths suspected of a crime when they entered the program. We also find positive effects for the male subpopulation, while the effects for females are close to zero. Being suspected of a crime is strongly correlated with gender: 75% of the suspected subpopulation is male. In addition, statistically significant effects for the younger subpopulation are found (but the difference between estimated effects for young and old individuals is not very large).

These findings indicate that the marginally significant effects of assignment to the NSP on criminal activity are driven by the effects on the subpopulation of suspected individuals. Table 4.A.3 in the Appendix presents all OLS and IV estimates for the subsample of suspected youths. Positive effects on crime rates are also found in the IV models that do not include a higher-order polynomial of the month of entry and in the OLS models. This provides additional evidence that assignment to the NSP increases criminal activity among youths who have been suspected of criminal activities in the past.

[13] We have chosen this definition of higher education because 70% of youths in our estimation sample completed only primary education.

Table 4.4 IV estimates of the effect of assignment to the NSP on educational/labour market position and crime for subgroups

	(1) Suspected	Not suspected	(2) Male	Female	(3) Old	Young	(4) Lower-educated	Higher-educated
	Dependent variable: Educational/labour market position							
IV (second stage)								
Assignment to the NSP	-0.002 (0.094)	0.135 (0.107)	0.161 (0.134)	-0.099 (0.138)	-0.024 (0.093)	0.097 (0.081)	-0.050 (0.070)	0.214 (0.167)
	Dependent variable: Crime rate							
IV (second stage)								
Assignment to the NSP	0.355** (0.127)	0.008 (0.027)	0.290** (0.118)	-0.005 (0.068)	0.124 (0.074)	0.252** (0.120)	0.117 (0.086)	0.281 (0.222)
	Dependent variable: Assignment to the NSP							
First-stage								
Intended treatment	0.600*** (0.077)	0.732*** (0.104)	0.721*** (0.054)	0.594*** (0.129)	0.695*** (0.084)	0.678*** (0.071)	0.698*** (0.056)	0.687*** (0.122)
F-test for instrument	60.72	49.54	178.27	21.20	68.46	91.19	155.35	31.71
Controls								
SES	yes	yes	yes	yes	yes	yes	yes	yes
Educational level	yes	yes	yes	yes	yes	yes	yes	yes
Pretreatment crime	yes	yes	yes	yes	yes	yes	yes	yes
Polynomial control for month of entry	cubic	cubic	cubic	cubic	cubic	cubic	cubic	cubic
Observations	199	184	232	151	194	189	266	117

Notes. Robust standard errors are in parentheses. Asterisks indicate that the estimates are statistically significant at the *** 1% level and **5% level. The SES control variables include gender, age, age squared, and country of birth. The educational level controls include the latest educational position and the highest completed education level. Pre-treatment crime refers to a variable indicating the yearly average number of crimes an individual is suspected of during the eight-year period before the consultation date.

4.5.3 Alternative outcome measures

In addition to our two main outcomes, we construct six related outcome variables: (1) a dummy variable that indicates whether an individual is enrolled in education on 1 November 2012; (2) a dummy variable that indicates whether an individual has work on 1 November 2012; (3) a dummy variable that indicates whether an individual has an educational and/or labour market position on 1 November 2012 or obtained an educational degree between the consultation date and 1 November 2012; (4) a dummy variable that indicates whether an individual has ever been in education or worked on either of the four reference dates (1 April 2011, 1 October 2011, 1 May 2012, and 1 November 2012); (5) the crime rate during a one-year period after the consultation date; and (6) a dummy variable that indicates whether an individual is suspected of at least one crime in the period between the consultation date and 1 November 2012. Table 4.5 presents the full-model (second-stage) IV estimates of the effect of assignment to the NSP for these six outcome variables. Each column corresponds to one outcome variable.

Models (1) to (4) are related to educational and employment outcomes. Since the program is relatively more focused on educational outflow, one might hypothesize that it affected school enrolment more than employment. We find no evidence that the NSP specifically affected any one of these outcomes; both estimated effects on school enrolment and employment are statistically insignificant. Furthermore, we investigate the effects on an extended measure of educational and/or labour market success by including youths who have obtained a formal educational degree. In addition to educational enrolment, we received data on educational completion from the YPO. Five persons completed intermediate vocational educational before the latest reference date, two of whom assigned to the NSP and three to a regular program. Including these 'successfully treated youths' in our outcome measure does not affect estimation results. Finally, we assess potential short-run effects by looking at educational or labour market position on at least one of the four reference dates. Previous studies on youth programs found that positive effects can fade away in the longer term (e.g., Schochet et al., 2008; Rodriguez-Planas, 2012). If assignment to the NSP initially increases school enrolment or employment, we expect to see this on the earlier reference dates. A complicating factor is that not all youths moved out of their programs on the earlier reference dates. Variation in the start dates and program durations of youths cause differences in outflow between both groups. Because of the longer program durations in the NSP, the fraction of youths who left the program is larger in the control group for each of the reference dates. Since youths still in their programs are, by definition, not enrolled in education or employed, the analysis seems to provide a lower bound on the NSP's potential short-run effects. The negative, statistically insignificant point estimate does not provide evidence of short-run gains.

Models (5) and (6) are related to criminal behaviour. Previous studies found that youth programs in a residential setting reduced crimes during the treatment period but the effects did not persist (e.g., Sochet et al., 2001; Sochet et al., 2008; Millenky et al., 2011). To shed light on potential short-run reductions in criminal activity, we estimate the effects on crime rates only during the first year after the consultation date. The positive (but statistically insignificant) point estimate of 0.11 suggests that criminal activity was not reduced during the treatment period.

Table 4.5 Estimates of the effect of assignment to the NSP on six alternative outcome measures

	Educational and/or labour market position				Criminal behaviour	
	(1) Educational position on the latest reference date	(2) Labour market position on the latest reference date	(3) Educational/ labour market position or educational degree on the latest reference date	(4) Educational/ labour market position on at least one of the four reference dates	(5) Crime rate in one-year period after the consultation date	(6) Suspected of at least one crime after the consultation date
IV (second stage)						
Assignment to the NSP	0.018 (0.062)	0.016 (0.046)	0.052 (0.078)	-0.061 (0.073)	0.115 (0.156)	0.017 (0.084)
Controls						
SES	yes	yes	yes	yes	yes	yes
Educational level	yes	yes	yes	yes	yes	yes
Pre-treatment crime	yes	yes	yes	yes	yes	yes
Polynomial control for month of entry	cubic	cubic	cubic	cubic	cubic	cubic
Observations	383	383	383	383	383	383

Notes. Robust standard errors are in parentheses. The SES control variables include gender, age, age squared, and country of birth. The educational level controls include the latest educational position and the highest completed education level. Pre-treatment crime refers to a variable indicating the yearly average number of crimes an individual is suspected of during the eight-year period before the consultation date.

If anything, a comparison with the estimated effect of 0.16 on criminal activity in the whole period after the consultation date indicates that the positive impact of assignment to the NSP on crime rates increases over time. We investigate these timing issues further in Section 4.6.

The impact on crime rates may be driven by effects on both the number of criminally active youths and the number of crimes among those already criminally active. The strong effect among the group of suspected individuals at the start of the program suggests that the latter is most prominent. We estimate the effect on a dummy variable for being suspected of a crime after the consultation date and find an insignificant point estimate close to zero. This finding implies that assignment to the NSP does not increase the number of criminally active youths and supports our finding that the overall increase in crime rates is driven by an increase in crimes among those who were already criminally active.

4.6 Mechanisms

Our estimation results indicate that assignment to the NSP increases criminal activity, especially for the subpopulation of suspected individuals. We now examine two mechanisms that may explain our findings. First, we try to distinguish between criminal behaviour during program participation and after leaving the program. In our main analyses, we estimate the effects on crime rates after the consultation date, which includes both the active treatment period and the period after leaving the program. Criminal activity can differ between these periods. During the treatment criminal behaviour can, for example, be hampered by the supervision of the coaches or the program's requirements. This can also lead to postponed criminal activity after leaving the program. On the other hand, if youths feel protected by their coaches when they get into trouble, they may be less deterred by potential sanctions and more inclined to risky criminal behaviour. A distinction in the crime rates provides insight into the timing of criminal activities.

Second, we investigate whether peer group effects can explain our findings. Within each of the two neighbourhood schools, around 100 individuals are treated simultaneously. This can provide the opportunity to learn from each other or to motivate each other with respect to criminal activities. If individuals are influenced by their peer group, social interaction between youths within the neighbourhood school can affect criminal outcomes. Previous studies point out the importance of peer group effects. Glaeser et al. (1996) find that social interaction in criminal behaviour is an important factor in explaining the variation in crime rates across US cities. Dishion et al. (1999) conclude that interventions for at-risk adolescents delivered in peer groups can reinforce delinquency and other kinds of problem behaviour. High-risk youths seem particularly vulnerable to peer group pressure. Dodge et al. (2007) review the literature on deviant peer influences and state that group-based interventions can increase problem behaviour, since participants can learn from interaction with their deviant peers.

To investigate these issues further, we have to restrict our estimation sample to youths who actually started in a program. For non-starters it is not possible to distinguish between the program and post-program period or to investigate the influence of program peers. The restricted sample of starters consists of 217 individuals, of whom 100 participated in the NSP and 117 in one of the regular reintegration programs. Table 4.A.4 in the Appendix shows the descriptive statistics of this selective subpopulation of starters. Both groups within this sample are quite similar on most observable characteristics, such as gender, educational attainment and criminal activity. We only observe a significant difference with respect to age: Starters in the NSP are younger than starters in a regular program.[14] Table 4.6 shows the

[14] Table 4.A.5 in the Appendix also reports OLS estimates of the effect of starting in the NSP compared to starting in a regular program on the educational/labour market position and crime rate. As discussed in Section 4.4, the

average crime rates during and after the program for both groups of starters, with standard deviations in parentheses.[15]

The crime rates are roughly equal during the program, but after the program we observe a remarkable difference: Crime rates in the NSP stay at the same level, whereas crime rates in the regular programs decrease.[16] This pattern does not suggest a reduction in crime rates during the program or the postponement of criminal activities.

Table 4.6 Crime rates during and after program participation

	starters in the NSP	starters in the control group
Crime rate during the program	0.44	0.41
	(0.12)	(0.16)
Crime rate after leaving the program	0.38	0.21
	(0.13)	(0.04)
Number of observations	100	117

Notes. The standard deviations of the average crime rates are in parentheses.

Table 4.7 shows the OLS estimates of models in which the crime rate during the program (left panel) and after the program (right panel) are regressed on a dummy variable for starting in the NSP and all the covariates. Since the impact of assignment to the NSP on criminal activity is mainly driven by the group of suspected individuals, we also perform our analyses on this subpopulation. Columns (1) and (2) show the estimated effects on crime rates during the program for the full sample and the subpopulation of

estimated impact of assignment to the NSP captures both the effect of participation in the NSP and the effect of a difference in starting probabilities. Although we cannot give a causal interpretation to the results the estimated effects suggest that starting in the NSP does not increase school enrolment or employment probabilities.

[15] This analysis used the actual start date to construct program duration rather than the consultation date. There can be slight differences between both dates, since it takes some time before an individual actually starts the program to which he or she is referred. Our main analyses use the consultation date, since non-starters do not have a start date.

[16] The differences in crime rates presented in Table 4.6 are not statistically significant. The *p*-value of the difference in crime rates after leaving the program equals 0.21.

suspected criminals. Similarly, columns (3) and (4) present the estimated effects on the post-program crime rates for both samples.

Whereas we find no significant effects on the crime rate during the program, we find that starting in the NSP has a positive and statistically significant effect at the 1% level on the crime rate after program exit. In the complete sample the estimated effect is 0.25. When restricting the sample to suspected individuals, we find an even larger effect, 0.43. These findings suggest that the positive effect of assignment to the NSP on criminal activity is mostly driven by differences in post-program criminal behaviour. Whereas we observe a drop in criminal activity after an individual leaves a regular reintegration program, crime rates for starters in the NSP stay roughly at the same level. This finding is in line with the increasing impact of the NSP on crime rates over time (see Table 4.5, column (5) and Table 4.3, column (6)).

Table 4.7 The effect of starting in the NSP (compared to starting in a control program) on the crime rates during and after the program

	Effects on crime rate during program		**Effects on crime rate after the program**	
	(1) Complete sample	(2) Subsample of suspected individuals	(3) Complete sample	(4) Subsample of suspected individuals
OLS				
Starting in the NSP	-0.065 (0.116)	-0.132 (0.268)	0.254*** (0.050)	0.426*** (0.127)
Controls				
SES	yes	yes	yes	yes
Educational level	yes	yes	yes	yes
Pre-treatment crime	yes	yes	yes	yes
Polynomial control for month of entry	cubic	cubic	cubic	cubic
Observations	217	121	217	121

Notes. Robust standard errors are in parentheses. Asterisks indicate that the estimates are statistically significant at the *** 1% level. The SES control variables include gender, age, age squared and country of birth. The educational level controls include the latest educational position and the highest completed education level. Pre-treatment crime refers to a variable indicating the yearly average number of crimes an individual is suspected of during the eight-year period before the consultation date.

Potentially positive influences of program employees or coaches may be mitigated in the NSP by peer group effects. To investigate the existence of peer effects, we try to exploit the variation in pre-treatment criminal activity across the two locations of the NSP (North or South). Individuals were assigned to the location based on their place of residence. Since interaction takes place within a location, we consider the subsample of youths at the specific location as the relevant peer group. We remove four individuals who were assigned twice from our sample, since they were referred to the NSP after the assignment period and started the program later. This leaves us with 96 starters during the assignment period. For the location South, the average pre-treatment crime rate (0.36) is 29% higher compared to that at North (0.28). Hence, an individual assigned to location South is more likely to face peers with criminal experience than an individual sent to location North.[17]

We now examine the existence of crime-related peer effects by estimating the effect of the average pre-treatment crime rate in the peer group on the individual crime rate after the consultation date, conditional on individual pre-treatment crime rate and all the other covariates. The left panel of Table 4.8 shows the OLS estimates of the effects of both individual and average pre-treatment crime rates. Model (1) includes all 96 starters in the NSP; model (2) restricts the sample to the subpopulation of suspected individuals. If peer effects related to criminal behaviour exist, we expect to find positive effects for the average crime rate, especially in the subsample of suspected individuals. Table 4.8 shows that the estimates of the average pre-treatment crime rate are positive. When we only take into account suspected individuals, we find much larger and statistically significant estimates. This may provide evidence for the existence of crime-related peer effects. Nevertheless, since we only use variation between locations, we cannot rule out that our measure for peer effects picks up other unobservable differences between both locations. If our findings truly reveal crime-related peer effects, we expect our measure to be less clearly correlated with educational and/or labour market outcomes. If they reflect other differences, it seems more plausible that our measure is correlated with other outcomes as well.

The right panel (models (3) and (4)) presents the estimation results of equivalent OLS models, in which a dummy variable for educational and/or labour market position is used as the outcome variable. We find insignificant effects of the average pre-treatment crime rate on educational and/or labour market position. Model (3) including all starters yields a positive point estimate, whereas model (4) including only suspected individuals yields a negative point estimate. Hence, the average pre-treatment crime rate seems to be related to individual criminal activity but not to individual success with respect to education or work.

[17] We also compared the fraction of suspected individuals at the start of the program, which turns out to be similar in both sites. Hence, the difference in pre-treatment criminal activity is caused by higher crime rates among suspected individuals in location South.

Table 4.8 The effect of the average crime rate of peers on individual outcomes (starters in the NSP)

	Dependent variable: Crime rate		**Dependent variable: Educational and/or labour market position**	
	(1) Complete sample	(2) Subsample of suspected individuals	(3) Complete sample	(4) Subsample of suspected individuals
OLS				
Individual crime rate	0.804*** (0.118)	0.915*** (0.176)	-0.273*** (0.098)	-0.216* (0.125)
Average crime rate in peer group	1.563 (1.457)	5.089** (2.453)	0.639 (1.205)	-0.729 (1.741)
Controls				
SES	yes	yes	yes	yes
Educational level	yes	yes	yes	yes
Pre-treatment crime	yes	yes	yes	yes
Polynomial control for month of entry	cubic	cubic	cubic	cubic
Observations	96	54	96	54

Notes. Standard errors are in parentheses. Asterisks indicate that the estimates are statistically significant at the *** 1% level, **5% level and *10% level. The SES control variables include gender, age, age squared, and country of birth. The educational level controls include the latest educational position and the highest completed education level. Pre-treatment crime refers to a variable indicating the yearly average number of crimes an individual is suspected of during the eight-year period before the consultation date.

This analysis does not allow a causal interpretation because the youths were not randomly assigned to the sites and we cannot fully rule out the possibility that the estimated effects of average pre-treatment crime rates reflect other differences between the two locations. Nevertheless, our results are consistent with crime-related peer effects and we interpret them as suggestive evidence that peer effects appear to explain the positive effect of assignment to the NSP on criminal activity.

4.7 Conclusions

The NSP aims to increase school enrolment and employment among multi-problem school dropouts. Previous studies provide mixed evidence on the success of training programs for disadvantaged youths. It seems therefore difficult to effectively serve the vulnerable target group of at-risk adolescents and realize persistent social gains. The NSP is largely designed in line with lessons from the literature and shares several components with other promising programs. It offers an intensive and integrated approach of educational and work services combined with professional care and personal mentoring. We evaluate the effects of the program by implementing a specific assignment rule that ensures that treatment status is determined by an individual's application date. The effects of assignment to the NSP are estimated three years after the start of the program.

We find no evidence that assignment to the NSP improves school enrolment or employment probabilities compared to the regular treatment. This finding is in line with a large body of the literature that shows no impact of training programs for at-risk youths on labour market outcomes (e.g., Carneiro and Heckman, 2003; LaLonde, 2003). In addition, we find evidence that assignment to the NSP increases criminal activity, especially among the subpopulation of youths suspected of a crime at the time of entry. This result is consistent with previous studies that document adverse effects of group counselling or group-based interventions on criminal activity (Dishion et al., 1999). Deviant peer effects caused by grouping at-risk adolescents together can explain the reinforcement of criminal behaviour (Dodge et al., 2007). Additional analyses provide suggestive evidence in line with this explanation. Hence, adverse peer effects due to placing at-risk youths together may have mitigated other promising elements of the NSP.

Appendix 4.A: Additional analyses

Table 4.A.1 Correlation table

	Male	Age	Low education level	Pre-treatment crime rate (8-year period)	Post-treatment crime rate	Educational/labour market position
Male	1.000					
Age	-0.138	1.000				
Low education level	0.115	-0.073	1.000			
Pre-treatment crime rate	0.308	0.001	0.157	1.000		
Post-treatment crime rate	0.287	-0.043	0.085	0.508	1.000	
Educational/labour market position	-0.174	-0.057	-0.134	-0.221	-0.209	1.000

*Notes.*Youths with a low education level have primary education as their highest completed education level. The pre-treatment crime rate is the average number of crimes an individual is suspected of in the eight-year period before consultation at the YPO. The post-treatment crime rate is the average number of crimes an individual is suspected of during the period between consultation at the YPO and 1 November 2012.

Table 4.A.2 OLS and IV estimates of the effects on crime rates (all models including a cubic polynomial of the month of entry)

	(1)	(2)	(3)	(4)
OLS				
Assignment to the NSP	0.145* (0.081)	0.143** (0.068)	0.139** (0.065)	0.101 (0.061)
IV (second stage)				
Assignment to the NSP	0.155 (0.137)	0.179 (0.115)	0.193* (0.106)	0.163* (0.090)
Controls				
SES	no	yes	yes	yes
Educational level	no	no	yes	yes
Pre-treatment crime	no	no	no	yes
Polynomial control for month of entry	cubic	cubic	cubic	cubic
Observations	383	383	383	383

Notes. Robust standard errors are in parentheses. Asterisks indicate that the estimates are statistically significant at the ** 5% level and *10% level. The SES control variables include gender, age, age squared, and country of birth. The educational level controls include the latest educational position and the highest completed education level. Pre-treatment crime refers to a variable indicating the yearly average number of crimes an individual is suspected of during the eight-year period before the consultation date.

Table 4.A.3 Subpopulation of suspected individuals: OLS and IV estimates of the effects on crime rates

	(1)	(2)	(3)	(4)	(5)	(6)
OLS						
Assignment to the NSP	0.113 (0.122)	0.170* (0.093)	0.185* (0.092)	0.122 (0.085)	0.202** (0.089)	0.201** (0.089)
IV (second stage)						
Assignment to the NSP	0.012 (0.179)	0.139 (0.145)	0.221* (0.114)	0.187* (0.101)	0.355** (0.128)	0.355** (0.127)
Controls						
SES	no	yes	yes	yes	yes	yes
Educational level	no	no	yes	yes	yes	yes
Pre-treatment crime	no	no	no	yes	yes	yes
Polynomial control for month of entry	linear	linear	linear	linear	quadratic	cubic
Observations	199	199	199	199	199	199

Notes. Robust standard errors are in parentheses. Asterisks indicate that the estimates are statistically significant at the ** 5% level and *10% level. The SES control variables include gender, age, age squared, and country of birth. The educational level controls include the latest educational position and the highest completed education level. Pre-treatment crime refers to a variable indicating the yearly average number of crimes an individual is suspected of during the eight-year period before the consultation date.

Table 4.A.4 Descriptive statistics: Comparison of starters in the NSP and regular programs

	Started in the NSP	Started in a regular program	*p*-Value
Covariates			
Gender (male = 1)	0.65	0.56	0.20
Age	20.20	20.75	0.01
Country of birth			0.80
The Netherlands	0.74	0.68	
Morocco	0.06	0.05	
The Antilles	0.08	0.12	
Surinam	0.06	0.06	
Other	0.06	0.09	
Latest educational position			0.29
Primary education	0.00	0.00	
Practical education	0.02	0.00	
Special education	0.04	0.04	
Pre-vocational secondary education	0.04	0.07	
Intermediate vocational education (level 1)	0.30	0.19	
Intermediate vocational education (level 2)	0.45	0.44	
Intermediate vocational education (level 3)	0.07	0.13	
Intermediate vocational education (level 4)	0.05	0.09	
Higher secondary education	0.00	0.01	
Unknown	0.03	0.03	
Highest completed education level			0.45
Primary education	0.77	0.64	
Pre-vocational secondary education	0.10	0.18	
Intermediate vocational education (level 1)	0.05	0.07	
Intermediate vocational education (level 2)	0.05	0.06	
Intermediate vocational education (level 3)	0.00	0.00	
Intermediate vocational education (level 4)	0.00	0.01	
Higher secondary education	0.00	0.01	
Unknown	0.03	0.03	
Pre-treatment crime rate (2-year period)*	0.44	0.42	0.92
Pre-treatment crime rate (4-year period)*	0.42	0.40	0.92
Pre-treatment crime rate (8-year period)*	0.35	0.29	0.47
Outcome variables			
Educational and/or labour market position	0.33	0.33	0.96
Crime rate (after consultation)*	0.35	0.27	0.42
Total number of observations	117	100	

Notes. * Crime rates are the average yearly number of crimes an individual is suspected of during the relevant period.

Table 4.A.5 OLS estimates of the impact of starting in the NSP (compared to starting in a regular program) on educational/labour market position and the crime rate (subsample of starters)

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: Educational/labour market position					
OLS						
Starting in the NSP	0.011	0.002	0.011	0.011	0.014	0.014
	(0.110)	(0.087)	(0.087)	(0.089)	(0.088)	(0.088)
	Dependent variable: Crime rate					
OLS						
Starting in the NSP	0.028	0.022	0.031	0.033	0.080	0.079
	(0.101)	(0.081)	(0.075)	(0.062)	(0.058)	(0.056)
Controls						
SES	no	yes	yes	yes	yes	yes
Educational level	no	no	yes	yes	yes	yes
Number of crimes	no	no	no	yes	yes	yes
Polynomial control for month of entry	linear	linear	linear	linear	quadratic	cubic
Observations	217	217	217	217	217	217

Notes. Robust standard errors are in parentheses. The SES control variables include gender, age, age squared ,and country of birth. The educational level controls include the latest educational position and the highest completed education level. Pretreatment crime refers to a variable indicating the yearly average number of crimes an individual is suspected of during the eight-year period before the consultation date.

5

The impact of a comprehensive school reform policy for failing schools on educational achievement[1]

Abstract

This chapter estimates the effects of a comprehensive school reform program on high-stakes test scores in Amsterdam. The program implements a systematic and performance-based way of working within weakly performing primary schools and integrates measures such as staff coaching, teacher evaluations and teacher schooling, and the use of new instruction methods. Difference-in-differences estimates show substantial negative effects on test scores for pupils in their final year of primary school. The program decreased test scores with 0.17 standard deviations in the first four years after its introduction. A potential explanation for this finding is the intensive and rigorous approach that caused an unstable work climate with increased teacher replacement.

[1] This chapter is based on joint work with Suzanne Kok: Van Elk, R., S. Kok, 2014, The impact of a comprehensive school reform policy for failing school on educational achievement; Results of the first four years, CPB Discussion Paper 264.

5.1 Introduction

The improvement of weakly performing schools is an important issue in many countries. Comprehensive school reform (CSR) methods have been widely used to turn around failing schools.[2] These programs involve integrated changes at all levels within schools rather than incremental changes targeted at single aspects. The school is considered as the level of improvement and a program's content is tailored to its specific needs. [3] CSR models typically include various elements such as professional development of educators, an increased attention to instruction methods and the individual needs of pupils, improvement in classroom or school management, parental involvement, curriculum improvements and setting high achievement goals. Proponents of CSR methods argue that comprehensive changes are needed since the impact of isolated interventions can be distorted by dysfunction in other areas (Borman et al., 2004). However, the effectiveness of CSR programs in increasing student outcomes is by and large unclear.

The purpose of this chapter is to estimate the causal effects of a CSR policy on educational performance. We investigate the impact of the Amsterdam School Improvement Program (ASIP) in the Netherlands, which was introduced in 2008. The goal of the program is to improve the educational quality of failing primary schools in Amsterdam. The ASIP is an intensive two-year program that aims to implement a systematic and performance-based way of working. This includes data-driven teaching with educators that systematically measure pupil performance. This information is used to respond to the individual needs of the pupils. The ASIP uses improvement plans that are fitted to the particular needs of the schools and typically integrate measures such as staff coaching, teacher quality evaluations, teacher schooling, and the use of new instruction methods. The implementation of each improvement plan is guided by an expert team with strong educational experience.

As of April 2008, all primary schools in Amsterdam that were judged to perform below national quality standards were invited to voluntarily participate in the program. To examine the impact of the program we compare the development of pupil achievement in failing schools in Amsterdam to that in failing schools outside Amsterdam. Our assessment of educational achievement is based on the CITO test. This is a nationwide, high-stakes test that pupils take in the highest grade (eight) of primary education. The test includes questions on language, math, and information processing. We make use of administrative data on CITO test scores from 2005 to 2012, which enables us to compare the change in performance in

[2] The U.S. government has devoted over 2 billion dollars to the implementation of CSR programs in the 1990s and early 2000s (U.S. Department of Education, 2004; 2006).

[3] Over 800 variations of CSR models have been implemented in more than 5,000 schools in the US in the past decades (Rowan et al., 2004).

Amsterdam before and after the introduction of the CSR policy with the change in performance in other cities that did not introduce the CSR policy.

This study relates to the literature on the effects of CSR models. Borman et al. (2003) provide an extensive overview of existing studies with respect to 29 of the most widely implemented CSR models in the U.S. Although they find overall promising results, the authors conclude that both quality of research designs and quantity of studies are insufficient to draw strong conclusions on the effectiveness of CSR models. More recent studies show ambiguous effects of CSR models. Whereas some studies find positive effects on student performance (e.g. May and Supovitz, 2006), others find no significant effects (Gross et al., 2009; Bifulco et al., 2005) or mixed results across grades and subjects (Schwartz et al., 2004). While the overall record of CSR models appears to be encouraging, most results come from studies that do not use credible research designs to handle potential selection problems. Schools that adopt a comprehensive school model are likely to differ from other schools in a number of aspects, such as student composition, educational quality, or desire for innovation. Even if one controls for student characteristics or school fixed effects, estimated effects may still be biased because of unobserved heterogeneity.

Our main contribution is that we use a quasi-experimental research design to address potential endogeneity. We estimate difference-in-differences models to identify the effects of the introduction of the CSR policy. Difference-in-differences models have been frequently used in previous economic evaluation studies (see e.g. Ashenfelter and Card, 1985; Card and Krueger, 1994; Blundell et al., 1998; Jacob, 2005). We present intention-to-treat estimates and find substantially negative effects on test scores in the first four years after the introduction of the program. This result is robust to a variety of sensitivity analyses. In our preferred specification the introduction of the ASIP decreases test scores by 0.17 standard deviations. The detrimental impact on test scores is largest for language and is generally larger at the left part of the test score distribution. Interviews with school-leaders of participating schools provide a candidate explanation for our findings. The rigorous and demanding approach appears to have caused an increase in teacher replacement. The resulting loss of school specific knowledge, increase in recruitment and hiring costs, and uncertain work atmosphere felt by teachers may have negatively affected pupil achievement. We cannot exclude that our findings reflect adjustment costs during the transition from a failing to a successful school and that it takes longer before beneficial effects become manifest. In any case, we conclude that the introduction of the ASIP induced major costs in terms of substantial test score losses for at least four cohorts of pupils.

The rest of this chapter is organized as follows. Section 5.2 provides a brief overview of previous studies. Section 5.3 describes the ASIP. Sections 5.4 and 5.5 discuss the empirical strategy and the data. Section 5.6 presents the results and Section 5.7 discusses potential mechanisms that could explain our findings. Section 5.8 concludes.

5.2 Previous studies

The literature on CSR models largely consists of practitioner-oriented studies (see e.g. Herman et al., 1999; Traub, 1999; Slavin and Fashiola, 1998). Although some of these studies provide an assessment of CSR models, none of them makes use of control groups to identify causal effects of CSR models on student achievement. Borman et al. (2003) provide an overview of existing studies with respect to 29 of the most widely implemented CSR models in the U.S. Considering only studies that are performed by an independent third party and that make use of some form of control groups, the strongest evidence for positive effects is provided by three programs: the Direct Instruction Program (DIP), the School Development Program (SDP) and Success for All (SFA). Remarkably, two of these, DIP and SFA, are relatively narrow-targeted interventions mainly focusing on better instruction methods and curriculum improvements. For one program, Edison, statistically significant negative effects have been found. This program intended to create innovative schools with a challenging curriculum, instruction methods tailored to the needs of the pupils and an emphasis on computer technology. In addition, the authors report large heterogeneity in the estimated effects between CSR models that cannot be explained by the differences in the specific measures included in the program. This suggests that school-specific requirements and/or the level and quality of implementation are more important for determining success. The average school across all studies reviewed had implemented its CSR model for around three years and the authors point out that, if cumulative effects exist, the analyses may underestimate the impact of CSR models. Although they find overall effects that appear promising, Borman et al. (2003) conclude that both quantity and quality of studies are insufficient to draw reliable conclusions on the effectiveness of CSR models yet. They advocate new programs to be evaluated making use of (quasi-) experimental research designs, to obtain more evidence-based knowledge on the effects of CSR.

Some more recent, also non-experimental, studies show ambiguous results. In an 11-year longitudinal study May and Supovitz (2006) evaluate the impact of a CSR design, called ‘America’s Choice’, on student test performance in Rochester, New York. The authors find significant positive effects on student performance which accumulate over time. The impact of the reform seemed to be larger in later grades than in the early grades, which might be due to a more powerful influence of the program measures in later grades. The authors argue that the positive impact of the program is likely to be caused by instruction methods targeted towards the needs of individual students, ambitious expectations for student performance and a supporting organizational structure within school that facilitate this tailored way of working. Gross et al. (2009) investigate the effects of federal CSR funds on student achievement in Texas making use of student-level panel data. Since schools have to apply for these funds, selection bias is a concern for the identification of the effects. The authors deal with this issue by controlling for school fixed effects and find that CSR funds did not significantly affect student's reading performance. The effects on math performance varied across different student types. Bifulco et al. (2005) evaluate three reform programs in New York City, including the School Development Program (SDP), Success for All (SFA) and the More Effective Schools (MES) program. The authors selected control groups by a random sample of troubled schools and then adjusted the sample so that the treatment and control groups have a similar propensity to start a CSR program. In contrast to some previous studies (see Borman et al., 2003) they did not find evidence that the SDP or SFA program significantly contributed to student performance. These findings, however, do support results from other earlier evaluations of SDP in Maryland, Chicago and Detroit (Cook et al., 1998; 1999; Millsap et al., 2001). Positive effects on reading scores are found for the MES program in the short run, but these do not persist when the external program trainers leave the school. This finding may suggest that schools face difficulties to maintain progress on their own, after the end of the program. Schwartz et al. (2004) assess the impact of a CSR model on student performance in New York, called the New York Networks for School Renewal Project. Student test scores in these schools are compared to a control group of students attending a set of randomly selected New York public schools. The authors make use of three cohorts of students, who are in grades 4, 5 or 6 at the start of the program in 1995-1996. The authors find mixed effects across grades after two to three years: in grade 4 CSR significantly increased both math and reading test scores, while the effects in grade 5 are insignificant for reading and negative for math. In grade 6, the program did not significantly affect performance.

In sum, the number of studies that use a credible research design to examine the effect of CSR models on student achievement is limited and existing studies show mixed findings. Our study adds to the literature by investigating the impact of a CSR policy on high-stakes test scores in the Netherlands. The ASIP shares some of the elements of other promising programs, including the increased attention to the individual needs of pupils and the tailor-made improvement plans that are guided by external experts. Our main contribution lies in the use of a quasi-experimental difference-in-differences approach to identify the effects of the program on pupil achievement.

5.3 The Amsterdam School Improvement Program

5.3.1 Background

In the Netherlands, the quality of primary schools is judged by the Dutch Inspectorate of Education. This governmental agency periodically investigates whether schools provide an acceptable standard of education. Each primary school is judged on a yearly basis by means of a risk analysis. This risk analysis includes several aspects such as student test results, the level of exams, personnel management, the financial position of the school, and compliance with Dutch educational laws. If the outcomes of the risk analysis provide evidence of weak performance, a more extensive quality analysis follows to determine whether schools are failing to meet the required educational quality standards. Based on these analyses the Inspectorate of Education classifies schools as ‘basic’, ‘weak’ or ‘very weak’. Schools that satisfy national quality standards are classified as basic, while schools classified as (very) weak perform below national standards. Inspection reports in 2006 and 2008 showed that the educational quality of a relatively large proportion of schools in Amsterdam was below the national standards. In Amsterdam 13 percent of all primary schools was classified as weak and 2.4 percent as very weak in 2008. The nationwide fraction of weak and very weak schools was 9.2 and 1.4, respectively (Inspectorate of Education, 2009). Primary schools in Amsterdam also performed worse compared to primary schools in the other large cities in the Netherlands (Inspectorate of Education, 2008). Concerns on those weakly performing schools have led to an intensive policy effort by the municipality of Amsterdam to improve educational quality (see e.g. Municipality of Amsterdam, 2009). It invested in a comprehensive school reform program that was introduced in 2008. All schools in Amsterdam that were classified as weak or very weak in the beginning of 2008 were invited to voluntarily participate in the program. After the school year 2008-2009, all primary schools in Amsterdam became eligible for the program. Hence, the program was initially targeted at failing schools, but became also accessible for sound performing schools later on.

5.3.2 Content

The Amsterdam School Improvement Program (ASIP) is an intensive two-year program designed to improve educational quality of participating schools. It aims to implement a systematic and performance-based way of working within the whole school. This includes ‘data-driven teaching’, meaning that teachers systematically measure pupil performance and use this information to adjust lessons to the individual needs of the pupils. Schooling of teachers is used to improve their classroom practices, school-leaders and other school personnel take courses to improve their skills in performance-based working, and instruction methods are often replaced. A consistent way of working throughout the school should create an efficient organization in which teachers are optimally facilitated in their primary teaching tasks. The program is guided by an expert team with strong educational experience.[4]

The ASIP consists of three steps. First, a profound quality analysis is made by the educational experts together with the school. This analysis includes instruction methods, student performance, student care, didactical routines and management performance, including leadership, communication skills and the existence of a coherent vision of education. An important aspect of the analysis is the evaluation of teacher quality through observations of lessons. The experts use a specific teacher evaluation system to judge teacher quality on a variety of classroom practices, including pedagogical, didactical, and organizational competences. [5] Second, based on this analysis, the school sets up an improvement plan in collaboration with the experts. This improvement plan states the specific measures that have to be implemented to improve educational quality within two years. These measures are suited to the specific needs of the schools and typically involve schooling of teachers, coaching of school-leaders and the use of new instruction methods. Third, after the improvement plan has been tested and approved by the expert team, it has to be implemented. During the implementation period the schools are supported by the expert team. The experts serve as critical advisors and visit the school at least once every three months to regularly assess the progress. Each six months the improvements in educational quality is measured in classes based on the teacher evaluation system. The expert team evaluates the progress and educational quality of the school more broadly and extensively during the first formal audit one year after the start of the program. If needed, the plan may be adjusted. After two years, a second audit takes place, which can be considered as the end of the program.[6]

As part of the program the municipality of Amsterdam developed instruction courses for professional development of school personnel. These accredited courses focus on the development of performance-based working skills and became available in the school year 2009-2010. [7] The courses are not only used

[4] The expert team largely consists of former inspectors of the Dutch Inspectorate of Education.

[5] The teacher evaluation system (TES) is called '*Kijkwijzer*'. Van der Steeg and Gerritsen (2013) find that high teacher quality scores on this TES are associated with better pupil test scores. This suggests that the TES measures teacher practices that are important for the educational performance of pupils.

[6] In exceptional cases where educational goals are not achieved yet, the program can be extended by an additional third year.

[7] These courses are mainly targeted at school-leaders and supportive school personnel.

for schooling of educators at schools that participate in the ASIP, but are also accessible for other primary schools in Amsterdam that do not participate in the program.

At the start of the ASIP, the municipality of Amsterdam also introduced new achievement goals for primary schools. The standards aimed for are above the nationwide standards of the Inspectorate of Education. [8] The announcement of the achievement goals reflects the ambitious plans of the municipality and serves as a signal for primary schools in Amsterdam. However, no explicit sanctions follow if a school does not satisfy the standards.

5.3.3 Costs and participation

The costs of the ASIP depend on the specific measures in the improvement plan. The average costs of the two-year ASIP amount to around 300,000 Euros per school, of which 250,000 Euros for the implementation of the interventions and 50,000 Euros for counselling by the expert team. Costs are shared by the municipality and the schools. The amount of resources to be paid by the school is dependent on its financial position. On average, the matching percentage of the schools is around 25%. The total costs involved are substantial: on average the yearly ASIP investment is more than 10% of the total government funding for an average primary school. [9]

There are 209 primary schools in Amsterdam, of which 50 participated in the ASIP by the end of the 2010-2011school year: 16 schools started in the program during or before the 2008-2009 school year, 14 during the 2009-2010 school year, and 20 during the 2010-2011 school year. It should be noted that these concern both schools that are classified as (very) weak and schools that are classified as basic. In our main analyses we focus on the sample of weak-performing schools in the beginning of 2008, which was the initial target population of the program (see Sections 5.4 and 5.5).

[8] The achievement goals include (i) an average CITO test score of at least 534, (ii) at least 25% percent of the pupils assigned to higher secondary education, and (iii) at most 20% of the disadvantaged pupils assigned to specific secondary education levels that provide special care because of learning difficulties (called '*Praktijkonderwijs'* and '*Leerwegondersteunend Onderwijs'*).

[9] With an average primary school size of 220 pupils, the average yearly ASIP investment per pupil is around 680 Euros. This is more than 10% of the per-pupil government funding of around 5,000 Euros.

5.4 Empirical strategy

To assess the impact of the ASIP on educational performance, we adopt a difference-in-differences (DID) estimation approach. This approach essentially compares the change in educational performance after and before the start of the program in Amsterdam to the same change in other Dutch cities that did not implement a CSR program. We implement this strategy by estimating the following model:

$$Y_{ist} = \beta_0 + \beta_1 A_{ist} + \beta_2 T_t + \beta_3 A_{ist}{*}T_t + X_{ist} + \alpha_s + \tau_t + \varepsilon_{ist}, \tag{5.1}$$

where Y_{ist} is the test score of pupil i in school s in year t, A_{ist} is a dummy variable that takes value 1 if pupil i is at school in Amsterdam and 0 otherwise, T_t is a dummy variable that takes value 1 in case of a post-treatment year and value 0 in case of a pre-treatment year, X_{ist} is a vector with individual background characteristics, α_s are school fixed effects, τ_t are year dummies and ε_{ist} is the error term. The estimated coefficient β_3 is the parameter of interest.

For our main analysis we use data on the CITO test scores from 2005 to 2012. This is a nationwide, high-stakes test that pupils take in their final year of primary school (see Section 5.5). We define the years 2005 till 2008, before the start of the ASIP, as pre-treatment years. The years 2009 till 2012 are the post-treatment years. Regarding the two-year program duration, one might argue whether 2009 is an appropriate post-treatment year. Therefore, we will also present results in case of only including later post-treatment years.

We focus our main analysis on the sample of all schools in the Netherlands that were classified as ‘weak’ or ‘very weak’ by the Inspectorate of Education on 1 January 2008. These failing schools were the initial target group of the ASIP and constitute a homogeneous sample with respect to school quality according to the nationwide standards of the Inspectorate of Education. All of these schools in Amsterdam were eligible for participation in the ASIP, whereas similar weak-performing schools outside Amsterdam were not allowed to participate. We use all (very) weak schools outside Amsterdam as our main control group. In addition, we construct two alternative control groups that consist of only larger cities in the Netherlands. Schools in other large cities may be more similar to the schools in Amsterdam. The crucial assumption for identification of the treatment effect is the common trend assumption, which implies that the development of test scores in Amsterdam would have been the development in test scores in the control group in the absence of the ASIP. This rules out city-specific trends and composition effects. To investigate the validity of the common trend assumption we compare the pre-treatment test score development in Amsterdam to the pre-treatment test score development in the control groups. We address the potential effect of the policy on the composition of schools by presenting a sensitivity analysis that includes only those schools that participated in the CITO test both before and after the introduction of the policy. Furthermore, we argue and provide supportive evidence that our estimation results are not likely to be affected by changes in the composition of pupils within schools.

The estimated treatment effect should be interpreted as the effect of the introduction of the ASIP policy. The introduction of the policy offered all failing schools in Amsterdam the opportunity to participate in the program. Since not all of the eligible schools participated in the program, we estimate an intention-to-treat (ITT) effect. The ITT effect differs from the effect of actual participation in the program. The standard approach to estimate the effect of participation for those who participate would be to use eligibility for the program as an instrumental variable for participation in the program (Imbens and Angrist, 1994; Angrist et al., 1996). Estimation by two-stage-least-squares then yields the treatment-on-the-treated effect, which is essentially equal to the ITT effect divided by the compliance rate (Bloom, 1984). This analysis assumes that eligibility for the program, i.e. being at school in Amsterdam in a post-treatment year, does not affect the outcomes of non-participating schools. This assumption is not likely to hold here because schools that do not participate in the ASIP have the opportunity to take the professional development courses from the 2009-2010 school year onwards, which may affect outcomes. It turns out that three schools in our estimation sample participated in the courses without following the complete ASIP program (see Section 5.5). This makes it impossible to strictly disentangle the effect of participation in the complete ASIP from participation in only the professional development courses. We therefore only present the estimated ITT effects, which pick up both effects.

The potential influence of other implemented policies in Amsterdam does not seem to be a main concern in our analysis, because of the relatively large size of the ASIP. One project with the goal of raising pupil's math performance was implemented in the south-east district of Amsterdam in 2008.[10] Outside Amsterdam, the municipality of Rotterdam started an action program to improve the educational quality of primary schools in 2011.[11] To address the potential impact of these other programs we present sensitivity analyses in which we leave out the schools in the corresponding areas.

[10] This program was called '*Omdat elk kind telt in Zuidoost*'.

[11] This action program is called '*Beter Presteren'* and focuses on additional school time, professionalization of schools and parental involvement.

When calculating standard errors we take into account the presence of common group errors (Moulton, 1986). In all estimation results we present robust standard errors corrected for clustering at the school-year level. Still, standard errors may be too small in our case where we use multiple years of data. Bertrand et al. (2004) show that ignoring serial correlation in outcomes can lead to over-rejection of the null hypothesis of no effect. To address this potential issue of serial correlation, we also present estimation results of a model in which we collapse the data before and after the introduction of the policy (Bertrand et al., 2004).

5.5 Data

We received information on schools that were classified as 'weak' or 'very weak' on 1 January 2008 from the Inspectorate of Education. For this sample of schools we obtained data on CITO test scores from the CITO organization. The CITO test is a nationwide, high-stakes test that pupils take in the highest grade of primary education (grade eight). It contains questions on language, math, and information processing. Test taking takes place during three days in the beginning of February and the test is used for the assignment of pupils to different levels of secondary education. Teachers use the test results to advice pupils on the most appropriate secondary education level and secondary schools often use threshold values for enrolment in more advanced types of secondary education. Test results at the school level are used by the Inspectorate of Education to judge the quality of primary schools. Each year more than 80 percent of all primary schools participate in the CITO test. When a school chooses to participate in the CITO test, in principle all pupils in grade eight have to take the test. [12]

Our dataset contains information on CITO test scores at the pupil level from 2005 to 2012 for all failing schools on the reference date 1 January 2008. The total sample includes 614 schools that are comparable with respect to educational quality according to the nationwide standards of the Inspectorate of Education. It should be noted that our sample contains only schools that have participated in the CITO test at least once during the period 2005-2012.[13] The total number of observations equals 78,545. Our estimation sample contains 35 schools in Amsterdam that were eligible for the ASIP. A total of 24 of these 35 schools have participated in the ASIP. Table 5.1 provides a more detailed overview of the timing of relevant events.

[12] An exception is made for pupils in special categories such as foreign students that have been in the Netherlands for a short time and students that are expected to be assigned to secondary education types with special care (see also Table 5.3).

[13] The total number of primary schools that were classified as 'weak' or 'very weak' on 1 January 2008 equals 751. We do not observe the schools that never participated in the CITO test.

Table 5.1 Timing of Events

Time	Event
February 2005	CITO test 2005
February 2006	CITO test 2006
February 2007	CITO test 2007
February 2008	CITO test 2008
April 2008	Municipality of Amsterdam introduces the ASIP
April 2008 - February 2009	7 schools start in the ASIP
February 2009	CITO test 2009
February 2009 - September 2009	5 schools start in the ASIP
September 2009	Municipality of Amsterdam introduces the professional development courses
September 2009 - February 2010	3 schools start in the ASIP
February 2010	CITO test 2010
February 2010 - September 2010	3 schools start in the ASIP
September 2010 - February 2011	5 schools start in the ASIP
February 2011	CITO test 2011
February 2011 - September 2011	1 school starts in the ASIP
February 2012	CITO test 2012

Notes. The presented number of schools that have started in the ASIP concern only those in our estimation sample.

The CITO tests of 2005, 2006, 2007, and 2008 have taken place before the introduction of the policy. As of April 2008, all failing schools in Amsterdam were invited to participate in the ASIP. It turns out that 12 schools started in the ASIP before or during the 2008-2009 school year, of which 7 before the CITO test of 2009; 6 schools started in the ASIP during the 2009-2010 school year, of which 3 before the CITO test of 2010; and 6 schools started in the ASIP during the 2010-2011 school year, of which 5 before the CITO test in 2011. From the 2009-2010 school year onwards, schools in Amsterdam could also participate in the professional development courses. In total 16 of the 35 schools in Amsterdam participated in these courses; 13 schools took the courses as part of the complete program, and 3 schools only took the courses without participation in the ASIP.

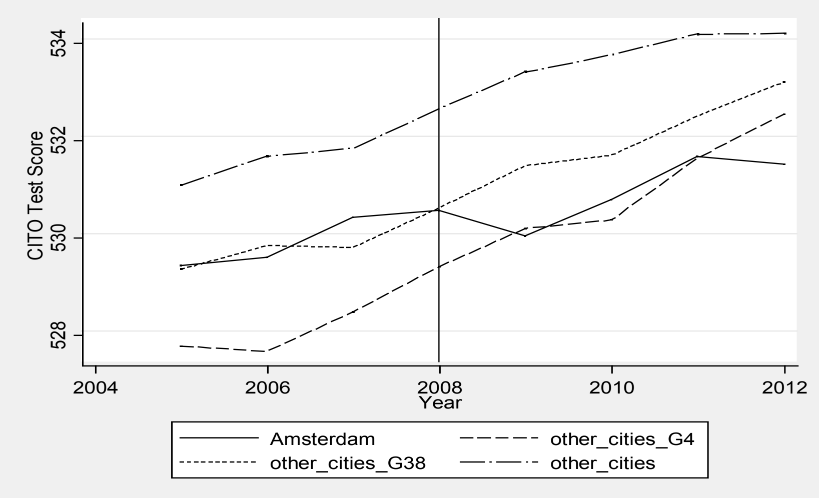
Table 5.2 presents summary statistics of the CITO test scores in our sample. The test consists of 200 questions: 100 questions on language, 60 questions on math and 40 questions on information processing. The test scores on language, math and information processing equal the number of correctly answered questions. The total score is a linear transformation of the total number of correctly answered questions and ranges from 501 (lowest score) to 550 (highest score) each year. The linear transformation is such that the total scores are comparable across years.

Table 5.2 Summary statistics of CITO test scores (total sample 2005-2012)

	Average	Standard deviation	Min	Max	Observations
Total score	532.60	10.30	501	550	78,545
Language	70.27	14.08	10	100	78,545
Math	40.16	11.60	1	60	78,545
Information processing	28.42	6.57	0	40	78,545

Figure 5.1 presents the total test score developments from 2005 to 2012 for Amsterdam, and for three control groups. The control groups contain failing schools outside Amsterdam. We use all schools outside Amsterdam that were classified as (very) weak by the Inspectorate on reference date 1 January 2008 as our main control group. In addition, we use two alternative control groups that consist of subsamples of large cities in the Netherlands: the so-called 'G38' and 'G4' cities. The G38 consists of 38 medium- and large cities and the G4 consists of the four largest cities in the Netherlands (Amsterdam, Rotterdam, The Hague and Utrecht). Schools in other large cities may be more similar to the schools in Amsterdam.

Figure 5.1 CITO test scores for weakly performing schools in the Netherlands.



The availability of CITO test scores back to 2005, 3 years prior to the introduction of the policy, allows us to compare the pre-treatment test score trend in Amsterdam with that in the control groups.
The performance of pupils in Amsterdam increased from 2005 till 2008. The increase in test scores is well comparable to that in other cities. The test score trend indicates that the schools in Amsterdam did not experience a pre-treatment performance dip that would invalidate the common trend assumption (Ashenfelter, 1978).[14] The level of test scores in Amsterdam is larger compared to that in the other three large cities in the Netherlands, almost identical to that in the group of 37 other medium and large cities, and smaller compared to the nationwide test score level. To test for differential trends between Amsterdam and the control group, we regressed the first-difference of the test score on a time trend and an interaction term between the time trend and a dummy variable for Amsterdam. This yields an insignificant coefficient for the interaction term (with a *t*-value of -0.46), indicating that there is no evidence for differential trends. This supports the credibility of our identifying common trend assumption. A comparison of the post-treatment development of test scores in Amsterdam to the post-treatment development of test scores outside Amsterdam provides a first impression of the impact of the program. After the introduction of the ASIP in 2008, we observe that the increase in test scores in Amsterdam becomes smaller compared to that in the other cities.

The CITO dataset also contains information on individual background characteristics, such as gender, birth date and subsidy factor. In our empirical analyses we use these as covariates to control for observable changes in the pupil population over time. Table 5.3 presents the sample means of both the covariates and the outcome variables in 2008 (the most recent pre-treatment year) and 2012 (the most recent post-treatment year) for Amsterdam and the three control groups. The variable gender takes the value of one in case of a male and the value of zero in case of a female. We dispose of the year and month of birth of each pupil, from which we construct the age in years at the time of the test. We lack data on gender for 372 observations and on age for 359 observations. For this small fraction of our estimation sample, we impute missing values by the average value in the estimation sample. Furthermore, we dispose of categorical variables for the language spoken at home (seven categories), subsidy factor (six categories) and pupil category (six categories). The subsidy factor is an indicator of socioeconomic background. The Dutch funding scheme for primary schools distinguishes several groups of disadvantaged pupils, for whom primary schools receive additional funding. The subsidy factor depends

[14] In case of a so-called 'Ashenfelter dip' one would expect schools in Amsterdam to improve after the introduction of the program, because of mean reversion. One might then incorrectly conclude that this improvement would be caused by the program.

on parental education level and can take values 0.3 and 1.2.[15] A subsidy factor of 0.3 implies that a school receives 30% of additional funding and a subsidy factor of 1.2 implies 120% of additional funding. The subsidy factor takes value 0 in case of a non-disadvantaged pupil. In earlier years, before 2010, other rules for the subsidy factor were used that depended not only on parental education level, but also on profession and ethnic background. This explains that the factors can also take values 0.4 or 0.9 in the years 2005-2009. The pupil category refers to specific groups of pupils for whom participation in the CITO test is not compulsory. Category 'I' stands for foreign pupils that have been in the Netherlands for less than four years; category 'J' for pupils that are expected to be assigned to special education; and category 'K' for pupils that are expected to be assigned to vocational secondary education with additional care.[16] All of these categorical variables contain one category to refer to an unknown value. The language spoken at home is unknown for around 12% of the observations in the years 2005-2011. In 2012 this information was not collected, implying that it is unknown for all observations. The subsidy factor is unknown for around 30% of the observations in the years 2005-2009 and for around 12% of the observations in the years 2010-2012.

Asterisks indicate that the sample mean in the control group differs significantly from that in Amsterdam in the corresponding year. Pupils in Amsterdam are well comparable to those in the other three groups with respect to gender. The age of the pupils when taking the CITO test in Amsterdam is comparable to that in the G38, but somewhat below (above) that in the G4 (rest of the Netherlands) in the pre-treatment year. Amsterdam is less comparable to the control groups with respect to the other covariates. Amsterdam, the largest city in the Netherlands, typically has a large fraction of disadvantaged pupils. This explains the smaller fraction of pupils with Dutch as their home language, the larger fraction of pupils with a high subsidy factor and the larger fraction of pupils belonging to a special category. We observe that in the G38 and the G4, differences on these variables are smaller, though still statistically significant in most cases. Only the pre-treatment difference in pupil category between Amsterdam and the G4 is insignificant. In our empirical analyses we use these variables as covariates to control for observable changes in the student population over time in Amsterdam and the control groups. The bottom panel presents the CITO test scores that have been standardized to have a mean of zero and standard deviation of one in the full sample. Pre-treatment test scores in Amsterdam are very similar to those in the G38, and below (above) those in the rest of the Netherlands (G4).

[15] The subsidy factor equals 1.2 in case the highest completed education level is primary education for at least one of the parents and lower secondary education for the other. The subsidy factor equals 0.3 in case lower secondary education is the highest completed education level for both parents (or the parent which is responsible for daily care).

[16] This type of education is called *'Leerwegondersteunend Onderwijs' (LWOO)*. Pupils in this category generally suffer from learning arrears, low IQ and/or social or emotional problems.

Table 5.3 Sample means in 2008 and 2012

	Amsterdam		Rest Netherlands		G38		G4	
	2008	2012	2008	2012	2008	2012	2008	2012
Covariates								
Gender (male = 1)	0.48	0.47	0.50	0.49	0.50	0.49	0.48	0.50
Age	12.12	12.00	12.06***	12.00	12.14	12.07***	12.21***	12.09***
Home language								
Dutch	0.51	0.00	0.81***	0.00	0.66***	0.00	0.59***	0.00
Other Western-Europe	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00
Arabic	0.13	0.00	0.02	0.00	0.05	0.00	0.06	0.00
Surinam	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.00
Turkish	0.09	0.00	0.03	0.00	0.09	0.00	0.10	0.00
Other	0.06	0.00	0.03	0.00	0.04	0.00	0.05	0.00
Unknown	0.18	1.00	0.10	1.00	0.14	1.00	0.18	1.00
Subsidy factor								
0	0.30	0.61	0.62***	0.71***	0.47***	0.63***	0.33***	0.57***
0.3	0.00	0.10	0.00	0.09	0.00	0.11	0.00	0.12
0.4	0.05	0.00	0.12	0.00	0.12	0.00	0.12	0.00
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.2	0.00	0.25	0.00	0.08	0.00	0.17	0.00	0.23
Unknown	0.66	0.04	0.26	0.11	0.41	0.10	0.55	0.08
Pupil category								
No special category	0.80	0.79	0.92***	0.90***	0.88***	0.89***	0.85	0.86***
I	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.03
IJ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IK	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
J	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
K	0.17	0.18	0.07	0.07	0.10	0.08	0.14	0.09
Outcome variables								
Total score	-0.21	-0.11	0.00***	0.15***	-0.20	0.05***	-0.32**	-0.01**
Language	-0.18	-0.36	0.02***	-0.08***	-0.18	-0.19***	-0.29**	-0.26**
Math	-0.05	0.03	0.08***	0.19***	-0.06	0.14***	-0.18**	0.11*
Information processing	-0.25	-0.14	-0.02***	0.18***	-0.19	0.04***	-0.26	-0.04**
Schools	34	34	500	487	112	132	32	38
Observations	901	991	8,739	9,275	2,411	2,996	733	1,023

Notes. Asterisks indicate that the sample mean in the control group differs significantly from that in Amsterdam in the corresponding year at a *10% level, **5% level, and ***1% level. Tests of significant differences for gender, age, and the outcome variables are based on a two-tailed *t*-test. Tests of significant differences for the categorical variables are based on a chi-squared test.

We observe that the total test scores have improved over time, both in Amsterdam and in the three control groups. The difference in test scores between 2012 and 2008 is smaller in Amsterdam compared to the three control groups for each of the test subjects.

5.6 Results

5.6.1 Main findings

Table 5.4 presents the estimates of the effect of the introduction of the ASIP on CITO test scores for three model specifications. The first model (column 1) regresses the standardized CITO test score on a dummy for Amsterdam, a dummy for a post-treatment year, an interaction term that indicates a post-treatment year in Amsterdam, and year dummies.[17] The second model adds individual pupil background characteristics such as gender, age, age squared, home language, subsidy factor and pupil category. The third model additionally includes school fixed effects.

The top panel reports the estimation results for the complete sample. The middle panel shows similar results for the subsample of 38 middle and large Dutch cities and the bottom panel presents them for the subsample containing the four largest cities in the Netherlands. For each of these samples we present the estimated effects on the total CITO score as well as on the specific subjects language, math, and information processing. Since all test scores are standardized, the estimated effects can be interpreted in terms of standard deviations.

In the complete sample we find negative effects of the introduction of the ASIP on CITO test scores in all model specifications. The addition of controls increases the size of the point estimates. This can be explained by a more favourable development of covariates (related to higher test scores) in Amsterdam compared to the control groups.[18] When estimated with all individual background characteristics and school fixed effects, the estimated effect implies that the introduction of the ASIP decreases the total CITO test score by 0.17 standard deviations. The estimated coefficient is statistically significant at the 1% level. Statistically significant negative effects are also found for each of the subjects of the test. The negative impact is largest for language and smallest for math, with estimated effects of -0.19 and -0.09,

[17] Estimated treatment effects in models without the year dummies are very similar.

[18] For example, the increase in the share of non-disadvantaged pupils with a subsidy factor of 0 between 2008 and 2012 is larger in Amsterdam than that in other cities (see Table 5.3). Since non-disadvantaged pupils are more likely to perform well on the CITO test, inclusion of this control variable decreases the estimated treatment effect.

respectively. The G38 and G4 samples yield similar results. We find negative and statistically significant effects in all models, within a range from -0.11 to -0.20. These results indicate that the ASIP has negatively affected educational performance in the highest grade of primary education in the first four years after its introduction.

Table 5.4 Difference-in-differences estimates of the introduction of the ASIP

	(1)	(2)	(3)
A. Complete Sample			
Total score	-0.105*(0.062)	-0.144***(0.049)	-0.170***(0.036)
Language	-0.109*(0.060)	-0.152***(0.046)	-0.188***(0.034)
Math	-0.052 (0.056)	-0.078*(0.047)	-0.086**(0.036)
Information processing	-0.127**(0.061)	-0.160***(0.049)	-0.179***(0.035)
Observations	78,545	78,545	78,545
Schools	614	614	614
B. G38			
Total score	-0.127*(0.068)	-0.154***(0.053)	-0.190***(0.040)
Language	-0.117*(0.067)	-0.155***(0.050)	-0.200***(0.037)
Math	-0.107*(0.061)	-0.114**(0.052)	-0.136***(0.040)
Information processing	-0.114*(0.067)	-0.139***(0.053)	-0.164***(0.039)
Observations	27,882	27,882	27,882
Schools	173	173	173
C. G4			
Total score	-0.181**(0.084)	-0.145**(0.066)	-0.149***(0.049)
Language	-0.162*(0.085)	-0.140**(0.063)	-0.158***(0.046)
Math	-0.175**(0.074)	-0.131**(0.064)	-0.120**(0.049)
Information processing	-0.143*(0.086)	-0.113*(0.068)	-0.107**(0.050)
Observations	13,597	13,597	13,597
Schools	74	74	74
Individual characteristics	no	yes	yes
School fixed effects	no	no	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level.

5.6.2 Heterogeneity

We proceed by investigating the impact of the introduction of the ASIP across different groups of pupils. We distinguish between male and female, higher and lower socio-economic status, and Dutch and foreign pupils. We define all pupils with a subsidy factor larger than 0 to have low socio-economic status and all other pupils to have high socio-economic status. Pupils with home language other than Dutch are defined as ‘foreign’ and all other pupils as ‘Dutch’. We leave out pupils with missing values on these variables.

Table 5.5 reports the estimated full model effects for each of the subgroups, taking into account the complete sample containing all weakly performing schools in the Netherlands. The estimated effects on the test scores are reasonably in line with the total sample estimates and are similar across subgroups.[19] With respect to the specific subjects, the estimated effects on math turn insignificant in models (3), (5) and (6), but the differences in effects sizes across the subsamples are not large. Table 5.A.1 in the Appendix presents the results for the G38 and G4 samples. The results in the G38 sample are in line with our findings of no differential effects across subgroups in the complete sample. The results in the G4 suggest that the introduction of the policy has been more detrimental for non-disadvantaged pupils than it has been for disadvantaged pupils. This finding holds for each of the subjects and differs from our findings in the other samples. In sum, we conclude that we find no strong evidence that specific groups of pupils are particularly affected by the policy. When taking into account only the four largest cities, the policy seems to have had a more detrimental impact for non-disadvantaged pupils.

Table 5.5 Heterogeneous treatment effects: Estimated effects of the introduction of the ASIP

	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Low socio-economic status	High socio-economic status	Foreign	Dutch
Total score	-0.164***	-0.164***	-0.205***	-0.196***	-0.106*	-0.130***
	(0.042)	(0.040)	(0.069)	(0.046)	(0.057)	(0.048)
Language	-0.186***	-0.178***	-0.231***	-0.230***	-0.107*	-0.151***
	(0.041)	(0.037)	(0.066)	(0.045)	(0.058)	(0.046)
Math	-0.074*	-0.088**	-0.075	-0.091**	-0.068	-0.038
	(0.040)	(0.043)	(0.078)	(0.046)	(0.058)	(0.046)
Information processing	-0.178*** (0.045)	-0.174*** (0.039)	-0.277*** (0.073)	-0.211*** (0.048)	-0.123** (0.061)	-0.171*** (0.050)
Observations	38,607	39,566	12,006	48,527	7,150	53,385
Schools	614	614	556	608	419	601
Individual characteristics	yes	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level.

[19] In models (3) and (4) the estimated effects are both larger (in absolute value) than the total sample estimates, while models (5) and (6) both yield smaller estimates. This can be explained by the fact that we leave out pupils with missing values on socioeconomic status or home language.

In addition to the effects for different groups, we investigate the impact of the ASIP on different parts of the test score distribution by estimating quantile regressions. Table 5.6 presents the estimated effects for various quantiles of the test score distributions. With respect to the total test score, the estimated coefficients differ across quantiles. The impact is most detrimental at the lower tale of the distribution. The introduction of the ASIP decreases the lower quartile of the total test score distribution by 0.19 standard deviations, the median by 0.18 standard deviations (which is close to the OLS coefficient of -0.17) and the upper quartile by 0.11 standard deviations. The impact is smallest at the upper decile, -0.06, and largest at the lower decile, -0.25. This pattern of decreasing estimated effect sizes with quantile also shows up for the specific subjects language and information processing. The pattern is less clear for math, though also here the estimated effects are smaller (in absolute value) in the upper tail of the test score distribution. Table 5.A.2 in the Appendix presents quantile regression results for the G38 and G4 samples. The findings in the G38 are consistent with the observed pattern in the complete sample. The picture in the G4 is less clear. The largest effect sizes are not always found to be at the lowest decile, but in most cases the smallest effect sizes are found at the upper decile. Hence, our finding of a less detrimental impact at the upper part of the test score distribution is confirmed in both other samples as well.

Table 5.6 Quantile regressions results: Estimated effects of the introduction of the ASIP for quantiles of the test score distributions

	(1)	(2)	(3)	(4)	(5)
	quantile	quantile	quantile	quantile	quantile
	regression	regression	regression	regression	regression
	0.1			0.75	0.90
		0.25	0.50		
Total score	-0.247***	-0.193***	-0.183***	-0.108***	-0.062**
	(0.038)	(0.032)	(0.029)	(0.03)	(0.029)
Language	-0.246***	-0.192***	-0.212***	-0.161***	-0.093***
	(0.041)	(0.032)	(0.028)	(0.027)	(0.029)
Math	-0.097**	-0.101***	-0.111***	-0.063**	-0.052**
	(0.041)	(0.036)	(0.031)	(0.026)	(0.026)
Information	-0.309***	-0.240***	-0.186***	-0.138***	-0.077***
processing	(0.043)	(0.035)	(0.029)	(0.025)	(0.026)
Observations	78,545	785,45	78,545	78,545	78,545
Schools	614	614	614	614	614
Individual characteristics	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes

Notes. Each cell represents a separate quantile regression. Standard errors are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level.

5.6.3 Sensitivity

Table 5.7 presents several sensitivity analyses to probe the robustness of our main findings. First, we restrict our sample to schools that participated in the CITO test in each of the years 2005 to 2012. A potential concern is that the introduction of the policy affected participation in the CITO test. For instance, it might be that non-treated schools outside Amsterdam were closed due to persistently bad performance, in which case they do not show up in our data in all post-treatment years. This might bias our estimates downwards if the weakest performing schools outside Amsterdam drop out of the estimation sample. Column (1) presents estimation results for the sample consisting of only schools that participate each year. This includes 399 schools containing 63,165 pupils. The estimated effects are slightly larger (in absolute value) than the main estimates. This indicates that our results are unlikely to be biased by selective attrition of schools.

Table 5.7 Sensitivity: Estimated effects of the introduction of the ASIP on CITO test score (full model).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Sample of schools that participate in the CITO test in all years 2005-2012	Sample excluding the south-east district in Amsterdam	Sample excluding schools in Rotterdam	Sample excluding the year 2009	Sample excluding the years 2009, and 2010	Sample excluding the years 2009, 2010 and 2011	Sample collapsed to before/after observations at school level
Total score	-0.190***	-0.128***	-0.169***	-0.169***	-0.208***	-0.299***	-0.144**
	(0.037)	(0.039)	(0.035)	(0.039)	(0.043)	(0.057)	(0.063)
Language	-0.200***	-0.155***	-0.189***	-0.187***	-0.217***	-0.321***	-0.177**
	(0.035)	(0.037)	(0.034)	(0.037)	(0.041)	(0.053)	(0.059)
Math	-0.109***	-0.042	-0.081**	-0.100**	-0.153***	-0.204**	-0.075
	(0.037)	(0.040)	(0.036)	(0.039)	(0.043)	(0.059)	(0.064)
Information processing	-0.201*** (0.037)	-0.140*** (0.036)	-0.184*** (0.035)	-0.165*** (0.038)	-0.181*** (0.044)	-0.274*** (0.057)	-0.144** (0.065)
Observations	63,165	76,956	75,955	69,030	59,438	49,462	1,145
Schools	399	607	601	612	612	609	614
Individual characteristics	yes	yes	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level.

Second, we exclude schools whose results might have been affected by other programs from the sample. We leave out schools in the south-east of Amsterdam to address the potential influence of the program that was raised to improve pupil's math performance in this district. The district contains 7 primary schools in our sample including 1,589 pupils. If the program has improved results, we would expect the estimated effects to become more detrimental. Instead, we find negative point estimates that are smaller (in absolute value) than our main estimates and the estimated effect for math turns insignificant (see column 2). This suggests that our main results are not biased upwards because of the impact of the program in the south-east district. Furthermore, we exclude schools in Rotterdam and find estimated effects that are very similar to our main estimates (see column 3). This suggests that the action program launched by the municipality of Rotterdam in 2011 does not affect our findings.[20]

Third, we perform similar analyses on a sample in which we leave out the year 2009. Since the first schools that participated in the ASIP started in 2008, one might argue whether 2009 is an appropriate post-treatment year. Excluding 2009 from the sample, using only 2010 to 2012 as the post-treatment years, yields similar result (see column 4). Regarding the two-year program duration, it might take even longer before the impact of the program was felt in the CITO test scores. To provide insight into the timing of effects, we further reduce the number of post-treatment years taken into account. In model (5) we leave out the years 2009 and 2010 whereas model (6) additionally excludes the year 2011 from our sample. In case of an improvement in test scores over time, we expect to find better results in models that include only the most recent post-treatment years. Instead, the negative point estimates of the effect of the ASIP become larger in models (5) and (6). Hence, we find no evidence for cumulative effects over time in the first four years after the introduction of the program.

A complicating factor for analysing cumulative effects is that not all schools started in the program at the same time. If participation in the ASIP causes an initial decrease in test scores that is followed by an upward development, a concern might be that an improvement in test scores of participating schools that started in 2008-2009 is negated by a decrease in test scores in schools that started in later years. This, however, seems a less plausible interpretation of our findings because of the relatively large share of schools in our sample that started in 2008-2009 (see Table 5.1). We further address this issue by excluding all schools in Amsterdam that did not start in the ASIP before or during the 2008-2009 school year. Including only the 12 schools in Amsterdam that participated in 2008-2009 (and all other schools outside Amsterdam), we would expect to find increasing effects when restricting the sample to more

[20] Excluding the schools in Rotterdam in the G4 sample yields an estimated effect of -0.167***(0.061) on the total CITO test score.

recent post-treatment years in case of cumulative effects. Table 5.A.3 in the Appendix presents the estimation results for the four sets of post-treatment years. We find that the negative point estimates of the impact of the ASIP become larger when we restrict the sample to more recent post-treatment years. This is consistent with our conclusion of no improvement in test scores over time.

In addition, we have estimated four separate models that each include one post-treatment year. The results are shown in Table 5.A.4 in the Appendix. We find mostly statistically significant negative estimates for each of the years. The estimated effects on test scores are smallest (in absolute value) in the years 2010 and 2011, and largest in 2012. Table 5.A.5 presents similar results when estimated on the sample that excludes schools in Amsterdam that did not start in the ASIP before or during the 2008-2009 school year. The estimated effects on test scores are smaller (in absolute value) and statistically insignificant for the year 2010, while larger and statistically significant for the other post-treatment years. We conclude that also these separate estimates for each post-treatment year do not provide evidence of increasing effects over time.

Fourth, we address concerns on over-significance of our estimates because of potential serial correlation problems, by collapsing the data before and after the introduction of the ASIP at the school level (Bertrand et al., 2004). This leaves us with 1145 observations.[21] We find a statistically significant effect on the total test score of -0.14 at the 5%-level.

A final concern might be that the policy has affected the testing pool within schools that participate in the CITO test. More specifically, if the policy caused a relative increase in the CITO test participation of weak students (with low unobserved ability) in Amsterdam compared to other cities, this might have biased our estimated effects downwards. A change in the tested population induced by the program, however, does not seem very likely since a larger CITO test participation was no clear element or goal of the ASIP. Hence, the program did not explicitly stimulate additional participation of (low ability) students in the test. Still, one might be concerned that the ASIP implicitly affected the testing pool if increased scrutiny limited opportunities to exclude weak students from the test. [22] In our data we do not observe a particularly large increase in the number of students that take the test in Amsterdam relative to other cities after the introduction of the policy. In addition, the share of pupils placed in special categories stays

[21] The sample contains 614 schools and 2 time periods. Not all 614 schools are present in both the before and the after period: there are 561 schools in the before period and 584 schools in the after period.

[22] Since the CITO test scores are important for judging educational quality, schools may have an incentive for shaping the testing pool. Previous studies have shown that schools can respond strategically to the implementation of accountability policies by excluding weak students from the test (e.g. Jacob, 2005).

reasonably constant over time, both in Amsterdam and other cities (see also Table 5.3). Ideally, we would have disposed of the total number of students in grade eight for each school and year. This would have enabled us to compare the development of participation shares in the CITO test between Amsterdam and other cities, and to estimate the effect of the policy on CITO test participation. However, our CITO data only contain information on the pupils that participated in the test. Therefore, we have performed an additional analysis making use of another dataset, called COOL, that includes a representative sample of around 10% of all Dutch primary schools in the pre-treatment year 2007-2008 and post-treatment year 2010-2011 (Driessen at al., 2009; 2012). These data include information on CITO test participation for the pupils in grade eight. Regressing a dummy variable for CITO test participation on a dummy variable for Amsterdam, a dummy variable for the post-treatment year, an interaction term between Amsterdam and the post-treatment year, and a set of pupil background characteristics yields an insignificant estimated effect for the interaction term that is close to zero.[23] Although this analysis is performed on a different sample that concerns not only failing schools but also sound performing schools, we interpret our finding of no effect as supportive evidence that the introduction of the policy did not affect participation in the CITO test.[24]

A related issue is the impact of grade retention. If the policy has affected retention, this could have changed the pupil composition in grade eight. Since we do not dispose of formal information on grade retention, we investigate this issue further by comparing the age of the pupils after and before the introduction of the policy in Amsterdam and in the other cities. We perform two analyses. First, we regress age on a dummy for Amsterdam, a dummy for a post-treatment year, an interaction term that indicates a post-treatment year in Amsterdam, year dummies and all other covariates. Second, we use a similar specification to estimate the effect of the introduction of the policy on a dummy variable indicating whether a pupil is older than 12.5. We use this dummy variable as an indicator for grade retention since pupils aged above 12.5 are most likely to have retained in grade. The first column of Table 5.A.6 in the Appendix presents the estimation results. We find statistically significant effects in both models: the age at which pupils take the test decreases with around 0.08 years (row 1) and the probability of grade retention decreases with around 3 percentage points (row 2). These results suggest that the retention probability has decreased following the introduction of the policy. This may have biased our

[23] The estimated effect for the interaction term is 0.001 (0.029). The included pupil background characteristics are gender, age, age squared, and a categorical socioeconomic status variable that distinguishes six categories based on parental education level and ethnic origin. The total sample includes 18,887 pupils in grade eight divided over 676 different schools.

[24] Restricting the sample to only those that were classified as weak or very weak on 1 January 2008 leaves us with only 6 schools in Amsterdam, of which 3 participated in the ASIP. A similar analysis on this subsample yields an estimated effect for the interaction term of 0.005 (0.006).

estimates downwards if an additional year in education increases test scores for weak pupils. To address this issue, we proceed with two robustness analyses that are presented in columns 2 - 5 of Table 5.A.6. First, we estimate the effect of the introduction of the policy on the CITO test score for the subsample of pupils who have not been retained. Excluding pupils who are aged above 12.5 yields a statistically significant negative effect, -0.14, which is somewhat below (in absolute value) the estimated effect in the complete sample, -0.17 (see column 2).[25] Second, we estimate the impact of the policy for three subsamples that exclude schools in Amsterdam with the largest decrease in age (see columns 3-5). We define the decrease in age as the average age of test taking in the years after the introduction of the policy minus the average age of test taking in the years before the introduction of the policy. In the first subsample we exclude all schools in Amsterdam for which the average age of test taking decreases with 0.2 years or more (eight schools). The second subsample leaves out all schools in Amsterdam with a decrease in age of at least 0.15 years (twelve schools) and the third subsample leaves out all schools in Amsterdam with a decrease in age of at least 0.10 years (sixteen schools). In this way schools that are most likely to have faced a reduction in retention following the introduction of the policy are excluded from the analyses. Leaving out such schools obviously reduces the estimated impact of the policy on age and on the indicator for grade retention (see rows 1 and 2 in columns 3, 4, and 5). The estimated effect on the indicator for grade retention is statistically insignificant and close to zero in all models. In each of the three subsamples we find statistically significant negative effects of the introduction of the policy on the CITO test score, ranging from -0.14 to -0.17 (see row 3 in columns 3, 4, and 5). These results are close to our main estimates and suggest that our findings are not importantly affected by the potential impact of grade retention.

We conclude that our finding that the introduction of the ASIP negatively affected test scores in grade eight of primary education is robust to a variety of sensitivity tests.

5.7 Interviews and potential mechanisms

To gain further insight into the effects of the policy, we have taken interviews with school-leaders of participating schools. We focused on the schools that started in the ASIP before or during the 2008-2009

[25] Please note that this analysis is not fully informative on the magnitude of potential bias caused by the impact of grade retention. After all, it does not exclude those pupils in Amsterdam that have not retained after the introduction of the policy, but that would have retained in the absence of the policy. In addition, the lower estimated effect may well be explained by the exclusion of weak performing pupils for whom the impact of the policy on test scores is likely to be more detrimental (see Table 5.6).

school year and found seven school-leaders who were willing to provide information on their experiences with the program. The outcomes of these interviews reveal a potential explanation for our empirical findings. The overall opinions of the school-leaders were very similar and yield a consistent picture that is two-fold. Most of the school-leaders expected that the program would result in better educational quality in the longer term. They especially appreciated the use of teacher quality evaluations which provided insight into teacher behaviour that revealed current weaknesses in classroom practices and provided a clear view on potential improvements. In addition, the courses for school-leaders on performance-based working and the use of new instruction methods were mentioned as valuable elements of the program. At the same time, they experienced the ASIP as an intensive and radical program with a rigorous approach. Most of them reported severe resistance among teachers, for whom the program was especially demanding. The teachers were confronted with direct feedback on their classroom behaviour, changes in their instruction materials, and were expected to put in effort to improve their competences in addition to their regular teaching tasks. Almost all school-leaders report the exit of school personnel after the introduction of the program. Some of them left the school voluntarily because they did not want to go along with the changes induced by the program, but others were forced to leave because they appeared not capable to satisfy the required standards. The proportion of replaced teachers seems to be substantial. Three school-leaders explicitly mention the number of replaced teachers. Two of them report that around 25% of the initial teacher population was replaced; the other one reports that even around 90% of the initial teacher population was replaced. According to most school-leaders, the replacement of teachers also led to uncertainty among school personnel.

The outcomes of the interviews provide suggestive evidence that the ASIP has increased teacher mobility. This finding is consistent with existing literature on the impact of school reforms on school personnel. Figlio and Loeb (2011) discuss the relationship between school accountability and teacher labour markets. They refer to interview and survey research providing evidence that teachers value a cohesive and supportive work environment that acknowledges their efforts and competences, while they interpret increased scrutiny and/or high-stakes testing as a reduction in their classroom autonomy and a message of being viewed as incompetent (e.g. Luna and Turner, 2001). The authors state that school reforms that influence these aspects of the work place are likely to affect teacher mobility. In addition, they discuss empirical studies providing evidence that accountability systems especially increase teacher attrition in schools that are labelled as low performing (see e.g. Feng et al., 2010). The ASIP, initially targeted at weak performing schools, increased scrutiny by lesson observations and may have contributed to the feeling of an unsupportive work atmosphere. In addition, the teacher evaluations may have helped the school-leaders to identify and replace ineffective teachers.

The increased mobility can negatively affect pupil's test scores in the short term via two mechanisms (Figlio and Loeb, 2011). First, recruitment and hiring of new teachers can take time and resources away from the regular instruction tasks. Second, the leave of (experienced) teachers implies a loss of specific knowledge on the school's way of working, instruction program and pupils. It takes time before new teachers have developed this knowledge. In addition to these mechanisms, the changing and uncertain work environment may have disturbed an optimal focus on primary instruction tasks among teachers.

In sum, the program appears to have created an unstable work atmosphere with increased teacher mobility. This may explain the negative impact on educational achievement in the first four years after its introduction. In the longer term, however, teacher mobility need not be detrimental if the least effective teachers are replaced by more effective teachers. In that case one might expect better performances once the more effective teachers are hired and the new work environment within schools has been stabilized for a while. Our main analyses concern the impact on a high-stakes test during the first four years after the introduction of the program. We do not find evidence for an improvement in the CITO test scores over these years. Nevertheless, we cannot exclude that these findings reflect adjustment costs of the reform policy and that it takes longer before more beneficial effects become manifest in the CITO test scores.

5.8 Conclusion

CSR methods have been widely used as an instrument to improve failing schools, but the evidence on its effectiveness remains limited. We estimate the effects of the ASIP, a CSR policy introduced in the Netherlands in 2008 with the goal of improving the educational quality of weak-performing primary schools in Amsterdam. The program implements performance-based working at all levels within the school and typically integrates measures such as staff coaching, teacher observations and teacher schooling, and the use of new instruction methods. Each program is tailored towards the specific needs of the school and is guided by educational experts.

Difference-in-differences estimates show substantial and statistically significant detrimental effects on the educational achievement of pupils in the highest grade of primary education. This finding is robust to a broad range of sensitivity tests. In our preferred specification, test scores decrease by 0.17 standard deviations in the first four years after the introduction of the policy. The overall negative effects are larger for language scores than for math scores. The size of the estimated effects varies across different parts of the test score distribution. The largest negative effects are generally found at the left part of the test score distribution, and the least detrimental effects at the upper tail of the test score distribution.

Interviews with school-leaders of participating primary schools reveal a candidate explanation for our findings. Although most of the school-leaders expected that the program would result in better educational quality in the longer term, they experienced the ASIP as an intensive program with a rigorous approach. It was especially confronting and demanding for teachers, who were judged based on lesson observations and expected to improve their competences. All required efforts had to be made in addition to their primary teaching tasks. Almost all school-leaders report the replacement of teachers after the introduction of the program. Some of them left the school voluntarily because of disagreement with the program, but others were forced to leave because they appeared not capable to satisfy the required standards. The outflow of teachers implies a loss of school specific knowledge and an increased focus on hiring new personnel that may have gone at the cost of instruction tasks. In addition, it seems to have created uncertainty among school personnel which can have disturbed an optimal focus on instruction tasks. Altogether, an increased teacher mobility induced by the program is a potential explanation for our negative findings on educational achievement. In that case, one might expect more beneficial effects in the longer term if less effective teachers are replaced by more effective ones and once the work environment within schools has been stabilized for a while. We do not find evidence for increasing effects over time in the first four years after the introduction of the policy. Nevertheless, we still cannot exclude that our findings reflect adjustment costs of the reform, and that it takes longer before beneficial effects become manifest in the CITO test scores. Even in such a case, one may question whether the future gains will outweigh the initial losses. In any case, we conclude that the introduction of the comprehensive school reform induced large costs in terms of a substantial decrease in educational performance for at least four cohorts of pupils.

Appendix 5.A: Additional results

Table 5.A.1 Heterogeneous treatment effects: Estimated effects of the introduction of the ASIP in the G38 en G4 samples

	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Low socio-economic status	High socio-economic status	Foreign	Dutch
A. G38						
Total score	-0.190***	-0.184***	-0.200***	-0.208***	-0.103*	-0.130**
	(0.046)	(0.044)	(0.076)	(0.052)	(0.060)	(0.052)
Language	-0.200***	-0.192***	-0.211***	-0.226***	-0.091	-0.147***
	(0.045)	(0.040)	(0.072)	(0.050)	(0.060)	(0.049)
Math	-0.135***	-0.134***	-0.098	-0.144**	-0.077	-0.063
	(0.044)	(0.048)	(0.085)	(0.053)	(0.061)	(0.051)
Information	-0.166***	-0.158***	-0.263***	-0.182***	-0.130**	-0.143***
processing	(0.049)	(0.043)	(0.081)	(0.053)	(0.064)	(0.054)
Observations	13,534	14,198	5,374	13,892	5,108	14,553
Schools	173	173	169	171	156	169
B. G4						
Total score	-0.169***	-0.120**	-0.051	-0.200***	-0.124	-0.095
	(0.058)	(0.055)	(0.095)	(0.064)	(0.076)	(0.062)
Language	-0.171***	-0.137***	-0.092	-0.202***	-0.081	-0.124**
	(0.058)	(0.049)	(0.086)	(0.062)	(0.073)	(0.059)
Math	-0.143***	-0.088	0.038	-0.152**	-0.146*	-0.027
	(0.054)	(0.060)	(0.102)	(0.063)	(0.079)	(0.060)
Information	-0.131**	-0.080***	-0.100	-0.193***	-0.126	-0.119*
processing	(0.061)	(0.056)	(0.105)	(0.065)	(0.079)	(0.064)
Observations	6,692	6,817	2,877	5,630	3,034	6,047
Schools	74	74	72	73	72	73
Individual characteristics	yes	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level.

Table 5.A.2 Quantile regressions results: Estimated effects of the introduction of the ASIP for quantiles of the test score distributions in the G38 and G4 samples

	(1)	(2)	(3)	(4)	(5)
	quantile	quantile	quantile	quantile	quantile
	regression	regression	regression	regression	regression
	0.1			0.75	0.90
		0.25	0.50		
A. G38					
Total score	-0.246***	-0.208***	-0.220***	-0.130***	-0.093***
	(0.042)	(0.033)	(0.031)	(0.029)	(0.034)
Language	-0.236***	-0.211***	-0.227***	-0.174***	-0.099***
	(0.047)	(0.035)	(0.030)	(0.030)	(0.032)
Math	-0.142***	-0.158***	-0.166***	-0.127***	-0.067**
	(0.042)	(0.040)	(0.034)	(0.030)	(0.030)
Information	-0.266***	-0.201***	-0.164***	-0.140***	-0.066**
processing	(0.047)	(0.036)	(0.033)	(0.029)	(0.029)
Observations	27,822	27,822	27,822	27,822	27,822
Schools	173	173	173	173	173
B. G4					
Total score	-0.156***	-0.159***	-0.172***	-0.130***	-0.086**
	(0.055)	(0.041)	(0.039)	(0.039)	(0.040)
Language	-0.135***	-0.166***	-0.219***	-0.144***	-0.065
	(0.050)	(0.043)	(0.039)	(0.039)	(0.040)
Math	-0.058	-0.155***	-0.166***	-0.116***	-0.086**
	(0.052)	(0.050)	(0.041)	(0.038)	(0.039)
Information	-0.217***	-0.117***	-0.090**	-0.132***	-0.039
processing	(0.058)	(0.046)	(0.041)	(0.039)	(0.038)
Observations	13,597	13,597	13,597	13,597	13,597
Schools	74	74	74	74	74
Individual characteristics	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes

Notes. Each cell represents a separate quantile regression. Standard errors are in parentheses. * / ** / *** denotes significance at a 10 / 5 / 1 % significance level.

Table 5.A.3 Estimated effects of the introduction of the ASIP: Different post-treatment years included. Sample without schools in Amsterdam that did not start in the ASIP in 2008-2009

	(1)	(2)	(3)	(4)
	Post-treatment years: 2009, 2010, 2011, 2012	Post-treatment years: 2010, 2011, 2012	Post-treatment years: 2011, 2012	Post-treatment years: 2012
Total score	-0.237***	-0.208***	-0.287***	-0.353***
	(0.067)	(0.071)	(0.081)	(0.122)
Language	-0.236***	-0.221***	-0.283***	-0.380***
	(0.065)	(0.069)	(0.081)	(0.116)
Math	-0.179***	-0.171**	-0.262***	-0.279**
	(0.066)	(0.069)	(0.076)	(0.117)
Information processing	-0.198***	-0.143**	-0.200***	-0.270**
	(0.063)	(0.069)	(0.078)	(0.114)
Observations	73,182	64,352	55,469	46,214
Schools	591	589	589	587
Individual characteristics	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level.

Table 5.A.4 Estimated effects of the introduction of the ASIP for separate post-treatment years

	(1)	(2)	(3)	(4)
	Post-treatment year: 2009	Post-treatment year: 2010	Post-treatment year: 2011	Post-treatment year: 2012
Total score	-0.189***	-0.094	-0.128***	-0.299***
	(0.053)	(0.059)	(0.050)	(0.057)
Language	-0.206***	-0.130**	-0.125***	-0.321***
	(0.049)	(0.056)	(0.045)	(0.053)
Math	-0.071	0.002	-0.109**	-0.204**
	(0.051)	(0.058)	(0.051)	(0.059)
Information processing	-0.225***	-0.143**	-0.100*	-0.274***
	(0.051)	(0.059)	(0.053)	(0.057)
Observations	48,711	48,788	49,172	49,462
Schools	575	586	599	609
Individual characteristics	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level.

Table 5.A.5 Estimated effects of the introduction of the ASIP for separate post-treatment years. Sample without schools in Amsterdam that did not start in the ASIP in 2008-2009

	(1) Post-treatment year: 2009	(2) Post-treatment year: 2010	(3) Post-treatment year: 2011	(4) Post-treatment year: 2012
Total score	-0.339***	-0.053	-0.226***	-0.353***
	(0.092)	(0.081)	(0.085)	(0.122)
Language	-0.299***	-0.088	-0.187**	-0.380***
	(0.093)	(0.074)	(0.079)	(0.116)
Math	-0.237**	-0.001	-0.256***	-0.279**
	(0.099)	(0.091)	(0.091)	(0.117)
Information processing	-0.361***	-0.042	-0.137*	-0.270**
	(0.070)	(0.088)	(0.078)	(0.114)
Observations	45,475	45,528	45,900	46,214
Schools	553	564	576	587
Individual characteristics	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level.

Table 5.A.6 Additional analyses: Grade retention

	(1) Total sample	(2) Sample excluding pupils with age > 12.5	(3) Sample excluding schools in Amsterdam with decrease in age > 0.2	(4) Sample excluding schools in Amsterdam with decrease in age > 0.15	(5) Sample excluding schools in Amsterdam with decrease in age > 0.10
	Dependent variable: age				
Introduction of the ASIP	-0.079*** (0.015)		-0.039** (0.015)	-0.026 (0.017)	-0.017 (0.018)
	Dependent variable: dummy variable indicating age > 12.5				
Introduction of the ASIP	-0.032*** (0.011)		-0.011 (0.012)	-0.001 (0.013)	0.002 (0.014)
	Dependent variable: total CITO test score				
Introduction of the ASIP	-0.170*** (0.036)	-0.139*** (0.037)	-0.174*** (0.038)	-0.140*** (0.039)	-0.159*** (0.044)
Observations	78,545	66,823	76,820	76,027	75,195
Schools	614	614	606	602	598
Individual characteristics	yes	yes	yes	yes	yes
School fixed effects	yes	yes	yes	yes	yes

Notes. Each cell represents a separate regression. Robust standard errors, clustered at the school-year level, are in parentheses. * / ** / *** denotes significance at a 1 / 5 / 10 % significance level.

6

Conclusions

This thesis aims to provide new evidence on the effectiveness of educational policies. It presents four empirical evaluations of educational policies in the Netherlands. The thesis applies design-based research methods in real policy cases and contributes to several topics in the literature of the economics of education. The main findings per chapter, including potential policy implications, are presented below, followed by a brief discussion on the limitations of the presented studies and suggestions for further research. The chapter ends with some overall concluding remarks.

6.1 Summary of main findings

Chapter 2 focuses on the effects of the timing of tracking on higher education completion. The age at which students are tracked into different education types is one of the most remarkable differences across countries. In our analysis, we use data from the Secondary Education Pupil Cohorts of 1989, and exploit the coexistence of a tracked and a comprehensive system within the Dutch education system. Some schools in the Netherland directly track 12 year old pupils that leave primary education into categorial classes of a particular education level, while other schools offer comprehensive classes that postpone the time of tracking with one or two years. This causes variation in the timing of tracking, which we exploit to analyse the effects of early tracking by comparing higher education completion of pupils who start secondary education in a categorial class (the ‘tracked’ pupils) to those who start in a comprehensive class (the ‘non-tracked’ pupils). To deal with selection problems we restrict our estimation sample to a particular school advice group that is homogeneous with respect to ability, use a large set of covariates, and adopt an instrumental variables approach. In the instrumental variable model we use the variation in tracking that is caused by the regional supply of school types. Our main analysis focuses on pupils at the margin of the Dutch high and low tracks. This group is likely to be most affected by the timing of tracking. We find negative effects of early tracking on completion of higher education. The OLS estimates are supported by the IV estimates. The OLS estimates show that early tracking decreases the probability of completion of higher education by approximately 5 percentage points. In the 1989 cohort, average completion of higher education for the tracked pupils is around 21%. Hence, our estimation results suggest that pupils can increase their probability of completing higher education by around 25% by postponing the timing of tracking. The IV estimates yield even more detrimental effects for early tracking. Additional analyses suggest that higher ability pupils are not negatively affected by being grouped in a comprehensive class together with the lower ability ones. This indicates that postponing tracking to older children would increase the number of people completing higher education. Our study contributes to the understanding of optimal tracking ages. It supports the conclusion of the OECD (2007) that the early tracking regime in the Netherlands causes a constraint for the growth of higher education participation. The results suggest that postponing tracking to children older than 12 would increase the number of people completing higher education in the Netherlands.

Chapter 3 contributes to the literature on the role of incentives in education. Since empirical work has found difficulties to show a relationship between increased resources and improved educational achievement, the literature on the use of incentives in education has expanded rapidly in recent years. Previous studies have mainly focused on incentives for students, teachers and schools. This study analyses the effectiveness of a new type of financial incentive scheme targeted at the level of school districts. We exploit the gradual introduction of the incentive scheme for local education authorities to reduce school dropout in the Netherlands in 2006. In the first year, covenants with financial incentives were offered to 14 out of 39 school districts with the highest number of dropouts. This selection rule, and a unique dataset containing individual information on the whole student population in both the pre- and post-treatment year, allows us to evaluate the effects of the treatment by means of a (local) difference-in-differences approach. We find statistically insignificant effects of the policy on dropout probability. These findings are robust to a variety of specifications and sensitivity tests. An additional analysis provides suggestive evidence for manipulation of outcomes in response to the program. This study provides no evidence that an incentive scheme targeted at regional education authorities has been effective in improving educational outcomes. The ineffectiveness of the scheme might be due to strategic behaviour at the cost of purposive actions. This once more points out the importance of a proper design with well-thought prior conditions for a successful implementation of incentive programs in education. Further research to improve knowledge on optimal designs seems highly valuable regarding the promise of incentive-based policies to increase educational quality at relatively low costs.

Chapter 4 investigates the effects of the Neighbourhood School Program (NSP), a comprehensive program for multi-problem school dropouts, by means of a field experiment. Previous studies provide mixed evidence on the success of training programs for disadvantaged youths. It seems therefore difficult to effectively serve the vulnerable target group of at-risk adolescents and realize persistent social gains. The NSP is largely designed in line with lessons from the literature and shares several components with other promising programs. It offers an intensive and integrated approach of educational and work services combined with professional care and personal mentoring. To assess the effects of the program we make use of capacity constraints at the NSP and implement a specific assignment rule such that treatment status is determined only by an individual's application date. We make use of administrative data on school enrolment, employment, and criminal behaviour, and investigate the impact of the program three years after its introduction. We estimate various regression models, which essentially compare outcomes between youths assigned to the NSP and youths assigned to standard intervention activities, conditional on the time of application. To address noncompliance with the assignment rule, we use an instrumental variables approach in which actual assignment is instrumented by intended assignment. We find statistically insignificant effects of assignment to the NSP on school enrolment or employment. This finding is in line with a large body of the literature that shows no impact of training programs for at-risk youths on labour market outcomes (LaLonde, 2003; Carneiro and Heckman, 2003). Most important, we find evidence that assignment to the NSP increases criminal activity compared to standard intervention, especially among the youths who were suspected of a crime at the time of entry. This result is consistent with previous studies that document adverse effects of group counselling or group-based interventions on criminal activity (Dishion et al., 1999). Deviant peer effects caused by grouping at-risk adolescents together can explain the reinforcement of criminal behaviour (Dodge et al., 2007). Additional analyses provide suggestive evidence in line with this explanation. Hence, adverse peer effects due to placing at-risk youths together may have negated other promising elements of the NSP. This study provides additional evidence that it is difficult to design effective programs that help vulnerable young dropouts back on track. In addition, the findings suggest that one should be cautious to treat youths-at-risk simultaneously, since peer effects can reinforce adverse behaviour.

Chapter 5 investigates the impact of comprehensive school reform (CSR) policy for failing schools on student achievement. CSR policies have been widely used as an instrument to improve failing schools, but the evidence on its effectiveness remains limited. We estimate the effects of the Amsterdam School Improvement Program (ASIP), a CSR policy introduced in the Netherlands in 2008. The program implements performance-based working at all levels within the school and typically integrates measures such as staff coaching, teacher observations and teacher schooling, and the use of new instruction methods. Each program is tailored towards the specific needs of the school and is guided by educational experts. Failing schools in Amsterdam were eligible for the program, whereas similar schools outside Amsterdam were not allowed to participate. Our assessment of student achievement is based on the high-stakes CITO test that pupils take in their final year of primary school. We make use of administrative data on CITO test scores from 2005 to 2012, which enables us to compare the development of student achievement in failing schools in Amsterdam to that in failing schools outside Amsterdam in a difference-in-differences framework. Difference-in-differences estimates show substantial and statistically significant detrimental effects on CITO test scores. This finding is robust to a broad range of sensitivity tests. In our preferred specification, test scores decrease by 0.17 standard deviations in the first four years after the introduction of the policy. Interviews with school-leaders of participating schools provide a candidate explanation for our findings. The rigorous and demanding approach appears to have caused an increase in teacher replacement. The resulting loss of school specific knowledge, increase in recruitment and hiring costs, and uncertain work atmosphere felt by teachers may have negatively affected pupil achievement. We cannot exclude that our findings reflect adjustment costs during the transition from a failing to a successful school and that it takes longer before beneficial effects become manifest. For instance, if teacher outflow indeed explains our findings, one might expect better performances once more effective teachers are hired and the new work environment within schools has been stabilized for a while. In any case, we conclude that the introduction of the CSR policy induced major costs in terms of substantial test score losses for at least four cohorts of pupils.

6.2 Limitations

The focus in the empirical studies in this thesis is on the identification of causal policy effects by using evaluation methods that address selection problems. These methods contribute to the credibility of the presented results. The presented studies also have their limitations.

First, the impact evaluations not always reveal the underlying mechanisms or critical factors that can explain the outcomes. For example, the analysis on the effects of tracking age is not informative on the mechanisms that translate early tracking to outcomes. Potential mechanisms include peer effects, teacher quality effects, or consequences for teachers related to the ability variance in the classroom. Low ability pupils may benefit from the interaction with their high-achieving peers in comprehensive classes. If better teachers prefer to teach relatively high ability classes, teacher sorting may result in a higher quality of education in these classes. In tracked classes with a lower ability variance, teaching may be easier since course material can be better adapted to the needs of the pupils. In addition to these differences between a tracked and a comprehensive class, our empirical results also reflect differences between tracking ages in the ability of educators to allocate students to appropriate school types. Tracking pupils at an early age may imply more uncertainty with respect to the pupil's true capabilities and hence a higher risk of sending them to an inappropriate school type. The role of such mechanisms remains a valuable topic for further research which may contribute to a more profound understanding of the effects of the timing of tracking.

Second, there may be concerns on the external validity of the presented results. The focus in our studies on the internal validity by addressing selection problems does not take away potential concerns on the predictive power of the results in a different context. Obviously, each empirical study provides evidence on the policy effects in the particular educational environment under consideration. If heterogeneity in responses is limited, these results are informative for other environments as well. If not, additional evidence from other studies can contribute to a more general picture of the effects of the policy. All of the studies in this thesis are concerned with policies implemented in the Dutch education system. It is not obvious that our findings can be directly translated to other countries as well. For example, the effectiveness of centralized incentives can be dependent on specific characteristics of the education system, such as the large school autonomy in the Netherlands. Since school autonomy might limit the power of education authorities, it is conceivable that this type of incentive scheme will be more successful in other education systems. Additional research efforts across countries can shed light on the effectiveness of incentive programs in different education systems.

6.3 Concluding remarks

The findings in this thesis provide insight into the effects of various types of educational policies. Our results point out that it is by no means obvious that policy interventions always work out the way they were meant to. We find that two intensive programs for troubled youths and schools did not succeed in improving outcomes. Instead, for both programs we find adverse effects. Such programs require substantial resources and human efforts. Our results fit in a large body of studies that show that additional resources do not guarantee an improvement of educational outcomes (Hanushek, 2006; Woessmann 2003). These findings can explain the current interest in other, more cost-effective ways to improve the education system. Policies based on incentives are generally less costly compared to resource-based interventions. Such policies include incentives for teachers and students (Angrist and Lavy, 2009; Kremer et al., 2009; Lavy, 2002, 2009) as well as accountability systems for schools (Hanushek and Raymond, 2005; Jacob, 2005; Dee and Jacob, 2011). This thesis contributes to this literature by investigating a new type of financial incentives for regional educational authorities. We find no evidence that the incentive scheme improved educational outcomes. In addition to incentive-based policies, policies targeted at institutional features of the education system have the potential to improve outcomes at negligible costs. This thesis provides evidence that the number of people that graduate from higher education can be increased by postponing the age at which students are tracked into different types of secondary education from 12 to 13 or 14. Such institutional changes would require (almost) no additional resources. Naturally, the findings in this thesis do not imply that intensive programs or resource-based policies will never work. They illustrate that the more expensive interventions not necessarily perform better than interventions that require less resources. This suggests that putting in additional resources is not always the best way to address problems or to improve the system. The findings that policy instruments can have no, or even detrimental, effects on outcomes point out the importance of a proper knowledge on the impact of policy interventions. Now that the empirical evaluation methods have largely matured and found credible ways to address selection problems, policymakers can benefit from the insights provided by new evaluation studies. These insights can be used to adjust the content of interventions, to select the most promising interventions or to abandon policies that have been shown to be ineffective. In this way additional design-based evaluations can valuably contribute to a more efficient allocation of resources and an improvement of the education system.

Bibliography

Acemoglu, D., J. Angrist, 2001, How large are human capital externalities? Evidence from Compulsory Schooling Laws, in B. Bernanke and K. Rogoff, eds., *NBER macroeconomics annual 2000*. Cambridge, MA: MIT Press, 9-59.

Ammermüller, A., J.-S. Pischke, 2009, Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study, *Journal of Labor Economics*, 27 (3), 315-348.

Angrist, J., 1990, Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records, *American Economic Review*, 80 (3), 313-336.

Angrist, J. , E. Bettinger, E. Bloom, B. King, M. Kremer, 2002, Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment, *American Economic Review*, 92 (5), 1535–58.

Angrist, J., G. Imbens, D. Rubin, 1996, Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association*, 91 (434), 444–55.

Angrist, J., A. Krueger, 1991, Does compulsory school attendance affect schooling and earnings, *Quarterly Journal of Economics*, 106 (4), 979-1014.

Angrist, J., A. Krueger, 2001, Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments, *Journal of Economic Perspectives*, 15 (4), 69-85.

Angrist, J., D. Lang, P. Oreopoulos, 2009, Incentives and Services for College Achievement: Evidence from a Randomized Trial, *American Economic Journal: Applied Economics*, 1(1), 136-163.

Angrist, J., V. Lavy, 1999, Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement, *Quarterly Journal of Economics*, 114 (2), 533-575.

Angrist, J., V. Lavy, 2009, The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial, *American Economic Review*, 99 (4), 1384-1414.

Angrist, J., J. Pischke, 2009, *Mostly Harmless Econometrics*, Princeton University Press, Princeton.

Angrist, J., J. Pischke, 2010, The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics, NBER Working Paper 15794.

Ariga, K., G. Brunello, 2007, Does secondary school tracking affect performance? Evidence from IALS, IZA Discussion Paper 2643.

Ashenfelter, O., 1978, Estimating the effect of training programs on earnings, *Review of Economics and Statistics*, 60, 47-50.

Ashenfelter, O., D. Card, 1985, Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs, *Review of Economics and Statistics*, 67 (4), 648-660.

Ashenfelter, O., C. Harmon, H. Oosterbeek, 1999, A review of estimates of the schooling/earnings relationship with tests for publication bias, *Labour Economics*, 6 (4), 453 - 470.

Becker, G., 1964, *Human capital: A theoretical and empirical analysis with special reference to education*, First Edition ed. New York, NY: National Bureau of Economic Research.

Bertrand, M., E. Duflo, S. Mullainathan, 2004, How Much Should We Trust Differences-in-Differences Estimates, *Quarterly Journal of Economics*, 119, 249-275.

Betts, J., J. Shkolnik, 2000, The effects of ability grouping on student math achievement and resource allocation in secondary schools, *Economics of Education Review*, 19 (1), 1–15.

Bifulco, R., W. Duncombe, J. Yinger, 2005, Does whole-school reform boost student performance? The case of New York City, *Journal of Policy Analysis and Management*, 24, 47-72.

Blair, T., J. Cunningham,1999, *Modernising Government*, Prime Minister and Minister for the Cabinet Office, London, UK.

Bloom, H., 1984, Accounting for No-shows in Experimental Evaluation Designs, *Evaluation Review*, 8, 225-246.

Bloom, H., L. Orr, S. Bell, G. Cave, F. Doolittle, W. Lin, J. Bos, 1997, The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study, *Journal of Human Resources*, 32, 549-576.

Blundell, R., A. Duncan, C. Meghir, 1998, Estimating Labor Supply Responses Using Tax Reforms, *Econometrica*, 66(4), 827-861.

Borman, G., G. Hewes, L. Overman, S. Brown, 2003, Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research,* 73 (2), 125–230.

Borman, G., G. Hewes, L. Overman, S. Brown, 2004, Comprehensive school reform and achievement: A meta-analysis. In C.Cross, ed., *Putting the pieces together: Lessons from comprehensive school reform research* (pp. 53–108). Washington, DC: The National Clearinghouse for Comprehensive School Reform.

Bound, J., D. Jager, R. 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, 443-450.

Browning, M., E. Heinesen, 2007, Class Size, Teacher Hours and Educational Attainment, *Scandinavian Journal of Economics*, 109, 415–438.

Brunello, G., D. Checchi, 2007, Does tracking affect equality of opportunity? New International Evidence, *Economic Policy*, 52, 781-861.

Burgess, S., C. Propper, M. Ratto, E. Tominey, 2004, Incentives in the Public Sector: Evidence from a Government Agency, CMPO Working Paper Series No. 04/103.

Card, D., 1995, Using Geographic Variation in College Proximity to Estimate the Return to Schooling, in L.N. Christofides, E. Kenneth Grant, and R. Swidinsky, eds., *Aspects of labour market behaviour: Essays in honour of John Vanderkamp.* Toronto: University of Toronto Press, 201-222.

Card, D., 1999, The causal effect of education on earnings. In O. Ashenfelter, D. Card, eds., *Handbook of Labor Economics,* Vol 3A, Amsterdam: Elsevier, 1801-1863.

Card, D., A. Krueger, 1994, Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, *American Economic Review*, 84 (4), 772-793.

Carneiro, P., J. Heckman, 2003, Human Capital Policy, NBER Working Paper 9495.

Cave, G., F. Doolittle, 1991, *Assessing JOBSTART: Interim Impacts of a Program for School Dropouts*, New York: Manpower Demonstration Research Corporation.

Coe, D., E. Helpman, A. Hoffmaister, 2009, International R&D spillovers and institutions, *European Economic Review*, 53, 723 – 741.

Cohen, D., M. Soto, 2007, Growth and human capital: good data, good results, *Journal of Economic Growth*, 12, 51-76.

Cook, T., 2008, "Waiting for Life to Arrive": A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics, *Journal of Econometrics*, 142 (2), 636-654.

Cook, T., E. Habib, M. Phillips, R. Settersten, S. Shagle, S. Degirmencioglu, 1999, Comer's School Development Program in Prince George's County, Maryland: A theory-based evaluation, *American Education Research Journal*, 36 (3), 543-597.

Cook, T., H. Hunt, R. Murphy, 1998, Comer's School Development Program in Chicago: A theory-based evaluation. WP-98-24. Evanston, IL: Institute for Policy Research, Northwestern University.

Couch, K., 1992, New Evidence on the Long-Term Effects of Employment and Training Programs, *Journal of Labor Economics,* 10, 380–88.

Currie, J., E. Moretti, 2003, Mother's education and the intergenerational transmission of human capital: Evidence from college openings, *Quarterly Journal of Economics*, 118 (4), 1495–1532.

Dearden, L., C. Emmerson, C. Frayne, C. Meghir, 2009, Conditional Cash Transfers and School Dropout Rates, *Journal of Human Resources*, 44 (4), 827-857.

Dee, T., B. Jacob, 2011, The Impact of No Child Left Behind on Student Achievement, *Journal of Policy Analysis and Management*, 30 (3), 418-446.

De la Fuente, A., R. Doménech, 2006, Human capital in growth regressions: how much difference does data quality make?, *Journal of the European Economic Association*, 4, 1-36.

Deloitte Accountants, 2006, *Audit over het gebruik van de informatiebronnen voortijdig schoolverlaten*.

Dishion, T., J. McCord, F. Poulin, 1999, When Interventions Harm: Peer Groups and Problem Behaviour, *American Psychologist,* 54, 755–64.

Dodge, K., T. Dishion, J. Lansford (eds.), 2007, *Deviant Peer Influences in Programs for Youth: Problems and Solutions*, New York: Guilford Press.

Donald, S., K. Lang, 2007, Inference with Difference-in-Differences and Other Panel Data, *Review of Economics and Statistics*, 89, 221-233.

Driessen, G., L. Mulder, G. Ledoux, J. Roeleveld, I. van der Veen, 2009, *Cohortonderzoek COOL5-18. Technisch rapport basisonderwijs, eerste meting 2007/08.* Nijmegen: ITS / Amsterdam: SCO-Kohnstamm Instituut.

Driessen, G., L. Mulder, J. Roeleveld, 2012, *Cohortonderzoek COOL5-18. Technisch rapport basisonderwijs, tweede meting 2010/11.* Nijmegen: ITS / Amsterdam: SCO-Kohnstamm Instituut.

Driessen, G., M. van der Werf, 1992, Het functioneren van het voortgezet onderwijs: beschrijving steekproef en psychometrische kwaliteit van de instrumenten, Groningen/ Nijmegen: RION/ITS.

Duflo, E., P. Dupas, M. Kremer, 2008, Peer effects and the impact of tracking: Evidence from a Randomized Evaluation in Kenya, NBER Working Paper 14475.

Eisenkopf, G., 2007, Tracking and incentives. A comment on Hanushek and Woessman, Thurgau Institute of Economics, Research Paper Series No. 22.

Eisenkopf, G., 2010, Peer effects, motivation and learning, *Economics of Education Review*, 29, 364-374.

Feng, L., D. Figlio, T. Sass, 2010, School Accountability and Teacher Mobility, NBER Working Paper 16070.

Figlio, D., L. Getzler, 2006, Accountabiliy, ability and disability: Gaming the system? In T. Gronberg, D. Jansen, eds., *Advances in Microeconomics, Vol. 14: Improving School Accountability - Checkups or Choice?*, Amsterdam: Elsevier, 35-49.

Figlio, D., M. Page, 2002, School choice and the distributional effects of ability tracking: does separation increase inequality?, *Journal of Urban Economics*, 83, 497-514.

Figlio, D., S. Loeb, 2011, School Accountability. In E. Hanushek, S. Machin, L. Woessmann, eds., *Handbook of the Economics of Education,*Vol. 3, Amsterdam: Elsevier, 383-421.

Frey, B., F. Oberholzer-Gee, 1997, The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-out, *American Economic Review*, 87, 746-755.

Fryer, R., 2011, Financial Incentives and Student Achievement: Evidence from Randomized Trials, *Quarterly Journal of Economics,* 126(4), 1755-1798.

Fryer, R., 2011, Teacher incentives and student achievement: Evidence from New York City Public Schools, NBER Working Paper 16850.

Galindo-Rueda, F., A. Vignoles, 2005, The heterogeneous effect of selection in secondary schools: Understanding the changing role of ability, CEE Discussion Paper No. 52.

Glaeser, E., B. Sacerdote, J. Scheinkman, 1996, Crime and social interactions, *Quarterly Journal of Economics*, 111(2), 507–48.

Glewwe, P., N. Ilias, M. Kremer, 2003, Teacher incentives, NBER Working Paper 9671.

Granger, R., R. Cytron, 1998, *Teenage Parent Programs a Synthesis of the Long-Term Effects of the New Chance Demonstration, Ohio's Learning, Earning and Parenting (LEAP) Program, and the Teenage Parent Demonstration (TPD)*, Manpower Demonstration Research Corporation.

Gross, B., T. K. Brooker, D. Goldhaber, 2009, Boosting Student Achievement: The Effect of Comprehensive School Reform on Student Achievement, *Educational Evaluation and Policy Analysis*, 31 (2), 111-126.

Hahn, J., P. Todd, W. van der Klaauw, 2001, Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design, *Econometrica*, 69 (1), 201-209.

Hanushek, E., 1986, The Economics of Schooling: Production and Efficiency in Public Schools, *Journal of Economic Literature*, 24 (3), 1141–1177.

Hanushek, E., 2006, School Resources. In E. Hanushek, F. Welch, eds., *Handbook of the Economics of Education*, Vol. 2, Amsterdam: Elsevier, 865-908.

Hanushek, E., 2011, The economic value of higher teacher quality, *Economics of Education Review*, 30, 466-479.

Hanushek, E., M. Raymond, 2004, The effect of school accountability systems on the level and distribution of student achievement, *Journal of the European Economic Association*, 2 (2-3), 406–415.

Hanushek, E., M. Raymond, 2005, Does School Accountability Lead to Improved Student Performance?, *Journal of Policy Analysis and Management*, 24 (2), 297-327.

Hanushek, E., S. Rivkin, 2006, Teacher quality. In E. Hanushek, F. Welch, eds., *Handbook of Economics of Education*, Vol. 2, Amsterdam: Elsevier, 1051-1078.

Hanushek, E., L. Woessmann, 2006, Does education tracking affect performance and inequality? Differences in differences evidence across countries, *The Economic Journal*, 116, C63-C76.

Hanushek, E., L. Woessmann, 2008, The role of cognitive skills in economic development, *Journal of Economic Literature*, 46 (3), 607–668.

Hanushek, E., L. Woessmann, 2012, Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation, *Journal of Economic Growth*, 17, 267-321.

Harmon C., H. Oosterbeek, I. Walker, 2003, The returns to education: microeconomics, *Journal of Economic Surveys*, 17, 115-155.

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

Heckman, J., L. Lochner, P. Todd, 2006, Earnings functions, rates of return, and treatment effects: The Mincer equation and beyond. In E. Hanushek, F.Welch, *Handbook of the Economics of Education,* Vol. 1, Amsterdam: Elsevier, 307 - 458.

Heller, S., H. Pollack, R. Ander, J. Ludwig, 2013, Preventing Youth Violence and Dropout: A Randomized Field Experiment, NBER Working Paper 19014.

Herman, R., D. Aladjem, P. McMahon, E. Masem, I. Mulligan, A. O'Malley, S. Quinones., A. Reeve, D. Woodruff, 1999, *An educator's guide to schoolwide reform*. Arlington, VA: American Institutes for Research.

Hill, P., B. Roberts, J. Grogger, J. Guryan, K. Sixkiller, 2011, Decreasing Delinquency, Criminal Behavior, and Recidivism by Intervening on Psychological Factors Other than Cognitive Ability: A Review of the Intervention Literature, NBER Working Paper 16698.

Holmstrom, B., P. Milgrom, 1991, Multi-Task Principal-Agent Problems: Incentive Contracts, Asset Ownership and Job Design, *Journal of Law, Economics and Organization*, 7, 24-52.

Imbens, G., J. Angrist, 1994, Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62 (2), 467–75.

Imbens, G., T. Lemieux, 2008, Regression Discontinuity Designs: A Guide to Practice, *Journal of Econometrics*, 142 (2), 615-635.

Imbens, G., J. Wooldridge, 2009, Recent Developments in the Econometrics of Program Evaluation, *Journal of Economic Literature*, 47 (1), 5-86.

Inspectorate of Education, 2008, *Regionale analyse; Een analyse van het Amsterdamse basisonderwijs*. Utrecht.

Inspectorate of Education, 2009, *De Staat van het Onderwijs, Onderwijsverslag 2007/2008.* Utrecht.

Jacob, B., 2005, Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools, *Journal of Public Economics*, 89, 761-796.

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

Jacob, B., S. Levitt, 2003, Rotten apples: An Investigation of the Prevalence and Predictors of Teacher Cheating, *Quarterly Journal of Economics*, 3, 843-877.

Jakubowski, M., 2007, Effects of tracking on achievements growth, exploring difference-in-difference approach to PIRLS, TIMMS and PISA data, unpublished manuscript.

Johnson, A., 1999, *Sponsor-A-Scholar: Long-term Impacts of a Youth Mentoring Program on Student Performance*, Princeton: Mathematica Policy Research.

Kane, T., C. Rouse, 1995, Labor-Market Returns to Two- and Four-Year College, *American Economic Review,* 85 (3), 600-614.

Kremer, M., E. Miguel, R. Thornton, 2009, Incentives to Learn, *Review of Economics and Statistics*, 91(3), 437-456.

Krueger, A., 1999, Experimental estimates of education production functions, *Quarterly Journal of Economics*, 114, 497–532.

Ladd, H., R. Walsh, 2002, Implementing value-added measures of school effectiveness: Getting the incentives right, *Economics of Education Review*, 21, 1-17.

Lalonde, R., 1986, Evaluating the econometric evaluations of training programs with experimental data, *American Economic Review*, 76 (4), 604-620.

LaLonde, R., 2003. Employment and training programs. In: Feldstein, M., Moffitt, R. (Eds.), *Means Tested Transfer Programs in the U.S.*, Chicago, IL: University of Chicago Press for the National Bureau of Economic Research.

Lavy, V., 2002, Evaluating the effect of teachers' group performance incentives on pupil achievement, *Journal of Political Economy*, 110 (6), 1286-1317.

Lavy, V., 2009, Performance Pay, Teachers' Effort, Productivity and Grading Ethics, *American Economic Review*, 99 (5), 1979-2021.

Lazear, E.P., 1999, Personnel economics: Past lessons and future directions, Presidential address to the Society of Labor Economists, San Francisco, May 1, 1998, *Journal of Labor Economics*, 17 (2), 199-236.

Lazear, E., S. Rosen, 1981, Rank-order tournaments as optimum labor contracts, *Journal of Political Economy*, 89 (5), 841-864.

Leuven, E, M. Lindahl, H. Oosterbeek, D. Webbink, 2007, The effect of extra funding for disadvantaged pupils on achievement, *Review of economics and statistics*, 89 (4), 721-736.

Lleras-Muney, A., 2005, The relationship between education and adult mortality in the United States, *Review of Economic Studies*, 72 (1), 189-221.

Lochner, L., E. Moretti, 2004, The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports, *American Economic Review*, 94 (1), 155-89.

Luna, C., Turner, C., 2001. The impact of the MCAS: Teachers talk about high-stakes testing, *English Journal*, 91 (1): 79–87.

Machin, S., O. Marie, S. Vujic, 2011, The Crime Reducing Effect of Education, *The Economic Journal*, 121, 463-84.

Maxfield, M., A. Schirm, N. Rodriguez-Planaz, 2003, *The Quantum Opportunities Program Demonstration: Implementation and Short-Term Impacts*, Washington, DC: Mathematica Policy Research.

May, H., J. Supovitz, 2006, Capturing the Cumulative Effects of School Reform: An 11-Year Study of the Impacts of America's Choice on Student Achievement, *Educational Evaluation and Policy Analyis*, 2006, 28 (3), 231-257.

Manning, A., J. Pischke, 2006, Comprehensive versus selective schooling in England and Wales: What do we know?, IZA Discussion Paper 2072.

Millenky, M., D. Bloom, S. Muller-Ravett, J. Broadus, 2011, Staying on Course: Three-Year Results of the National Guard Youth Challenge Evaluation, Manpower Demonstration Research Corporation.

Millsap, M., A. Chase., D. Obiedallah, A. Perez-Smith, 2001, Evaluation of the Comer School Development Program in Detroit, 1994-1999: Methods and results. Washington, DC.

Mincer, J., 1958, Investment in human capital and personal income distribution, *Journal of Political Economy*, 66, 281 - 302.

Mincer, J., 1962, On the job training: Costs, returns and some implications, *Journal of Political Economy*, 70, 50 - 79.

Ministry of Education, 2008a, *Aanval op de uitval: Uitvoeren en doorzetten*.

Ministry of Education, 2008b, *Evaluatie convenantactie voortijdig schoolverlaten schooljaar 2006-2007*, Directorate BVE.

Ministry of Education, 2010, *Bijlage VSV-brief 2010, Nieuwe voortijdig schoolverlaters convenantjaar 2008-2009.*

Minne, B., M. Rensman, B. Vroomen, D. Webbink, 2007, Excellence for productivity?, The Hague: CPB Netherlands Bureau for Economic Policy Analysis.

Moulton, B., 1986, Random Group Effects and the Precision of Regression Estimates, *Journal of Econometrics*, 32 (3), 385–97.

Municipality of Amsterdam, 2009, *Kwaliteitsaanpak Basisonderwijs Amsterdam; programmaplan 2009-2014.* Amsterdam.

Municipality of Rotterdam, 2011,*'Ga gewoon door!' De Rotterdamse aanpak van de jeugdwerkloosheid 2011-2012.*

OECD, 2007, Thematic Review of Tertiary Education – The Netherlands, Country Note.

Oreopoulos, P., 2006, Estimating average and local average treatment effects of education when compulsory schooling laws really matter, *American Economic Review*, 96 (1), 152-175.

Oreopoulos, P., 2007, Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling, *Journal of Public Economics*, 91, 2213-2229.

Orr, L., H. Bloom, S. Bell, W. Lin, G. Cave, F. Doolittle, 1994, *The National JTPA Study: Impacts, Benefits, and Costs of Title II-A. Report to the U.S. Department of Labour*, Cambridge, MA: Abt Associates.

Park, C., C. Kang, 2008, Does education induce healthy lifestyle?, *Journal of Health Economics*, 27 (6), 1516-1531.

Parliamentary investigation committee on education reform, 2008, *Tijd voor onderwijs*, The Hague.

Pekkarinen, T., R. Uusitalo, S. Kerr, 2009, School tracking and development of cognitive skills, IZA Discussion Paper No. 4058.

Pischke, J-S., 2007, The Impact of Length of the School Year on Student Performance and Earnings: Evidence from the German Short School Years, *Economic Journal*, 117, 1216-1242.

Rees, D., D. Brewer, L. Argys, 2000, How should we measure the effect of ability grouping on student performance? *Economics of Education Review*, 19 (1), 16–20.

Research voor Beleid, 2008, RMC Analyse 2007; voortijdig schoolverlaten en RMC functie 2006/2007.

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

Roder, A., M. Elliot, 2011, A Promising Start: Year-Up's Initial Impacts of Young Adults' Careers, Economic Mobility Corporation.

Rodriguez-Planas, N., 2012, Longer-Term Impacts of Mentoring, Educational Services, and Learning Incentives: Evidence from a Randomized Trial in the United States, *American Economic Journal: Applied Economics,* 4 (4), 121–39.

Rowan, B., C. Barnes, E. Camburn, 2004, Benefiting from comprehensive school reform: A review of research on CSR implementation. In C. T. Cross (Ed.), *Putting the pieces together: Lessons from comprehensive school reform research*, 1–52. Washington, DC:National Clearinghouse for Comprehensive School Reform.

Rubin, D., 1974, Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, 66, 688–701.

Rubin, D., 1977, Assignment to treatment group on the basis of a covariate, *Journal of Educational Statistics*, 2, 1–26.

Sardes, 2006, *De uitkomsten van de RMC analyse 2005*, Utrecht.

Schochet, P., J. Burghardt, S. Glazerman, 2001, *National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes*, Washington, DC: U.S. Department of Labour, Employment and Training Administration.

Schochet, P., J. Burghardt, S. McConnell, 2008, Does Job Corps Work? Impact Findings from the National Job Corps Study, *American Economic Review* 98 (5), 1864–86.

Schuetz, G., H. Ursprung, L. Woessmann, 2005, Education policy and equality of opportunity, CESIFO Working Paper 1518.

Schwartz, A., L. Stiefel, D. Kim, 2004, The impact of school reform on student performance: Evidence from the New York Network for School Renewal Project, *Journal of Human Resources*, 39 (2), 500-522.

Slavin, R., O. Fashola, 1998, *Show me the evidence!* Thousand Oaks, CA: Corwin Press.

Staiger, D., J. Stock, 1997, Instrumental variables regression with weak instruments, *Econometrica*, 65 (3), 557-586.

Springer, M., D. Ballou, L. Hamilton, V. Le, J. Lockwood, D. McCaffrey, M. Pepper, B. Stecher, 2010, Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching, Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Statistics Netherlands, 1991, Schoolloopbaan en achtergrond van leerlingen: cohort 1989, Voorburg/ Heerlen: CBS.

Statistics Netherlands, 2011, http://www.cbs.nl/nl-NL/menu/themas/arbeid-sociale-zekerheid/publicaties/artikelen/archief/2011/2011-05-30-jeugdwerkloosheid-tk.htm

Sund, K., 2009, Estimating peer effects in Swedish high school using school, teacher and student fixed effects, *Economics of Education Review*, 28, 329-336.

Taggart, R., 1995, *The Quantum Opportunities Program: Second Post-Program Year Impacts*, Philadelphia: Opportunities Industrialization Centers of America.

Tierney, J., J. Grossman, N. Resch, 1995, *Making a Difference: An Impact Study of Big Brothers Big Sisters*, Philadelphia: Public/Private Ventures.

Traub, J., 1999, *Better by design? A consumer's guide to schoolwide reform*. Washington, DC: Thomas B. Fordham Foundation.

Urquiola, M., 2006, Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia, *Review of Economics and Statistics*, 88, 171–176.

U.S. Department of Education, 2004, *Implementation and early outcomes of the Comprehensive School Reform Demonstration (CSRD) Program* (No. 2004-15). Jessup, MD: Policy and Program Studies Service, U.S. Department of Education.

U.S. Department of Education, 2006, *Comprehensive school reform program: Funding status*. Washington, DC: Author.

Van der Klaauw, W., 2008, Regression Discontinuity Analysis: A Survey of Recent Developments in Economics, Labour, 22 (2), 219-245.

Van der Steeg, M., S. Gerritsen, 2013, Teacher evaluations and pupil achievement; Evidence from classroom observations, CPB Discussion Paper 230.

Webbink, D., P. Koning, S. Vujic, N. Martin, 2013, Why Are Criminals Less Educated than Non-Criminals? Evidence from a Cohort of Young Australian Twins, *Journal of Law, Economics, and Organization*, 29 (1), 115-144.

Woessmann, L., 2003, Schooling resources, educational institutions and student performance: The international evidence, *Oxford Bulletin of Economics and Statistics*, 65 (2), 117-170.

Samenvatting (Summary in Dutch)

Motivatie en doel

Dit proefschrift richt zich op het identificeren van effectief beleid in het onderwijs. Dit is belangrijk vanwege meerdere redenen. Ten eerste heeft het recente verleden laten zien dat de invoering van nieuw beleid niet altijd even succesvol is geweest. De parlementaire onderzoekscommissie onderwijsvernieuwingen, onder leiding van de huidige minister van Financiën Jeroen Dijsselbloem, evalueerde in 2008 enkele van de meest ingrijpende vernieuwingen in het onderwijs in de jaren negentig van de vorige eeuw. Het ging daarbij om de invoering van het vmbo, de basisvorming en de tweede fase. De commissie concludeerde dat de overheid haar kerntaak, het zeker stellen van deugdelijk onderwijs, ernstig had verwaarloosd. De commissie merkte onder meer op dat verschillende - mogelijk conflicterende - wijzigingen tegelijkertijd werden ingevoerd en dat beleidskeuzen vaak werden bepaald door het financiële kader. Daarnaast wees zij op het gebrek aan een grondige probleemanalyse en wetenschappelijke onderbouwing voor de nieuwe maatregelen. Eén van de adviezen voor de toekomst betrof het gebruik van beleidsevaluaties om nieuw beleid wetenschappelijk te valideren. Wanneer een dergelijke onderbouwing ontbreekt, zou nieuw beleid eerst kleinschalig moeten worden ingevoerd ten behoeve van een wetenschappelijke evaluatie.

De toegenomen aandacht voor de wetenschappelijke onderbouwing van beleid is zichtbaar in meer landen. In het Verenigd Koninkrijk bepleitte de regering Blair onder het motto ‘what matters, is what works’ zich bij beslissingen niet langer te laten leiden door politiek-ideologische argumenten, maar door objectieve wetenschappelijke analyses (Blair en Cunningham, 1999). In de Verenigde Staten legde de regering Bush met de ‘No Child Left Behind-wet’ uit 2001 de nadruk op het gebruik van onderwijsmethoden- en activiteiten waarvan de werking wetenschappelijk is aangetoond. Het invoeren van ineffectief beleid kan kostbaar zijn en ongewenste gevolgen met zich meebrengen voor de onderwijskwaliteit.

Ten tweede is onderwijs een belangrijke determinant van economisch en maatschappelijk succes. De interesse van economen in onderwijs gaat terug naar studies over menselijk kapitaal in de jaren ’50 en ’60 van de vorige eeuw (Becker, 1964; Mincer, 1958, 1962). De theorie van het menselijk kapitaal beschouwt onderwijs als een investering in het verwerven van kennis van vaardigheden. De kosten die hiermee gepaard gaan, betreffen onder meer de uitgaven voor het volgen van onderwijs en het misgelopen loon vanwege het feit dat men niet werkt, maar op school zit. Een belangrijke bate van de investering is het hogere toekomstige loon ten gevolge van de verworven vaardigheden. Sinds de eerste studies over menselijk kapitaal is een omvangrijke literatuur ontstaan die het belang van onderwijs voor economische en maatschappelijke uitkomsten aantoont. Hoger opgeleiden hebben in het algemeen een hoger loon (Card, 1999; Ashenfelter et al, 1999), zijn gezonder (Lleras-Muney, 2005; Oreopoulos, 2007) en vertonen minder crimineel gedrag (Lochner en Moretti, 2004; Machin et al., 2011; Webbink et al., 2012). Een groot aantal empirische studies vindt private rendementen van onderwijs tussen de 5 en 10% (Card, 1999; Ashenfelter et al, 1999; Harmon et al., 2003; Heckman et al., 2006). Dit betekent dat het volgen van een extra jaar onderwijs leidt tot een 5 tot 10% toename van inkomsten over de levensloop. Daarnaast wijst een toenemend aantal studies op het belang van onderwijs voor economische groei. De meeste van deze studies relateren het gemiddeld behaalde onderwijsniveau in een land aan de economische groei en vinden sociale rendementen die iets hoger liggen dan de private rendementen (De la Fuente en Domenench 2006; Cohen en Soto 2007; Coe et al. 2009). Meer recente studies hebben zich ook gericht op de rendementen op cognitieve vaardigheden. Deze cognitieve vaardigheden worden gemeten door middel van toetsscores. Het meten van de hoeveelheid menselijk kapitaal op basis van cognitieve vaardigheden in plaats van het behaalde onderwijsniveau maakt het mogelijk om rekening te houden met verschillen in de onderwijskwaliteit tussen landen. Deze studies verschaffen bewijs dat cognitieve vaardigheden sterk gerelateerd zijn aan individuele opbrengsten en de economische groei van landen (Hanushek en Woessmann, 2008, 2012).

Het belang van menselijk kapitaal voor economisch en maatschappelijk succes verklaart de wereldwijde belangstelling voor goed onderwijsbeleid. In de meeste landen is de voorziening en financiering van onderwijs voornamelijk een taak van de overheid. Beleidsmakers hebben te maken met beperkte budgetten en moeten belangrijke keuzes maken over de inzet van schaarse middelen. Dit vereist kennis over de werking van verschillende beleidsopties. Beleidsevaluaties kunnen inzicht verschaffen in de effecten van overheidsinterventies op onderwijsuitkomsten. Deze inzichten kunnen gebruikt worden voor het versterken van het onderwijsbeleid door effectief gebleken maatregelen te introduceren of uit te breiden, of door ineffectief gebleken maatregelen aan te passen of stop te zetten.

De complexe onderwijsproductiefunctie bemoeilijkt een betrouwbare beleidsevaluatie. Onderwijsuitkomsten worden bepaald door diverse factoren die het leerproces van leerlingen of studenten beïnvloeden, zoals kenmerken van de school, de leraar, ouders of familie, medeleerlingen en omgeving (Hanushek, 1986). De veelheid aan observeerbare en niet observeerbare factoren die onderwijsuitkomsten beïnvloeden, maakt het niet eenvoudig om het oorzakelijk effect van een specifieke beleidsinterventie te isoleren. In de empirische economische literatuur is de afgelopen decennia veel aandacht geweest voor het vaststellen van causale effecten van beleid. Dit heeft geleid tot een verandering van onderzoeksmethoden op basis van controlestrategieën naar methoden op basis van specifieke onderzoeksdesigns die goede oplossingen kunnen bieden voor dit probleem (Imbens en Wooldridge, 2009; Angrist en Pischke, 2010). Met behulp van deze 'design-based' onderzoeksmethoden kunnen de oorzakelijke effecten van beleid op een geloofwaardige wijze worden vastgesteld.

Dit proefschrift beoogt een bijdrage te leveren aan de zoektocht naar effectieve beleidsmaatregelen in het onderwijs. Het proefschrift presenteert vier empirische evaluaties van onderwijsbeleid. Hierbij worden moderne onderzoeksdesigns toegepast op concrete beleidsinterventies in Nederland. In elk van de studies ligt de nadruk op een geloofwaardige identificatie van het oorzakelijk effect van het beleid. Het proefschrift draagt bij aan verschillende stromingen binnen de literatuur van de onderwijseconomie door nieuwe kennis te verschaffen over de effecten van verschillende typen beleidsinterventies. Dit betreft zowel institutioneel beleid, beleid gebaseerd op financiële prikkels, als beleid waarbij extra middelen worden ingezet voor de implementatie van nieuwe programma's gericht op specifieke doelgroepen. De eerste studie richt zich op het effect van het tijdstip van selectie op de afronding van het hoger onderwijs. De tweede studie evalueert de effectiviteit van een nieuw type financiële prikkel voor regionale onderwijsautoriteiten gericht op het terugdringen van voortijdig schoolverlaten. De derde studie analyseert de effecten van de wijkschool, een speciaal programma voor schoolverlaters met een meervoudige problematiek in Rotterdam, op een aantal sociaaleconomische uitkomsten. De vierde studie evalueert de effecten van een veelomvattend programma voor zwak presterende scholen in Amsterdam op de onderwijsprestaties van leerlingen.

Ontwikkelingen op het terrein van beleidsevaluaties

De veelheid aan factoren die bepalend zijn voor onderwijsuitkomsten maakt het niet eenvoudig om het oorzakelijk effect van een beleidsinterventie vast te stellen. Te midden van al deze, mogelijk niet observeerbare, invloeden is het immers moeilijk om het effect van het beleid te isoleren. Een betrouwbare beleidsevaluatie vereist dat het beleidseffect op een geloofwaardige manier kan worden gescheiden van de overige factoren.

Beleidsevaluaties rusten op een essentieel principe dat volgt uit het zogenoemde 'potentiële uitkomstenmodel' (Rubin, 1974, 1977). Dit model gaat er van uit dat elk individu twee potentiële uitkomsten heeft: een uitkomst in de situatie met het nieuwe beleid en een uitkomst in de situatie zonder het nieuwe beleid. Het oorzakelijk effect van het beleid is vervolgens het verschil tussen deze twee mogelijke uitkomsten. Wanneer beleidseffecten heterogeen zijn, volgt het gemiddelde beleidseffect uit het gemiddelde verschil tussen beide uitkomsten in de populatie. In de praktijk observeren we echter maar één van de twee mogelijke uitkomsten. De potentiële uitkomst zonder het nieuwe beleid is niet observeerbaar voor de groep die wordt blootgesteld aan het beleid. Voor deze uitkomst zijn we dus aangewezen op een andere groep waarbij het nieuwe beleid niet is ingevoerd. De aanname die vervolgens nodig is om het oorzakelijke effect vast te stellen, is dat de waargenomen uitkomst van deze controlegroep gelijk is aan die van de experimentgroep in de situatie zonder het nieuwe beleid. De belangrijkste uitdaging voor een beleidsevaluatie is dus om een betrouwbare controlegroep te vinden waarvoor deze aanname geldt.

Zelfselectie in een experiment- of controleconditie vormt een bedreiging voor de validiteit van deze aanname. Individuen die er voor kiezen om blootgesteld te worden aan het beleid verschillen per definitie van individuen die daar niet voor kiezen. Deze, mogelijk niet observeerbare, verschillen kunnen een vergelijking van uitkomsten tussen experiment- en controlegroep vertekenen. Wanneer men bijvoorbeeld het effect van klassengrootte op leerprestaties wil evalueren, moet men zich er van bewust zijn dat keuzes van scholen of ouders kunnen leiden tot een selectieve toewijzing van leerlingen aan grote en kleine klassen. Scholen kunnen er de voorkeur aan kunnen geven om zwakkere leerlingen in een kleinere klas te plaatsen en betrokken ouders kunnen moeite doen om hun kind in een kleine klas te plaatsen. Hierdoor kunnen verschillen in prestaties tussen leerlingen in grote en kleine klassen ook worden veroorzaakt door andere, niet geobserveerde kenmerken in plaats van door de klassengrootte.

Selectieprocessen zijn het belangrijkste probleem in beleidsevaluaties. Wanneer de beleidsvariable (die aangeeft of een individu wel of niet te maken heeft met het nieuwe beleid) is gecorreleerd met niet observeerbare factoren die de uitkomst beïnvloeden, leveren standaard regressiemethoden geen betrouwbare schatting van het effect van het beleid. In de traditionele literatuur over de productie van onderwijsuitkomsten was nog weinig aandacht voor dit probleem. Hanushek (1986) vatte de resultaten van 147 schattingen van onderwijsproductiefuncties samen. De geschatte effecten van factoren zoals klassengrootte en diverse kenmerken van leraren op onderwijsprestaties liepen sterk uiteen. De resultaten voor elk van de factoren verschilden zowel in teken als in statistische significantie. De auteur concludeerde uit deze tegenstrijdige resultaten dat er blijkbaar geen duidelijk verband is tussen onderwijsinputs en onderwijsprestaties. In geen van de studies werd echter nog rekening gehouden met selectieproblemen. Het is dus ook goed mogelijk dat een vertekening van de geschatte effecten door selectie een verklaring is voor de inconsistentie van de onderzoeksresultaten. Het selectieprobleem is pas echt onder de aandacht gekomen sinds een aantal studies op het gebied van arbeidseconomie problemen ondervonden bij het vaststellen van causale effecten van trainingsprogramma's voor werklozen (Ashenfelter, 1978; Ashenfelter en Card, 1985; Lalonde, 1986). Een invloedrijke studie van Lalonde (1986) was een belangrijke mijlpaal in dit bewustwordingsproces. De studie evalueerde de effecten van een trainingsprogramma op twee manieren: één op basis van een gerandomiseerd experiment waarbij rekening kon worden gehouden met zelfselectie en één op basis van een econometrische evaluatie waarbij geen rekening kon worden gehouden met zelfselectie. Het bleek dat de resultaten van beide evaluaties erg verschillend waren en de auteur concludeerde dan ook dat studies die niet goed rekening houden met zelfselectie vertekende resultaten kunnen opleveren. Als reactie op deze bevindingen heeft het empirisch onderzoek zich vanaf begin jaren negentig sterk gericht op geloofwaardige onderzoeksdesigns, waarmee op een goede manier rekening kan worden gehouden met het selectieprobleem. Dit heeft geresulteerd in een sterke toename van 'design-based' studies waarin expliciet aandacht wordt besteed aan de identificatie van het beleids- of behandeleffect. De verschuiving van controlestrategieën en econometrische methoden naar 'design-based' studies is een belangrijke ontwikkeling geweest in de empirische economie in de afgelopen decennia.

Deze ontwikkeling, door Angrist en Pischke (2009) betiteld als 'de geloofwaardigheidsrevolutie in de empirische economie', heeft sterk bijgedragen aan de kennis over onderzoeksmethoden die geloofwaardige oplossingen bieden voor het selectieprobleem. De meest overtuigende oplossing vanuit onderzoeksperspectief is het gecontroleerde experiment, waarbij individuen door middel van loting worden toegewezen aan een experiment- of controlegroep. De aselecte toewijzing verzekert dat de beleidsvariabele niet samenhangt met niet-geobserveerde factoren. In een dergelijk gerandomiseerd experiment is het verschil in uitkomsten tussen de experiment- en controlegroep een zuivere schatter van het gemiddelde beleidseffect. Anders gesteld levert een regressie van een uitkomstvariabele op een constante en de beleidsvariabele een zuivere schatter voor het beleidseffect. Door covariaten aan de regressiespecificatie toe te voegen kan de precisie van de schatting worden verbeterd. Het aantal gerandomiseerde experimenten in de empirische economie is de afgelopen decennia sterk toegenomen. Eén van de grootste experimenten in het onderwijs is het Tennesse STAR experiment (Krueger, 1999). Deze studie evalueert de effecten van klassengrootte door leerlingen in de onderbouw van de basisschool aselect toe te wijzen aan kleine of grote klassen. Op deze manier wordt gevonden dat een kleine klas leidt tot een bescheiden verbetering van de leerprestaties. Meer recent zijn ook sociale experimenten uitgevoerd om de effecten van het gebruik van financiële prikkels in het onderwijs te onderzoeken (Angrist en Lavy, 2009; Angrist et al., 2009; Kremer et al., 2009).

Experimenten in het onderwijs zijn lang niet altijd mogelijk. Experimenten kosten veel tijd en geld, en zijn ook niet altijd gewenst of geschikt. Daarnaast kunnen ethische overwegingen een belemmering vormen voor het starten van een experiment. Wanneer een gecontroleerd experiment niet mogelijk is, kunnen ook zogenoemde quasi-experimenten of natuurlijke experimenten een geloofwaardige oplossing bieden. Bij een natuurlijk experiment ontstaat de aselecte toewijzing aan een experiment- of controlegroep niet door de hand van de onderzoeker, maar door een toevallige situatie in de werkelijkheid. Bij natuurlijke experimenten wordt gebruik gemaakt van exogene variatie in behandeling veroorzaakt door bestaande instituties, regels, of natuurlijke krachten. We onderscheiden drie soorten natuurlijke experimenten: instrumentele variabele analyses, regressiediscontinuïteiten en difference-in-differences modellen.

Instrumentele variabele (IV) analyses maken gebruik van een situatie waarin een bepaalde variabele, het instrument, zorgt voor exogene variatie in de toewijzing aan de interventie. Dit betekent dat het instrument wel invloed heeft op de toewijzing aan de experiment- of controlegroep, maar geen invloed op de uitkomsten. Doordat enkel gebruik wordt gemaakt van de variatie in toewijzing die veroorzaakt wordt door het instrument, biedt deze methode een oplossing voor het selectieprobleem. IV modellen maken het mogelijk de effecten van beleidsinterventies consistent te schatten met behulp van ‘two-stage-least-squares’ (2SLS). Wanneer de effecten homogeen zijn, geeft dit het gemiddelde effect van de interventie. Imbens en Angrist (1994) laten zien dat een IV schatting in het geval van heterogene effecten moet worden geïnterpreteerd als een lokaal gemiddeld effect van de interventie. Zij introduceren de monotoniciteitsaanname, die inhoudt dat de toewijzing aan de interventie op monotone wijze wordt beïnvloed door het instrument. Onder deze aanname en de standaard IV aanname dat het instrument alleen invloed heeft op de uitkomst via de toewijzing aan de interventie, kunnen IV schattingen geïnterpreteerd worden als het oorzakelijk beleidseffect voor de deelgroep van individuen wier toewijzing aan de interventie wordt beïnvloed door het instrument. Voor het toepassen van IV modellen is het noodzakelijk geschikte instrumenten te vinden die aan deze voorwaarden voldoen. Een speciaal geval betreft studies waarin een gerandomiseerde toewijzing gebruikt wordt als instrument voor uiteindelijke behandeling (zie bijvoorbeeld Angrist, 1990; Bloom et al., 1997). Daadwerkelijke deelname aan de interventie kan afwijken van de toewijzing als individuen zich niet houden aan de toewijzing. In dergelijke gevallen met een binair instrument verdeelt de monotoniciteitsaanname de populatie in drie groepen: de zogenoemde ‘compliers’, ‘never-takers’ en ‘always-takers’. De ‘compliers’ zijn de individuen die hun toewijzing altijd naleven. De ‘never-takers’ en ‘always-takers’ kiezen er nooit of altijd voor om deel te nemen aan de interventie, onafhankelijk van hun toewijzing. Aangezien hun deelname aan de interventie niet varieert met het instrument, verschaffen IV schattingen geen informatie over het effect van de interventie voor deze deelgroepen. In gevallen met eenzijdige naleving kan worden aangetoond dat het lokale gemiddelde effect van de interventie gelijk is aan het gemiddelde effect voor de deelnemers aan de interventie (Bloom, 1984). Deze situatie ontstaat wanneer individuen die zijn toegewezen aan de interventie de mogelijkheid hebben om deel te nemen, terwijl het voor individuen die niet zijn toegewezen onmogelijk is om deel te nemen. De intuïtie hierachter is dat er in dergelijke gevallen geen 'always-takers' bestaan, waardoor alle deelnemers aan de interventie 'compliers' zijn. Overzichten van het gebruik van IV modellen in de economische literatuur worden gepresenteerd in Angrist en Krueger (2001) en Angrist en Pischke (2009). Een voorbeeld waarbij IV modellen vaak zijn toegepast betreft de literatuur over de rendementen op onderwijs. Het behaalde onderwijsniveau is een endogene variabele die afhangt van vele factoren, zoals niet-observeerbare vaardigheden of motivatie. Het gebruik van instrumenten voor behaald onderwijsniveau kan helpen om vertekende schattingsresultaten van het rendement op onderwijs te voorkomen. De studies in deze literatuur maken gebruik van variatie in behaald onderwijsniveau die wordt veroorzaakt door verschillen in het regionale aanbod van scholen (Card, 1995), schoolgeld (Kane and Rouse, 1995) of leerplicht (Angrist en Krueger, 1991; Acemoglu en Angrist, 2001; Oreopoulos, 2006).

Regressiediscontinuïteiten maken gebruik van een bepaalde afkapgrens die gehanteerd wordt in regelgeving of instituties. De toewijzing aan de interventie hangt dan af van een score op een bepaalde onderliggende variabele. Individuen die boven (of onder) een bepaalde grenswaarde scoren krijgen te maken met de interventie en individuen die lager (of hoger) scoren niet. Door individuen met vergelijkbare scores rondom de afkapgrens te vergelijken, kan een oorzakelijk effect van het beleid worden vastgesteld. Hierbij wordt dan in de analyses gecontroleerd voor een continue functie van de onderliggende variabele om te corrigeren voor de (kleine) verschillen tussen individuen met verschillende scores. De aanname die nodig is om een oorzakelijk effect te identificeren is dat individuen net onder de grenswaarde niet verschillen van de individuen net boven de grenswaarde, conditioneel op de continue functie van de onderliggende variabele en overige observeerbare covariaten. Deze aanname impliceert dat er geen andere verstorende discontinuïteiten rondom de grenswaarde mogen zijn. Om deze aanname zo geloofwaardig mogelijk te maken, worden regressiediscontinuïteiten vaak geschat binnen een klein interval rondom de grenswaarde. Er kan onderscheid gemaakt worden tussen zogenoemde 'sharp designs' en 'fuzzy designs' (Hahn et al., 2001). In een 'sharp design' is de toewijzing aan de interventie een deterministische functie van de onderliggende variabele. Dat wil zeggen dat alle individuen met een score boven de grenswaarde worden toegewezen, terwijl alle individuen met een score onder de grenswaarde niet worden toegewezen. In een 'fuzzy' design is de kans op toewijzing aan de interventie discontinu op de grenswaarde, maar verandert deze kans niet van 0 naar 1. Fuzzy design kunnen worden geanalyseerd met behulp van instrumentele variabele modellen. Een dummy variabele die aangeeft of de score boven de grenswaarde ligt kan daarbij gebruikt worden als instrument voor toewijzing aan de interventie.

Imbens en Lemieux (2008) en Van der Klaauw (2008) geven een overzicht van het gebruik van regressiediscontinuïteiten in de empirische economie. Op het gebied van onderwijs hebben meerdere studies de effecten van klassengrootte onderzocht door gebruik te maken van discontinuïteiten die werden veroorzaakt door regels voor maximum klassengrootte. De intuïtie is dat cohorten waarbij het totaal aantal leerlingen net groter is dan de maximum klassengrootte worden gesplitst in twee groepen, terwijl cohorten waarbij het totaal aantal leerlingen net kleiner is dan de maximum klassengrootte bij elkaar in één klas worden geplaatst. Op deze manier zorgt de regelgeving voor exogene variatie in klassengrootte. Angrist en Lavy (1999) gebruiken een dergelijke regel in Israël om de effecten van klassengrootte te onderzoeken. Zij vinden dat kleinere klassen leiden tot een beperkte verbetering van leerprestaties. Deze bevinding wordt ondersteund in andere studies die gebruik maken van soortgelijke regels (Urquiola, 2006; Browning en Heinesen, 2007). Vroegere studies naar de effecten van klassengrootte, waarin nog geen rekening werd gehouden met selectieproblemen, lieten een grote variatie aan uitkomsten zien. De recente studies op basis van (natuurlijke) experimenten laten veel minder variatie zien, waarbij de algemene conclusie lijkt te zijn dat kleinere klassen leiden tot een bescheiden verbetering van leerprestaties. Het is aannemelijk dat de consistentere onderzoeksresultaten mede voortkomen uit het gebruik van geschikte onderzoeksdesigns die op een goede manier rekening houden met selectieproblemen.

Difference-in-differences modellen maken gebruik van voor- en nametingen van individuen die te maken krijgen met nieuw beleid en individuen die niet, of in mindere mate, te maken krijgen met dat beleid. Deze modellen zijn goed toepasbaar wanneer beleid regionaal wordt ingevoerd. De ontwikkeling van uitkomsten in deze regio's (de experimentgroep) kan vervolgens vergeleken worden met de ontwikkeling van uitkomsten in regio's waar de maatregel niet is ingevoerd (de controlegroep). De identificerende aanname is de 'common trend'. Deze veronderstelt dat de ontwikkeling van uitkomsten in de controlegroep gelijk is aan de ontwikkeling in de experimentgroep als het beleid niet zou zijn ingevoerd.

Difference-in-differences modellen zijn veelvuldig toegepast in empirisch economisch onderzoek (zie bijvoorbeeld Ashenfelter en Card, 1985; Card en Krueger, 1994; Blundell et al., 1998). Voorbeelden op het gebied van onderwijs zijn Pischke (2007), die de effecten van de duur van een onderwijsjaar op leerprestaties bestudeert, en diverse studies die onderzoek doen naar de effecten van 'accountability' systemen in de Verenigde Staten (Hanushek en Raymond, 2004; Jacob, 2005; Dee en Jacob, 2011). Dit type beleid houdt in dat de prestaties van scholen worden beoordeeld volgens vooraf vastgestelde standaarden, waarbij consequenties worden verbonden aan het niet halen van bepaalde normen. De studies maken gebruik van het feit dat dergelijk beleid vaak in specifieke staten is ingevoerd en laten zien dat accountability beleid bij kan dragen aan een verbetering van onderwijsprestaties.

Resultaten

Dit proefschrift presenteert vier empirische evaluaties van verschillende typen onderwijsbeleid in Nederland. In elk van de studies ligt de nadruk op een geloofwaardige identificatie van het oorzakelijke beleidseffect, door de toepassing van 'design-based' onderzoeksmethoden. De interventies betreffen één institutionele beleidsmaatregel, één maatregel gebaseerd op financiële prikkels en twee maatregelen waarbij extra middelen worden ingezet voor de implementatie van nieuwe programma's gericht op specifieke doelgroepen.

Hoofdstuk 2 richt zich op de effecten van het tijdstip van selectie op het behalen van een diploma in het hoger onderwijs. Dit hoofdstuk past in de literatuur over institutionele kenmerken van onderwijsstelsels. Het tijdstip waarop leerlingen worden geselecteerd in verschillende onderwijsniveaus is één van de meest opvallende verschillen tussen landen. Eerdere studies naar de effecten van selectie hebben gebruik gemaakt van verschillen tussen landen (Schuetz et al., 2005; Hanushek en Woessmann, 2006). Het Nederlandse onderwijssysteem biedt de mogelijkheid om het effect van het tijdstip van selectie te onderzoeken binnen één land. Sommige scholen in Nederland bieden categoriale klassen aan, waarbij leerlingen na het verlaten van het primair onderwijs op 12-jarige leeftijd direct worden geselecteerd in een bepaald onderwijsniveau. Andere scholen bieden gecombineerde klassen aan waarin leerlingen bij elkaar geplaatst worden. Deze gecombineerde klassen stellen het tijdstip van selectie met één of twee jaar uit. Dit veroorzaakt variatie in het tijdstip van selectie die we benutten door het afronden van het hoger onderwijs onder vroeg geselecteerde leerlingen te vergelijken met het afronden van het hoger onderwijs onder later geselecteerde leerlingen. In deze studie maken we gebruik van VOCL (Voortgezet Onderwijs Cohort Leerlingen) data uit 1989. Het belangrijkste probleem voor het vaststellen van een oorzakelijk effect is zelfselectie in de verschillende type scholen. Hier wordt op een aantal manieren mee omgegaan. We beperken allereerst de onderzoeksgroep tot leerlingen die het primair onderwijs hebben verlaten met een mavo-advies. Dit is een homogene groep leerlingen wat betreft vaardigheden, die naar verwachting het sterkst zal worden beïnvloed door het tijdstip van selectie. Daarnaast maken we gebruik van een instrumentele variabele aanpak. Hierbij wordt het regionale aanbod van typen scholen gebruikt als instrument voor vroege selectie. In alle analyses wordt gecontroleerd voor een grote verzameling achtergrondkenmerken. De schattingsresultaten laten zien dat vroege selectie een negatief effect heeft op het afronden van het hoger onderwijs. OLS schattingen suggereren dat vroege selectie de kans op het behalen van een diploma in het hoger onderwijs met ongeveer 5 procentpunt verkleint. De IV schattingen laten nog grotere negatieve effecten van vroege selectie zien. Additionele analyses wijzen uit dat leerlingen met een havo-advies niet negatief worden beïnvloed wanneer zij samen met leerlingen met een mavo-advies in een gecombineerde klas geplaatst worden. De onderzoeksresultaten suggereren dat het aantal hoog opgeleiden kan worden vergroot door uitstel van het tijdstip waarop leerlingen worden geselecteerd in verschillende onderwijsniveaus.

Hoofdstuk 3 is gerelateerd aan de literatuur over het gebruik van financiële prikkels binnen het onderwijs. Een groot aantal empirische studies vindt geen relatie tussen extra middelen voor het onderwijs en een verbeterde onderwijskwaliteit (Woessmann, 2003; Hanushek, 2006). De conclusie dat additionele middelen geen garantie betekenen voor betere onderwijsprestaties, heeft geleid tot een sterk toegenomen belangstelling voor de mogelijkheden van financiële prikkels in het onderwijs. Wereldwijd zijn verschillende vormen van beleid gebaseerd op prikkels ingevoerd. Empirische studies laten zien dat dergelijke beleidsvormen kunnen bijdragen aan een verbetering van onderwijsprestaties (Angrist en Lavy, 2009; Kremer et al., 2009; Lavy, 2002, 2009). Tegelijkertijd keren leraren en hun vakbonden zich vaak sterk tegen het gebruik van financiële prikkels en laten diverse studies zien dat strategisch gedrag een bedreiging vormt voor de effectiviteit van dit type beleid (Glewwe et al., 2003; Jacob, 2005; Figlio en Getzler, 2006). Eerdere studies hebben zich voornamelijk gericht op financiële prikkels voor leerlingen, leraren en scholen. Deze studie beoogt een bijdrage te leveren aan de literatuur door zich te richten op een nieuw type financiële prikkel gericht op regionale onderwijsautoriteiten. Nederland kent 39 RMC-regio's (Regionale Meld en Coördinatiefunctie voor voortijdig schoolverlaten) die verantwoordelijk zijn voor de regionale registratie en bestrijding van voortijdig schoolverlaten. Met 14 van deze RMC-regio's zijn in 2006 convenanten afgesloten die een financiële prikkel bevatten om het aantal voortijdig schoolverlaters in de regio terug te dringen. Deze RMC's konden extra middelen ontvangen wanneer zij er in slaagden het aantal voortijdig schoolverlaters te verminderen. De 14 RMC's werden geselecteerd op basis van het aantal voortijdig schoolverlaters voorafgaand aan de invoering van het beleid. De introductie van het beleid in een deel van de RMC-regio's en het gebruik van de specifieke selectieregel maakt het mogelijk om lokale difference-in-differences modellen te schatten. Deze aanpak combineert een difference-in-differences model met een regressiediscontinuïteit. Hierbij wordt een difference-in-differences model geschat rondom de grenswaarde voor selectie voor het convenant. In het onderzoek wordt gebruik gemaakt van administratieve data over voortijdig schoolverlaten in het jaar voor en het jaar na de introductie van het beleid voor alle leerlingen in Nederland. We vinden statistisch insignificante effecten van het beleid op voortijdig schoolverlaten. Deze resultaten zijn robuust voor een groot aantal modelspecificaties en gevoeligheidsanalyses. Dit suggereert dat de financiële prikkel voor regionale onderwijsautoriteiten niet effectief is geweest in het terugdringen van het aantal voortijdig schoolverlaters. Additionele analyses lijken er op te wijzen dat de financiële prikkels hebben geleid tot een strategische reactie van de onderwijsautoriteiten.

De volgende twee hoofdstukken zijn gewijd aan de effecten van intensieve programma's voor specifieke doelgroepen. Hoofdstuk 4 evalueert de effecten van de wijkscholen in Rotterdam op een aantal sociaaleconomische uitkomsten door middel van een veldexperiment. De wijkscholen bieden een intensief programma voor schoolverlaters met complexe problemen op meerdere gebieden (zoals arbeid, financiën, gezondheid, huisvesting en sociale omgeving). Het programma heeft als doel hen te begeleiden naar het reguliere onderwijs of naar werk. Voortijdig schoolverlaters hebben een grotere kans op latere werkloosheid of betrokkenheid bij criminele activiteiten (zie bijvoorbeeld Lochner en Moretti, 2004; Machin et al., 2011). De hoge maatschappelijke kosten die daar mee gepaard gaan, legitimeren publiek gefinancierde interventies gericht op het verbeteren van de vooruitzichten voor deze jongeren. Eerdere studies hebben laten zien dat het niet eenvoudig is om effectieve programma's te ontwikkelen voor deze doelgroep (LaLonde, 2003; Carneiro en Heckman, 2003). De wijkschool is grotendeels vormgegeven in lijn met de meest veelbelovende interventies uit de literatuur. De wijkscholen bieden een veelomvattend programma dat bestaat uit een combinatie van onderwijs, werk en zorgdiensten, en persoonlijke begeleiding door coaches. Aangezien een aselecte toewijzing van doelgroepjongeren aan de wijkschool en een controleconditie niet mogelijk bleek, is een gecontroleerd natuurlijk experiment opgezet waarbij de toewijzing van jongeren gebaseerd werd op specifieke tijdsperiodes. Deze tijdsperiodes zijn vooraf vastgesteld op basis van capaciteitsrestricties op de wijkscholen. Gedurende bepaalde periodes zijn jongeren uit de doelgroep verwezen naar de wijkschool, terwijl vergelijkbare jongeren uit de doelgroep in andere periodes zijn verwezen naar een controleconditie. Deze laatste groep jongeren krijgt de standaardbehandeling die bestaat uit een verwijzing naar een regulier re-integratietraject. De toewijzing aan de wijkschool of controlegroep hangt op deze manier enkel af van het moment van aanmelding. Deze aanpak illustreert een mogelijke opzet van een veldexperiment in situaties waarin een gerandomiseerd experiment niet mogelijk is. De effecten van het programma worden vervolgens vastgesteld door het vergelijken van uitkomsten van jongeren die verwezen zijn naar de wijkschool met de uitkomsten van jongeren die verwezen zijn naar een regulier traject, conditioneel op het moment van aanmelding. We gebruiken een instrumentele variabele aanpak om te corrigeren voor het feit dat in de praktijk niet alle jongeren conform de verwijzingsregel zijn toegewezen. Hierbij wordt de daadwerkelijke verwijzing geïnstrumenteerd met de verwijzing zoals die had moeten plaatsvinden op basis van de regel. In het onderzoek wordt gebruik gemaakt van administratieve data over onderwijspositie, arbeidsmarktpositie en crimineel gedrag drie jaar na de start van de interventie. De data over criminaliteit zijn afkomstig van de politie Rotterdam-Rijnmond. We vinden insignificante effecten van het programma op de onderwijspositie of het hebben van een baan. Deze bevinding past in een groot deel van de literatuur waar geen effecten worden gevonden van trainingsprogramma's voor probleemjongeren op arbeidsmarktuitkomsten. Het belangrijkste resultaat is dat een verwijzing naar de wijkschool leidt tot een toename van criminele activiteit. Dit effect wordt veroorzaakt door een toename van criminele activiteit onder de groep jongeren die al verdacht werd van een misdrijf voor aanvang van het programma. Dit resultaat is consistent met eerdere studies die hebben aangetoond dat een groepsgewijze aanpak van probleemjongeren kan leiden tot een toename van criminaliteit (Dishion et al., 2007). Peer effecten vormen een mogelijke verklaring voor deze bevinding.

Hoofdstuk 5 evalueert de effecten van een beleidsprogramma van de gemeente Amsterdam gericht op het verbeteren de onderwijskwaliteit op zwak presterende basisscholen. Dit programma is gebaseerd op de zogenoemde 'comprehensive school reforms', die met name in de Verenigde Staten veel zijn gebruikt voor het verbeteren van zwakke scholen. Een dergelijke aanpak richt zich op de gehele school en bestaat uit een geïntegreerde verzameling maatregelen die op verschillende lagen binnen de school wordt ingevoerd. Empirisch bewijs voor de effectiviteit van dit type programma's is echter nog beperkt en eerder onderzoek is vaak niet gebaseerd op geloofwaardige onderzoeksdesigns die een oplossing bieden voor het selectieprobleem (Borman et al., 2003). Deze studie maakt gebruik van een 'difference-in-differences' aanpak om de effecten van de verbeteraanpak in Amsterdam op de onderwijsprestaties van leerlingen te schatten. De verbeteraanpak implementeert een systematische en opbrengstgerichte manier van werken binnen de school. Het programma kent een integrale aanpak bestaande uit diverse elementen, zoals evaluaties van de kwaliteit van leraren door middel van lesobservaties, cursussen voor leraren, schoolleiders en ander schoolpersoneel en het gebruik van nieuwe instructiemethoden. Vanaf 2008 konden alle basisscholen in Amsterdam die door de Inspectie van het Onderwijs als 'zwak' of 'zeer zwak' waren beoordeeld op vrijwillige basis deelnemen aan de aanpak. In het onderzoek wordt gebruik gemaakt van CITO toetsscores in groep 8 voor de jaren 2005 tot en met 2012. Dit maakt het mogelijk om de ontwikkeling van CITO toetsscores op zwak presterende scholen in Amsterdam te vergelijken met de ontwikkeling van CITO toetsscores op zwak presterende scholen buiten Amsterdam in een difference-in-differences model. De schattingsresultaten wijzen op een negatief effect op de toetsscores in de eerste vier jaar na de invoering van het beleid. De introductie van het beleid heeft de toetsscores met 0.17 standaarddeviaties verlaagd. Interviews met schooldirecteuren van deelnemende scholen verschaffen een mogelijke verklaring voor deze bevinding. Uit deze interviews komt naar voren dat de veeleisende aanpak van het programma heeft geleid tot een toename in de (uitgaande) mobiliteit van leraren. Het is denkbaar dat het verlies aan schoolspecifieke kennis, de benodigde aandacht voor het vinden van geschikte vervangers en het onzekere werkklimaat hebben bijgedragen aan de verslechtering van leerprestaties. Mogelijk reflecteren de bevindingen aanpassingskosten en kan het beleid op langere termijn betere effecten sorteren. De studie concludeert dat het beleid in ieder geval grote kosten met zich meebrengt in termen van substantieel lagere toetsscores voor ten minste vier cohorten leerlingen.

Concluderende opmerkingen

De resultaten in dit proefschrift verschaffen inzicht in de effecten van verschillende soorten beleidsmaatregelen. De bevindingen maken duidelijk dat beleid niet altijd uitpakt zoals vooraf beoogd. Twee intensieve programma's voor probleemjongeren en zwak presterende scholen blijken niet succesvol in het verbeteren van uitkomsten. We vinden zelfs negatieve effecten van beide programma's. Dergelijke beleidsmaatregelen kosten veel geld, tijd, en energie. De resultaten passen in het beeld uit eerdere studies dat extra middelen geen garantie zijn voor betere uitkomsten. Dit beeld verklaart de toegenomen belangstelling voor andere, meer kosteneffectieve manieren om het onderwijssysteem te verbeteren. Beleid gebaseerd op financiële prikkels is in het algemeen minder duur dan beleid waarbij extra middelen worden vrijgemaakt voor de implementatie van nieuwe programma's. Dergelijk beleid behelst zowel financiële prikkels voor studenten of leraren als zogenoemde 'accountability' systemen, waarbij scholen worden beoordeeld volgens vooraf vastgestelde normen. Dit proefschrift levert een bijdrage aan deze literatuur door onderzoek te doen naar een nieuw type beleid met financiële prikkels voor regionale onderwijsautoriteiten. Dit onderzoek levert geen bewijs dat dit beleid bijdraagt aan betere onderwijsprestaties. Naast beleid gebaseerd op financiële prikkels, heeft ook beleid dat zich richt op institutionele kenmerken van het onderwijssysteem de potentie om uitkomsten te verbeteren tegen geringe kosten. Zo laat dit proefschrift zien dat het aantal hoog opgeleiden kan worden vergroot door de leeftijd waarop leerlingen worden geselecteerd in verschillende onderwijsniveaus te verhogen van 12 naar 13 of 14. Een dergelijke institutionele wijziging zou nauwelijks extra kosten met zich meebrengen.

Vanzelfsprekend impliceren de resultaten in dit proefschrift niet dat beleid waarbij extra middelen worden ingezet voor de implementatie van specifieke programma's nooit succesvol kan zijn. Zij illustreren wel dat dure interventies niet noodzakelijkerwijs effectiever zijn dan interventies die minder middelen vragen. Dit suggereert dat het verhogen van budgetten of toekennen van extra middelen niet altijd de beste aanpak is om problemen op te lossen of om het systeem te verbeteren. De bevindingen dat beleidsinstrumenten geen, of zelfs negatieve, effecten kunnen hebben, geven het belang aan van een goede kennis van de werking van beleid. Nu de empirische evaluatietechnieken een zodanige ontwikkeling hebben doorgemaakt dat zij betrouwbare oplossingen kunnen bieden voor selectieproblemen, kunnen beleidsmakers de vruchten plukken van inzichten uit nieuwe evaluatiestudies. Deze inzichten kunnen gebruikt worden voor de aanpassing van beleidsinstrumenten, het opschalen van effectief beleid, of het stopzetten van maatregelen die niet werkzaam blijken. Op deze manier kunnen nieuwe beleidsevaluaties bijdragen aan een efficiëntere allocatie van middelen en een verbetering van het onderwijssysteem.

Curriculum Vitae

Roel van Elk was born in Nijmegen on 7 June 1979. He studied at Tilburg University and obtained a Bachelor of Science in Econometrics and Operations Research and a Master of Science in Mathematical Economics and Econometric Methods (cum laude) in 2005. From 2005 onwards, he has been employed at CPB Netherlands Bureau for Economic Policy Analysis. He has been involved in various research projects in the fields of transport economics, health economics and competition and regulation. Since 2007 his research focus has been on the economics of education, with a special interest in policy evaluations. He started his dissertation research at Erasmus University Rotterdam in 2010. His papers have been published in *Economics of Education Review* and *De Economist*. As of February 2014, he works as unit leader research and science policy at CPB.