Essays on Teacher Quality and Coaching

Marc van der Steeg

Essays on Teacher Quality and Coaching

Essays over de kwaliteit van leraren en coaching

Marc van der Steeg

ISBN/EAN: 978-94-6233-428-1

Printed by Gildeprint, Enschede

Essays on Teacher Quality and Coaching

Essays over de kwaliteit van leraren en coaching

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 van Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 24 november 2016 om 15.30 uur

Marc Willem van der Steeg

geboren te Leiderdorp

Erasmus University Rotterdam

Promotiecommissie

Promotor

Prof.dr. H.D. Webbink

Overige leden

Prof.dr. F. Corvers

Dr. O. Marie

Prof.dr. H. Oosterbeek

Copromotor

Dr. A.C. Gielen

Preface

My interest in doing research started when I was writing my Master thesis at the Institute of Economic Growth in India on India's vulnerability to macroeconomic shocks. This Master thesis turned out to be a somewhat larger and tougher project than expected. It took me almost two years to complete this thesis. My supervisor told me that if I would like to do a PhD, I would be already half underway with the work I had done for my Master Thesis. I thought about this for some time, but the idea of working on the same topic for a couple more years was not that appealing.

Instead, I got the opportunity to start my career at CPB Netherlands Bureau for Economic Policy Analysis. I started there in 2004 as a junior researcher at the Knowledge, Growth and Structure department. After some time I became involved in a project on early schoolleaving with Dinand Webbink, who was head of the Education Program at CPB at that time. This project and the interesting work and ideas of Dinand in the field of economics of education really triggered my interest in doing more research on education and education policy. Within the education program we started with a small and enthusiastic group of people to work on a number of interesting research projects using quasi-experimental and experimental research designs. First with Roel van Elk, and later on Sander Gerritsen joined the education program as well. At that time Dinand started to stimulate all three of us to think about starting a PhD track. It took me some time before I became convinced that this would be a good plan for me, since I was having some other interests and responsibilities that consumed quite a share of my time such as taking care of our young daughter and being the treasurer of my soccer club.

Nevertheless, I decided to really go for it three years ago, when I had written two CPB discussion papers that were considered suitable for a PhD thesis. I became an external PhD student under the supervision of Dinand at the Erasmus University Rotterdam. I completed the third paper of my PhD thesis in the beginning of 2015. That was just after I had decided to leave CPB and to switch to the Ministry of Education. I wrote my fourth paper when I was working at the Ministry. The fruitful cooperation with my former CPB colleagues Sander Gerritsen and Sonny Kuijpers in this research project helped me to finish my fourth paper by the end of 2015.

This thesis could not have been written without the help and support of many people. First of all I would like to thank Dinand for his inspiring guidance. He always had creative ideas concerning methodology and robustness checks and always gave constructive feedback on early versions of papers. Moreover, it was special to have a supervisor with whom I have worked pleasantly for several years at CPB and who has a great sense of humour. Second, I am indebted to my former employer CPB for enabling me to combine my work at CPB with writing a PhD thesis. I would like to thank Coen Teulings, George Gelauff, and Ruud Okker for giving me the opportunity to start my PhD thesis and combine this with my work at CPB. In later stages Bas ter Weel and Daniel van Vuuren gave me the opportunity to continue working on my PhD. I am very grateful to Bas ter Weel for his valuable comments on particularly the first two papers of my thesis. Furthermore, I would like to thank program leaders Debby Lanser and later on Karen van der Wiel for their flexibility to create space for my research ideas and projects into the work program of the education research program at CPB. My co-authors also deserve special mention. It was a great pleasure working with you guys, Roel, Sander and Sonny. Special thanks to Roel and Ib for being my paranimf. I would like to thank Rob Luginbuhl and Adam Elbourne for checking my English and my other former CPB colleagues for the great and unique work atmosphere at CPB. Furthermore, I would like to thank Vera Pieterman and Annemarie Sipkes for providing me some time to complete my PhD thesis. Furthermore, I am grateful to the members of the PhD comittee, Frank Corvers, Olivier Marie and Hessel Oosterbeek for their time and effort, and my copromotor Anne Chielen for valuable comments and suggestions on a draft of my PhD thesis.

Last but not least I would like to thank my family and friends who indirectly contributed to this thesis. In particular I would like to thank my father and mother who have always been supportive during my educational career. In particular during the time I was writing my Master thesis and I was having quite a hard time solving some modelling problems and completing this thesis. Finally I would like to express my gratitude and love for my wife Esther. The moral support she gave me during the years I was working on my PhD thesis was great. She also gave me the opportunity to work on my thesis during nights and some weekends. I am very proud of our kids Floor and Ruben who bring so much joy into our lives every day. I am looking very much forward to spend more time with my wife and kids now I completed my PhD!

Marc van der Steeg

Berkel en Rodenrijs, September 2016

Contents

1.

Introduction

1.1. Motivation and contribution

Teacher quality is key to the performance of pupils in education. Children assigned to a teacher with a one standard deviation higher quality score of 0.1 to 0.3 standard deviations higher on cognitive tests (see e.g. Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Kane & Staiger, 2008; Hanushek & Rivkin, 2010). Moreover, it was recently found that higher teacher quality may positively affect future outcomes, such as the probability of college attendance, the probability of working and annual earnings (Chetty et al., 2014a). Hanushek (2011) estimates that a teacher one standard deviation above the mean effectiveness annually generates marginal gains of over $400,000 in present value of student future earnings.

Improvements in teacher quality can therefore generate large returns. It is less clear however what drives teacher quality and how the quality of teachers can be improved. There is evidence that teacher quality rises with experience, whereas the evidence for other observable characteristics such as educational attainment and cognitive skills of teachers is at best mixed.[1] This leaves most of the variation in teacher quality unexplained.

This thesis aims to provide more insight into the determinants of teacher quality. The thesis consists of four essays that focus on the determinants of teacher quality and the effectiveness of policies targeted at raising teacher quality. In three papers I exploit experiments and quasi-experiments for identifying the causal effect of specific policies. Applying these experimental approaches is likely to be important and might yield new insights as most of the previous studies on teacher quality focuses on associations. The findings from these earlier studies might be biased by unobserved factors correlated with the determinants of teacher quality. I study three different potential determinants: i) teacher classroom practices, ii) continuous schooling by teachers, and iii) teacher pay. The fourth paper in this thesis deals with the impact of coaching, which can be considered a particular aspect of teaching.

Teacher classroom practices

This thesis starts with the question whether and to what extent a detailed observation rubric measuring teachers' pedagogical, didactical and classroom management skills is able to predict pupils' cognitive achievement.

[1] See Hanushek & Rivkin (2006) and Harris & Sass (2011) for literature reviews on the relationship between observable teacher characteristics and teacher quality.

The small share of variation in teacher quality explained by observable teacher characteristics has triggered education researchers in the last decade to focus more on whether observable teaching classroom practices and skills matter for teacher quality and hence pupil performance. One of the largest projects in this spirit is the Measures of Effective Teaching (or MET) project that has been carried out in the US in recent years.[2] I add to this small but growing literature by showing that the score on a detailed observation rubric is predictive of pupil achievement gains and that the rubric is particularly helpful in identifying the weaker teachers in terms of their impact on pupil achievement. This is an important finding, since this suggests that these measures have potential to be used for targeted feedback on and coaching of (weaker) teachers.

Teacher schooling

In the second paper, I investigate the impact of a public teacher schooling voucher scheme on enrollment in and completion of degree programs by teachers. The schooling vouchers consist of compensation for tuition fees of bachelor or master degree programs as well as compensation for their employers (i.e. the schools) for the costs of arranging a substitute teacher for the days at which the teacher is on study leave.

Investment in schooling of teachers is a regularly used policy measure, but studies on its impact are scarce. It is not obvious that extra funding for teacher schooling will significantly increase participation in these activities, since schools often already have regular budgets for schooling and training of their teachers and teachers may as well invest in schooling out of their own pocket. A large deadweight loss has been found in the case of schooling vouchers for employees in other sectors than the education sector (e.g. Schwerdt et al., 2012; Hidalgo et al., 2014). This paper contributes to the small literature on effects of schooling vouchers on schooling participation. In addition, I do not only look at effects on enrollment, but on completion as well. This is important since the vouchers can be used for participation in degree programs which may take one or more years of study. Important differences with earlier voucher effectiveness studies is that I study a voucher scheme with a much higher face value and that the target group is limited to high educated professionals.

[2] See http://www.metproject.org/index.php for more information on this project and the research findings.

Teacher pay

In the third paper I investigate the effect of higher teacher pay on teacher retention. The intervention consists of more funds for placing a higher share of teachers in a higher salary scale in a targeted urbanized teacher shortage region.

Higher teacher pay is often introduced with the aim to attract and retain quantitatively as well as qualitatively sufficient teachers in the teaching profession. Substantial public resources are often involved in raising teacher pay, which makes it important to know how the benefits relate to the costs. As far as I am aware only a few studies have succeeded in reliably investigating the effects of higher teacher pay on teacher retention. Notable exceptions are studies by Clotfelter et al. (2008) and Hendricks (2014) that both find positive effects on teacher retention. I add to this literature by focusing on effects of higher teacher pay on retention in the teaching profession as well as on retention in a targeted shortage region. Another contribution to the literature is that I also study the impact of teacher pay on participation in formal schooling activities.

Coaching

The fourth and last essay of this thesis investigates the impact of coaching on student dropout. The intervention consists of the assignment of a full-time coach to a class of students in post-secondary vocational education for one or two years.

Apart from the well-established impact of teachers on student performance there is small but growing evidence that coaching or mentoring can have an important impact on school success as well. See for instance Lavecchia et al. (2014) for a review of this literature. I contribute to this literature by evaluation of a randomized experiment with a relatively high-intensive coaching program.[3]

1.2. Empirical methods

The main aim of this thesis is to establish the determinants of teacher quality. Many of the studies in this area lack a credible identification strategy. Determining what improves teacher quality is difficult for various reasons.

[3] The findings of this experiment have been published in the Economics of Education Review: Steeg, M. van der, R. van Elk, and D. Webbink, 2015, Does intensive coaching reduce school dropout? Evidence from a randomized experiment, *Economics of Education Review*, vol. 48, October 2015, pp. 184-197.

One reason is that both teachers and pupils are often non-randomly assigned to classes. This implies that simply relating teacher characteristics to pupil performance is not likely to produce reliable estimates. For instance, more experienced teachers could be purposely assigned to classes consisting of weaker or more problematic pupils, thereby leading to an underestimation of the effect of teacher experience when the researcher fails to control for these often unobserved difference in classroom characteristics.

A problem with the evaluation of the effects of public policies on teacher quality is that these interventions are usually introduced simultaneously for all teachers, e.g. a pay raise for teachers or the provision of larger school budgets for the schooling of teachers. The evaluation of such measures are then likely to produce biased results as many other factors (e.g. other policies) may operate at the same time and affect the outcome as well. There are also cases where policies affect only a subsample of teachers. Simply comparing the outcome of treated with untreated teachers is also likely to produce biased results due to the (self-) selection of teachers or schools for these treatments. If for instance more motivated teachers apply for schooling vouchers, simply comparing their outcome with those teachers who did not apply is likely to produce biased results (i.e. an overestimation).

A solution to circumvent these difficulties is to search for (plausibly) exogenous variation in the determinant of interest. This variation can arise in various ways. In this thesis I do this by exploiting a randomized experiment with an intensive coaching intervention (chapter 5), exploiting regional variation in teacher pay (chapter 3), exploiting oversubscription in a teacher schooling voucher scheme and exploiting variation in the timing of applications (chapter 4), and by using a control strategy (chapter 2). I will now briefly discuss the empirical strategies used in the subsequent chapters of this thesis.

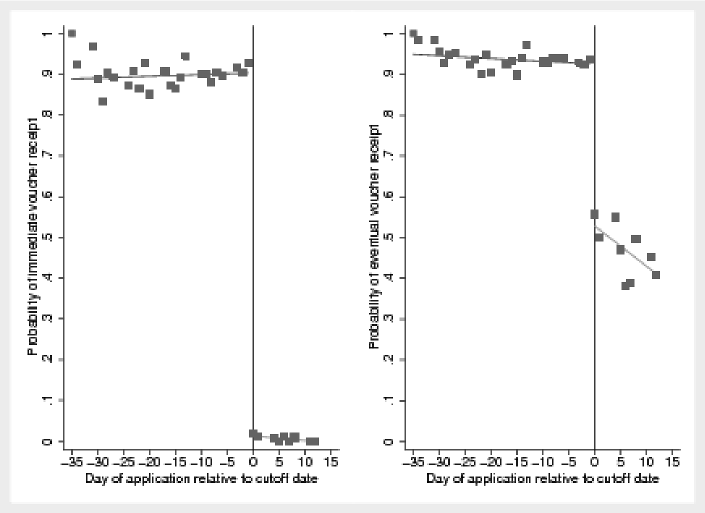
Teacher classroom practices and teacher value added: control strategy

In chapter 2 I use a teacher value-added type of model to estimate the relationship between teacher evaluation scores and their pupils' achievement gains. This type of model has been used in other recent studies, e.g. Kane et al. (2011) and Rockoff & Speroni (2011). The model we use aims to account for possible non-random assignment of teachers and pupils to classes by controlling for a large number of relevant pupil characteristics including socio-economic background and, previous test scores. Although unobservable characteristics could confound the estimates, several studies show that experimental estimates - in situations where teachers have been randomly assigned to classrooms - are consistent with value-added estimates that result from non-experimental value-added models, as long as these latter models control for students' prior test scores (see Nye et al., 2004; Kane & Staiger, 2008; and Kane et al., 2013; Chetty et al., 2014b).

Schooling vouchers and schooling participation: oversubscription

In chapter 3 I exploit two design features in a voucher program to investigate the effects of receiving a schooling voucher on schooling participation and completion. The voucher program had a yearly budget ceiling with a first-come-first-serve allocation mechanism when the yearly budget was exceeded. The program therefore involved a cutoff date after which applicants would not be assigned a teacher voucher, in the event that the budget ceiling was exceeded. This is clearly shown in the left-hand panel of figure 1.1, which depicts the probability of receiving a voucher in the first application period against the day of application.

Figure 3.1 Relationship between day of application and probability of immediate (left panel) and eventual (right panel) voucher assignment

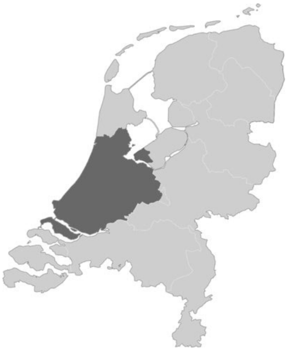


One could argue that these teachers, particularly the ones near the threshold date, were equally motivated and able to participate in schooling. The basic idea behind the empirical strategy is to compare the outcomes of teachers who applied in the vicinity of the threshold. In practice a substantial number of teachers who were too late in applying in the first round, managed to receive a voucher in later application periods. This can be seen in the righthand panel of figure 1.1 which depicts the probability of ever having received a schooling voucher as a function of the date of application in the first application period. I therefore exploit the variation in ever having received a voucher that is caused by the threshold date in the first application period by using this threshold date as an instrument for receiving treatment. This empirical evaluation approach is called a *fuzzy regression discontinuity* design (see Hahn et al., 2001; Angrist & Pischke, 2009).

Exploiting regional variation in teacher pay

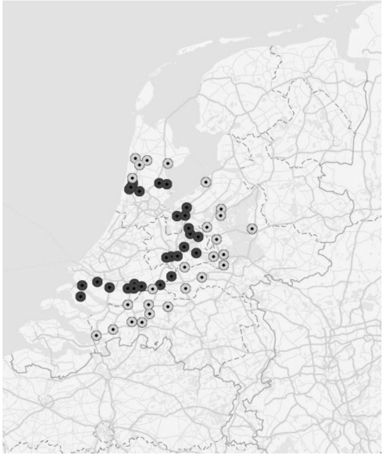
In order to estimate the effects of a higher teacher pay on teacher retention, which is the purpose of chapter 4, I exploit variation in teacher pay that was caused by the introduction of a regional teacher pay policy. Figure 1.2 below shows the region (marked dark-grey) where the extra funding for higher teacher pay was made available.

Figure 4.1 Randstad region within the Netherlands



To investigate effects of higher teacher pay on teacher retention I compare the evolution of teacher retention in treated and untreated regions. More specifically I compare the retention of teachers in the treatment and control schools just along the border of the policy intervention. The selection of municipalities for this local estimation sample is shown in figure 1.3. The idea is that time-varying (unobserved) teacher, school and pupil characteristics, as well as external labor market conditions should be similar and therefore these aspects should not bias the estimates. This so-called local difference-in-differences approach does not require that pre-treatment retention in treated and untreated regions be exactly the same. The main assumption made when employing a diff-in-diff approach is that the evolution of teacher retention in both regions would have been the same in the absence of the regional teacher pay policy. This is called the common-trend assumption.

Figure 1.3 Local estimation sample of border municipalities



Coaching and school dropouts: exploiting a randomized experiment

Chapter 5 investigates the effects of a randomized coaching intervention on the number of school dropouts. The coaching program was allocated to randomly selected classes of first-year students that have in turn been composed randomly. The control group constituted of classes of students that received care as usual. Because of the randomized assignment of the coaching treatment evaluation of its impact essentially comes down to simply comparing the outcomes of treated and untreated students. However, controlling for student characteristics helps to improve precision of the effect estimates. The randomized treatment assignment assures that treatment status is not correlated with omitted variables that could lead to biased estimates. Without an experimental design it could for instance have been that at-risk students with a larger probability of dropping out would have a larger probability of receiving the coaching intervention. This would lead to underestimating the impact of the coaching interventions. It could however also have been the case that the most problematic students are less likely to receive coaching (e.g. because of a higher probability of absence or less openness to coaching) and that coaching is provided to a less problematic group. In that case we would obtain an overestimate of the impact of coaching when simply comparing coached and non-coached students in the absence of an experimental design.

1.3. Summary of main findings

Chapter 2 examines the relationship between teacher evaluations and pupil performance gains in primary education. Teacher evaluations have been conducted by trained external evaluators who scored teachers on a detailed rubric containing 75 classroom practices. These practices reflect pedagogical, didactical and classroom organization competences considered crucial for effective teaching. Conditional on previous year test scores and several pupil and classroom characteristics the score on this rubric significantly predicts pupil performance gains on standardized tests in math, reading and spelling. Estimated test score gains are on the order of 0.4 standard deviations in math and spelling and 0.25 standard deviations in reading if a pupil is assigned a teacher from the top quartile instead of the bottom quartile of the distribution of the evaluation rubric. The observation rubric particularly seems to have potential to identify the weaker teachers.

Chapter 3 investigates the effects of schooling vouchers for teachers on their enrollment and completion of higher education programs, as well as on their retention. This is done by employing a so-called fuzzy regression discontinuity design. The discontinuity in the probability of being assigned a voucher arises due to budget constraints in the first application period. The estimates suggest that effects of voucher assignment on both higher education enrollment and completion rates are in the order of 10 to 20 percentage points as measured five and a half years after the application date for the voucher. Relative to a baseline enrollment rate of 77 percent and a baseline completion rate of 54 percent (i.e. of applicants that were not assigned a voucher), these estimates correspond to a 12 to 29 percent higher enrollment and to a 17 to 42 percent higher completion. The effects on enrollment and completion are relatively small for shorter studies (up to one year) and for teachers that had already started at the time of application. The teacher voucher largely crowds out both funding by schools out of their regular professional development budgets as well as financial contributions by teachers themselves. Our results suggest small positive effects of voucher assignment on retention in education as measured four years after application.

Chapter 4 investigates the effects of higher teacher pay for secondary school teachers on their teacher retention decision and enrollment in additional schooling. I exploit regional variation in teacher pay that is induced by the introduction of a new teacher remuneration policy. This policy provided schools in an urbanized region with extra funds to place a significantly larger share of their teachers in a higher salary scale. I exploit this policy in an instrumental variable setup to estimate the effects of higher teacher pay on our outcomes. I find no effect of higher teacher pay on the probability of remaining in the teaching profession. The policy however succeeded in keeping a slightly larger share of teachers in the targeted region. In addition, the findings suggest that the policy increased teachers' yearly probability of enrollment in bachelor or master degree programs from 2.3% to 3.2%. This finding is consistent with the setup of the policy in which one of the criteria for placement in a higher salary scale is that teachers would obtain extra qualifications or gain extra expertise.

Chapter 5 investigates the effect of an intensive coaching program aimed at reducing school dropout rates among students aged 16 to 20 in post-secondary vocational education. Within the coaching program students were offered fulltime support and guidance with their study activities, personal problems and internships in firms. The coaching program lasted one or two years. Students were randomly assigned to classes and the coaching program was randomly assigned to classes as well. I find that one year of coaching reduced school dropout rates by more than 40 percent from 17 to 10 percentage points. The second year of coaching further reduced school dropout by one percentage point. The program is most effective for students with a high ex-ante probability of dropping out, such as students no longer obliged to be in formal education, male students, and students not living with both parents. Cost-benefit analysis suggests that one year of coaching is likely to yield a net social gain.

2.

Teacher evaluations and pupil achievement gains: Evidence from classroom observations[4]

Abstract

This chapter investigates the relationship between teacher evaluations and pupil performance gains in primary education. Teacher evaluations have been conducted by trained external evaluators who scored teachers on a detailed rubric containing 75 classroom practices. These practices reflect pedagogical, didactical and classroom organization competences considered crucial for effective teaching. Conditional on previous year test scores and several pupil and classroom characteristics the score on this rubric significantly predicts pupil performance gains on standardized tests in math, reading and spelling. Estimated test score gains are in the order of 0.4 standard deviations in math and spelling and 0.25 standard deviations in reading if a pupil is assigned a teacher from the top quartile instead of the bottom quartile of the distribution of the evaluation rubric. The observation rubric particularly seems to have potential to identify weak teachers. These observations may stimulate targeted teacher improvement plans and personnel decisions.

[4] This is joint work with Sander Gerritsen: Steeg, M. van der, and S. Gerritsen, 2016, Teacher evaluations and pupil achievement gains: Evidence from classroom observations, *De Economist*, doi:10.1007/s10645-016-9280-5 , July 2016, pp. 1-25.

2.1 Introduction

Research on the impact of teacher quality on student achievement consistently shows that teachers matter for student achievement. Children assigned to a teacher with a one standard deviation higher quality gain in terms of achievement in the order of 0.10 to 0.25 standard deviations (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Kane & Staiger, 2008; Hanushek & Rivkin, 2010). In addition, the economic returns to higher quality teachers can be substantial. For example, Chetty et al. (2014a) show that children assigned to teachers with a higher value-added (i.e., teachers that produce larger achievement gains) attend college more often, earn more and live in better neighborhoods. Staiger & Rockoff (2010) predict a total gain of 330 to 760 thousand dollar in lifetime income for a class that has a one standard deviation better qualified teacher.

It is less clear which teacher characteristics or (sets of) teacher practices matter. Traditional observable characteristics of teachers - often used to determine teacher pay levels - have only little predictive power for measuring differences in teacher quality. With respect to teacher qualifications most studies do not find a relationship between the teacher's highest attained education level and teacher quality.[5] With respect to work experience most studies show that teachers gain in terms of effectiveness in the first two or three years of their career, but that this experience effect levels off after this period.[6]

A lack of knowledge about effective teacher characteristics and practices is problematic for policymakers and school leaders that aim to improve and reward teacher quality. Recent research in the United States reveals that teacher ratings or evaluations made by school principals, mentor teachers or trained evaluators have predictive power for student achievement (e.g. Jacob & Lefgren, 2008; Rockoff & Speroni, 2011; Tyler et al., 2010; Kane et al., 2011; Kane and Staiger, 2012; Grossmann et al., 2013; Kane et al., 2013; Harris & Sass, 2014). Estimates from these studies show that, depending on the tested domain (reading or math) and type of evaluation instrument, a one standard deviation higher evaluation score is related to 0.05 to 0.14 of a standard deviation higher student achievement scores. See Appendix table A.1 for a short overview of this literature and the main estimates. These

[5] See e.g. Hanushek & Rivkin (2006) and Harris & Sass (2011) for reviews of the literature.

[6] See, among others, Rivkin et al. (2005), Clothfelter et al. (2006) and Jacob (2007), and Staiger & Rockoff (2010). Notable exceptions are two recent papers by Harris & Sass (2011) and Wiswall (2013) that find that teacher productivity keeps on increasing with experience (far) beyond the first couple of years on the job.

evaluation systems seem an advancement over many teacher evaluation systems that hardly differentiate in scores between teachers.[7]

In this chapter we use teacher evaluations based on a detailed classroom observation instrument to estimate the predictive power of these teacher evaluation scores on pupil achievement gains. The evaluations were carried out by trained and experienced external evaluators in seven elementary schools in a large city in the Netherlands. The level of detail of the classroom observation protocol or rubric and the set of teacher practices measured is different from previous studies. Our measure includes a rubric of 18 standards and 75 associated teacher practices which are believed to reflect effective teaching. This is more than double the number of standards and associated practice descriptions relative to the studies conducted previously. For example, the Cincinatti's Teacher Evaluation System (TES) studied in Kane et al. (2011) and Tyler et al. (2010) has 8 standards and 29 associated practice descriptions. Another difference is that Cincinatti's TES has four scoring levels for each particular practice (from unsatisfactory to proficient), whereas our rubric has two. The evaluator just had to indicate whether or not the practice was demonstrated by the teacher.

Our main findings can be summarized as follows. We find that a higher score on the teacher evaluation rubric is related to higher pupil achievement in all tested domains. A one standard deviation higher score on this rubric is associated with a 0.15 standard deviation higher pupil test score in math, a 0.18 standard deviation higher score in spelling and a 0.11 standard deviation higher score in reading. Our estimates suggest that gains in pupil achievement are relatively large if a teacher from the bottom quartile of the teacher evaluation distribution is replaced by a teacher from the top quartile. Estimated gains range from 0.24 (reading) to 0.44 (spelling) standard deviations in pupil achievement. These gains are considerably larger than the ones found in Kane et al. (2011) for Cincinatti's TES. Consistent with earlier findings on other teacher evaluation schemes, our evaluation rubric seems to be particularly capable of identifying the weakest teachers, but seems less capable of differentiating between an average teacher and an excellent one.

Our contribution to the literature is twofold. First, we corroborate results in a European context that have been found in a couple of previous studies in the United States. In

[7] Weisberg et al. (2009) show in an analysis of teacher evaluation systems in 14 school districts in the US that most districts only have a binary rating system in which more than 98 percent of teachers rated the highest category (usually labeled "satisfactory").

particular, differences in teacher quality are large and teacher evaluations through classroom observations using a detailed rubric of classroom practices by trained evaluators are useful in identifying such differences. Second, the rubric used in this chapter seems to do a somewhat better job in identifying differences in teacher quality compared to rubrics assessed in earlier literature, especially in identifying the weakest teachers. This could be due to the higher level of detail (i.e., more teacher practices) or to differences in competences being assessed by the rubric. The rubric seems particularly capable of identifying weak teachers, which is important since weak teachers have a negative impact on a student's achievement and later socioeconomic outcomes. This study, together with the small but growing literature that has recently emerged, suggests that teacher evaluations carried out by trained experts have potential in identifying heterogeneity in teacher quality. The results of this research may be used to identify problems of low teacher quality in schools and to design and experiment with subsequent feedback and mentoring schemes to address these problems.

This chapter proceeds as follows. Section 2.2 describes the evaluation rubric and the data. Section 2.3 presents the empirical strategy. Section 2.4 shows the main estimation results. Section 2.5 presents a couple of sensitivity analyses. Section 2.6 concludes.

2.2 Data

We use data on pupils and teachers from grade 1 to 8 from seven elementary schools from a school district in Amsterdam, which is the largest city in the Netherlands. Elementary education in the Netherlands starts when children reach the age of 4 (in grade 1) and ends when they are 12 years old (in grade 8). School age starts at age 5, but the vast majority (>95 percent) of children enters at age 4. The seven schools participated in a large teacher evaluation project that was launched by the municipality. The pupil data contain information on math, spelling and reading test scores from the end of school year 2011/2012 and previous year test scores from the end of school year 2010/2011. Pupil test scores are from tests that are developed by the national test developing agency CITO. Primary schools in the Netherlands use these tests to monitor progress of their pupils throughout primary education. Besides information on test scores we have obtained detailed information on pupil background characteristics such as age, gender, highest attained education of the parents, nationality and whether the child lives in a one-parent family. Teacher data contain teacher experience and the scores on the rubric with the 75 classroom practices. Professionals have identified this list of classroom practices to reflect good teacher practices. Next to data on teachers we have obtained classroom information such as class size, the fraction of girls and the fraction of pupils whose parents are low educated.

In the empirical analysis we use standardized test scores for math, spelling and reading from the school year 2011/2012 as dependent variables. Test scores have been standardized by school year and grade.

Our main independent variable is the total score on the teacher evaluation system (TES). A detailed rubric has been constructed by educational professionals for the purpose of citywide monitoring of teacher quality. The rubric consists of 18 standards and 75 associated classroom practices that are believed by education experts to reflect good teacher practices.[8] These classroom practices are defined in three domains: pedagogical competence, didactical competence and classroom organization competence. The Cronbach's Alphas for all 75 items of the rubric and for the items in the three domains respectively are 0.96, 0.85 (15 items), 0.94 (46 items) and 0.84 (14 items). They are all larger than 0.8, suggesting that the internal consistency or construct validity of the rubric is sound.

Here we only discuss the most salient details of the rubric; appendix table A.2 provides an overview of the 18 standards of the rubric. While teaching a class, teachers were scored on this rubric by professional evaluators that have been specifically trained for the job. The evaluators were all experienced external coaches and/or (former) school leaders that had considerable experience with classroom observations of teachers. The training of the evaluators had a particular focus on consistency in scoring. All classroom practices of the rubric were discussed separately. The evaluations were announced and each evaluation was done by one single evaluator. The teachers were asked to teach a lesson in which they could demonstrate all 75 classroom practices. Teachers could either demonstrate a classroom practice or not, with the evaluator denoting a 1 if the teacher showed the competence and 0 if not. Hence, the score on the rubric may (in theory) range from 0 to 75.

Teachers were evaluated twice, once in the first period of the school year (September-November) and once at the end of the school year (June). No teacher was evaluated by the same rater twice. Following Kane & Staiger (2012), we take the average of the start and the

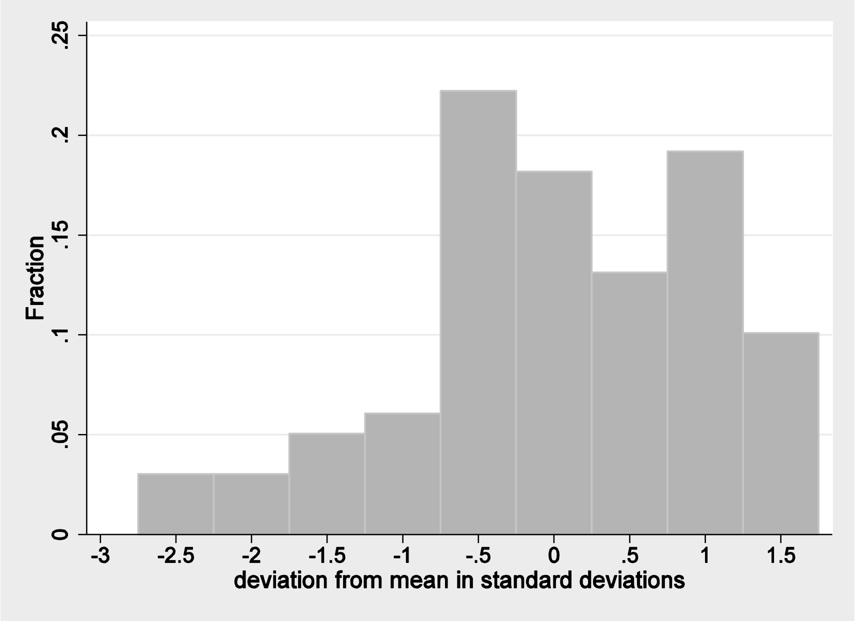
[8] The official competence requirements for teachers that are used by the Education Inspectorate of the Netherlands and that are part of the national Law on Occupations in Education (*Wet Beroepen in Onderwijs*) have been transferred to corresponding observable classroom practices in the rubric.

end score on the rubric. Kane and Staiger advise to use multiple classroom observations per teacher to obtain a more reliable picture of the true quality of the teacher. For 106 teachers we have both scores on the TES. For 19 teachers the end score is missing. There are multiple reasons for the missing values on the end score: some teachers left school during the school year 2011/2012, other teachers were not present in the week the evaluations were carried out due to illness or pregnancy, and yet others were in the middle of a dismissal procedure. These 19 teachers were relatively weak teachers as their score on the rubric was below average (i.e., on average 10 points lower). For these teachers we imputed the score from the end of the school year with the score from the start of the school year, and included an indicator for missing end score in our models. To investigate whether or not our results are influenced by these missing observations, we will present estimates for both the sample with full TES information (*n*=106) and the sample with imputed values for missing teachers (*n*=125). The set of estimates for the sample with full information (i.e. two evaluation scores per teacher) reduces the number of classrooms for which we have a TES-score from 99 to 88.

Pupils can have multiple part-time teachers during a school year. This is very common and related to the large share of female teachers in primary education in the Netherlands. In case a class has been taught by more than one teacher, we weigh the scores on the TES by their relative presence to calculate an average TES score for the classroom. For instance, if teacher *X* teaches three days a week in class *C*, and teacher *Y* the other two days of the week, the TES-score for class *C* is equal to (3/5)*TES score of teacher *X*+(2/5)*TES score of teacher *Y*.

In Figure 2.1 we show the distribution of the standardized TES-scores for our main sample of classes (*n*=99). Standardization has been done by subtracting the mean (52.41) from the original score and dividing it by the standard deviation (13.28) such that the standardized TES-score has mean 0 and standard deviation 1. The distribution is skewed to the left. The 25th percentile of the standardized distribution is equal to -0.63, the median is equal to -0.07 and the 75th percentile is equal to 0.82. The minimum is equal to -2.74 and the maximum is 1.51.

Figure 2.1 The distribution of the classroom teacher evaluation score (N=99 classrooms), average of two observations



We add a large set of covariates to the model. Our most important covariate is the previous test score derived from the end of school year 2010/2011. This previous test score is included as a control for ability differences between pupils. We also include a second degree polynomial of this variable in our models. The previous test scores contain missing values because some pupils only entered the particular school in 2011/2012. We put missing test scores to zero. This is equal to the average of a particular school year and grade because of standardization of the test scores by grade and year. We also include an indicator in the regression model when the previous test score is missing. Besides controlling for previous test scores, we control for other differences between pupils by including a second degree polynomial in age and dummies for gender, nationality, living in a one parent family, retention, and educational level of the parents. In our most comprehensive specification, we also control for observable classroom and teacher differences by including teacher experience[9], class size, the average of the previous test scores, average age, fraction of girls, fraction of pupils with Dutch nationality, fraction of pupils living in a one parent family,

[9] Teacher experience has been weighted in the same way as the TES-score for a classroom. We define this as the teacher experience a classroom of children is confronted with.

fraction of pupils that retained, fraction of pupils with low-educated parents[10], a dummy for classrooms that span multiple grades and a dummy for classrooms that have multiple teachers. We also include school- and grade-fixed effects.

Table 2.1 provides descriptive statistics of our variables. Panel A presents means and standard deviations of pupil characteristics based on the sample for which spelling test scores are available. The average fraction of pupils with low-educated parents equals 38 percent, which exceeds the average of 26 percent in this city. Almost half of the pupils live in a one parent family. Panel B shows descriptive statistics of classroom characteristics for the 99 classrooms for which TES-scores are available. The average of the unstandardized TES-score equals 52.4, with a standard deviation of 13.3. The average class size is 24. Panel C shows teacher characteristics. The vast majority of teachers (88 percent) is female, which is somewhat higher than the national average of 78 percent in Dutch primary education (Ministry of Education, 2013). Average work experience amounts to 19 years, with 13 percent of all teachers having five years or less of work experience. Two percent of the teachers in our sample have obtained a university degree. The remaining 98 percent has a degree at the level of higher vocational education, which is the standard requirement to become a teacher in primary education in the Netherlands. In table A.3 in the appendix we present all pair-wise correlations between our classroom variables.

[10] That is, parents who only finished the lowest level of secondary school or less.

Table 2.1 Descriptive statistics, restricted and unrestricted sample

	Unrestricted sample[c]		Restricted sample[d]	
Panel A: Pupil characteristics	**mean**	**sd**	**mean**	**sd**
Girl	0.50	0.50	0.50	0.50
Age	8.05	2.40	8.69	2.35
Dummy=1 if low educated parents[a]	0.38	0.49	0.38	0.49
Dummy=1 if from one parent family	0.49	0.50	0.51	0.50
Dummy=1 if Dutch nationality	0.90	0.30	0.90	0.30
Dummy=1 if retained	0.07	0.25	0.06	0.23
Number of pupils[b]	2110		1859	
Panel B: Classroom characteristics	**mean**	**sd**	**mean**	**sd**
Classroom teacher evaluation (unstandardized)	52.41	13.28	54.23	12.49
Classroom teacher experience	19.19	9.60	18.99	9.75
Class size	24.17	4.15	23.83	4.01
Classroom spans multiple grades (%)	20.20	40.35	18.18	38.79
Classroom has multiple teachers (%)	33.33	47.38	36.36	47.38
Fraction of girls (%)	50.92	7.74	50.99	7.87
Average age	8.01	2.41	8.08	2.38
Fraction of pupils with low educated parents (%)	38.69	12.64	38.82	12.23
Fraction of pupils from one parent family (%)	49,44	14.79	48,68	14.51
Fraction of pupils with Dutch nationality (%)	88.02	9.197	88.42	8.23
Fraction of pupils that retained a grade (%)	6.20	12.10	5.56	11.91
Number of classrooms	99		88	
Panel C: Teacher characteristics	**mean**	**sd**	**mean**	**sd**
Female (%)	88.24	32.37	89.62	30.64
Higher vocational education as highest level of educational attainment (%)	98.02	14.00	97.59	15.43
Experience in education (years)	19.45	11.61	19.68	11.53
Five years or less experience in education (%)	12.75	33.51	11.90	32.58
In higher pay scale (%)	7.14	25.88	8.54	28.11
Tenure (%)	92.16	27.02	95.24	21.42
Size of contract (% of FTE)	87.19	18.16	88.03	17.69
Number of teachers	125		106	

(a) Finished lowest track of secondary school or less.(b) The pupil characteristics are shown for the (estimation) sample of pupils for which spelling scores are available. (c) The unrestricted sample is the complete sample of teachers and their classes for which the start-of-the-year teacher evaluation score is available (but not necessarily the end-of-school year score). (d)The restricted sample is the sample of teachers and their classes for which both start and end-of-year teacher evaluation score is available.

2.3 Empirical strategy

The main goal of our empirical analysis is to estimate the relationship between the teacher evaluation scores and pupil performance gains. We employ a similar value-added type of model as used by Rockoff & Sperroni (2011) and Kane et al. (2011). These types of models account for the fact that teachers and pupils are not randomly assigned to classes within schools. We estimate a model in which the standardized test scores are related to standardized TES-scores in the following way:

$$\text{(M.1)} \quad Y_{ic} = \beta_0 + \beta_1 TES_c + \beta_2 Y_{t-1,ic} + \beta_3 Y_{t-1,ic}^2 + \beta_4' \boldsymbol{X}_{ic} + \beta_5' \boldsymbol{C}_c + \varphi_s + \theta_g + \varepsilon_{ic},$$

Herein is Y_{ic} the standardized test score of pupil i in school s in grade g in classroom c, and the previous test score is represented by $Y_{\mathrm{t-1,ic}}$. $X_{\mathrm{ic}}$is a vector of pupil characteristics and C_{c} is a vector consisting of classroom characteristics. This includes teacher experience and teacher experience squared. To ease notation, we leave out indices for schools s and grades g for pupil and classroom variables. The term φ_{s} represents school-fixed effects (6 dummies, because we have 7 schools) and θ_{g} represents grade fixed effects (7 dummies, because we have 8 grades). Note that, by including the school and grade-fixed effects, we use variation between classrooms within grades within schools. The parameter of interest is β_1, which represents the association between the test score of the pupil and the teacher evaluation score. The estimated coefficients can be interpreted in terms of standard deviations.

The 'value-added' type of model as presented in (M.1) aims to account for non-random assignment of teachers and pupils to classes. Rothstein (2010) criticizes these models because unobserved pupil characteristics make classrooms more easy or difficult to teach. These unobserved characteristics may play a role in the assignment of teachers to classes, which could yield biased estimates. Although unobservable characteristics could confound the estimates of β_1, Nye et al. (2004), Kane & Staiger (2008), and Kane et al. (2013) show that experimental estimates - in situations where teachers have been randomly assigned to classrooms - are consistent with value-added estimates that result from non-experimental value-added models as long as these non-experimental value-added models control for student's prior achievement. Chetty et al. (2014b) find that teacher value-added models that control for a student's prior test scores - as we do in this chapter - exhibit little bias in forecasting teachers' impact on pupil achievement.[11] In the next section we investigate to what extent our results are affected by (possible) non-random assignment of teachers to classes based on a large number of observable pupil and classroom characteristics.

Another concern may be that test scores are derived from the same school year as the TES-scores. This contemporaneous measurement could potentially confound the estimates of β_1 if there are unobserved class characteristics that independently and systematically affect both an evaluator's measurement and pupil achievement (Kane et al., 2011). For instance, if an evaluator encounters a teacher in a classroom with a high level of social cohesion, he may evaluate this teacher differently than he would have done if he encountered the same teacher in a classroom with a low level of social cohesion. At the same time higher social cohesion may result in positive peer effects that raise pupil achievement, causing the estimates of β_1 to be biased. Kane et al. (2011) propose to use pupil achievement data from the previous or next year compared to the year the evaluations are carried out. Unfortunately, pupil achievement gains data and information on assignment of teachers to classes is not available to us for other school years. However, Tyler et al. (2009) and Caridad et al. (2016) show that estimates are in the same order of magnitude when same-year pupil achievement data are used instead of previous or next-year data. Moreover, we also control for a large number of observable class characteristics that may be correlated with both the teacher TES score and pupil test score gains, such as the average previous classroom test scores, the fraction of pupils from a one parent family, the fraction of non-native pupils and the fraction of retained pupils.

2.4 Estimation results

Tables 2.2, 2.3 and 2.4 present the main estimates of the relationship between the standardized TES-score and pupils' math, spelling and reading achievement, respectively. Each table has 5 columns, which include different sets of covariates. In column (1) we present the association between the TES-score and the test scores without any controls; in column (2) we add school and grade fixed effects; in column (3) we include previous test scores and other pupil characteristics; in column (4) we add all other classroom information and in column (5) we also include teacher experience. By including the school- and grade-

[11] They conclude this from comparing estimates of teacher value added with and without controlling for previously unobserved parent characteristics, as well as from applying a quasi-experimental research design based on changes in teaching staff.

fixed effects we obtain an indication of the extent to which our results are affected by nonrandom sorting of teachers *across* schools and grades. By adding the previous test scores and the classroom variables we obtain an indication of the extent to which our results are affected by (possible) nonrandom sorting of teachers to classes *within* schools and grades.

Table 2.2 Relationship between standardized teacher evaluation score and pupil's math score

	Dependent variable: standardized math score				
Independent variable:	(1)	(2)	(3)	(4)	(5)
Standardized teacher evaluation score	0.120***	0.087**	0.173***	0.143***	0.154***
	(0.040)	(0.040)	(0.050)	(0.051)	(0.051)
School and grade fixed effects	no	yes	yes	yes	yes
Previous test scores and other pupil characteristics	no	no	yes	yes	yes
Classroom variables	no	no	no	yes	yes
Teacher experience	no	no	no	no	yes
Observations	2084	2084	2084	2084	2084
Number of classrooms	99	99	99	99	99
R-squared	0.017	0.054	0.416	0.440	0.449

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1. Other pupil characteristics include sex, a dummy variable indicating the education level of the parents (three categories), a dummy indicating whether the pupil lives in a single-parent family, a dummy indicating whether the pupil is from Dutch origin, and age. Classroom variables include class size, multi-grade classroom, the number of teachers teaching in that class, the share of female pupils and the share of pupils with low educated parents.

Table 2.3 Relationship between standardized teacher evaluation score and pupil's spelling score

	Dependent variable: standardized spelling score				
Independent variable:	(1)	(2)	(3)	(4)	(5)
Standardized teacher evaluation score	0.078*	0.123***	0.103**	0.152***	0.178***
	(0.041)	(0.046)	(0.047)	(0.047)	(0.046)
School and grade fixed effects	no	yes	yes	yes	yes
Previous test scores and other pupil characteristics	no	no	yes	yes	yes
Classroom variables	no	no	no	yes	yes
Teacher experience	no	no	no	no	yes
Observations	2110	2110	2110	2110	2110
Number of classrooms	99	99	99	99	99
R-squared	0.009	0.038	0.333	0.365	0.375

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1. Other pupil characteristics include sex, a dummy variable indicating the education level of the parents (three categories), a dummy indicating whether the pupil lives in a single-parent family, a dummy indicating whether the pupil is from Dutch origin, and age. Classroom variables include class size, multi-grade classroom, the number of teachers teaching in that class, the share of female pupils and the share of pupils with low educated parents.

Table 2.4 Relationship between standardized teacher evaluation score and pupil's reading score

	Dependent variable: standardized reading score				
Independent variable:	(1)	(2)	(3)	(4)	(5)
Standardized teacher evaluation score	0.039	0.076	0.034	0.083	0.107**
	(0.037)	(0.046)	(0.045)	(0.050)	(0.050)
School and grade fixed effects	no	yes	yes	yes	yes
Previous test scores and other pupil characteristics	no	no	yes	yes	yes
Classroom variables	no	no	no	yes	yes
Teacher experience	no	no	no	no	yes
Observations	2135	2135	2135	2135	2135
Number of classrooms	99	99	99	99	99
R-squared	0.002	0.026	0.366	0.387	0.398

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1. Other pupil characteristics include sex, a dummy variable indicating the education level of the parents (three categories), a dummy indicating whether the pupil lives in a single-parent family, a dummy indicating whether the pupil is from Dutch origin, and age. Classroom variables include class size, multi-grade classroom, the number of teachers teaching in that class, the share of female pupils and the share of pupils with low educated parents.

2.4.1 Main estimates

The estimated coefficient for the TES-score change when including the school and grade-fixed effects (in column 2), but the direction of the change is different for math than for reading or spelling (i.e., from 0.120 to 0.087 for math and from 0.078 to 0.123 for spelling). The direction of the change differs also between test domains when adding a pupil's previous test score (in column 3) and classroom variables (in column 4): while for math the estimated coefficient increases from 0.087 in column (2) to 0.173 in column (3) and decreases to 0.143 in column (4), the coefficient for spelling drops from 0.123 to 0.103 and increases from 0.103 to 0.152, respectively. These changes in the coefficients suggest that teachers are assigned to classes in a non-random manner. However, given the difference in the direction of change in the coefficients between the tested domains we find no indication that good teachers are systematically assigned to either particular good or weak classes.[12]

In any case, regardless of the type and direction of sorting of teachers into classrooms, it does not seem to affect the statistical significance of the estimated coefficients for math and spelling. They are statistically significant in all columns of Tables 2.2 and 2.3. This suggests that these associations are robust to the inclusion of a range of covariates. When we control for teacher experience, the estimates slightly rise because of the negative but not significant correlation between teacher experience and the average teacher evaluation score (correlation coefficient of -0.13, p-value of 0.19). Our estimates are robust to, alternatively, specifying experience with a series of indicator variables of experience (i.e. 10-year classes of experience).

Based on the results in column (5) with all relevant controls, which is our preferred specification, we find that a higher TES-score is associated with higher pupil achievement scores for all three test domains and that this association is statistically significant. For math we find that, on average, a pupil gains 0.15 standard deviation if he is assigned a teacher that has a one standard deviation higher score on the rubric. For spelling the estimated gain is about 0.18 standard deviations. The estimated coefficient for reading is 0.11 standard

[12] Appendix table A.4 shows the relationship between previous year test scores and the start-of-year teacher evaluation score based on a regression with school- and grade-fixed effects. It seems that better teachers (based on the start-of-year teacher evaluation score) are assigned to weaker pupils regarding math (see column 1). No significant relationship was found however for spelling and reading and point estimates are of the opposite sign.

deviations.[13] This coefficient points at a somewhat weaker association with reading scores, however still significant at a 5-percent significance level in the model with all controls.[14]

2.4.2 Nonlinear effects

The TES-score has been treated linearly in the analyses presented so far. To investigate the possibility of a nonlinear relationship we split up the TES-score in quartile dummies. Table 2.5 presents the results of regressions in which dummies for the quartiles of the TES-score have been included instead of the linear TES-score. The presented specification includes all covariates.

The estimates suggest that replacing a teacher from the lowest quartile of the TES-score distribution by a teacher from the upper quartile yields test score gains of 0.37 standard deviations in math, 0.44 in spelling, and 0.24 in reading.[15] These estimated gains are relatively large, both compared to findings in the earlier literature on the predicted performance gains by having a teacher with a one standard deviation higher score on other evaluation rubrics[16], as well as compared to the effects of a couple of well-known interventions such as reducing class size or an extra year in school.[17] The rubric particularly seems to differentiate between the weakest teachers and the rest. The estimated coefficients for the upper three quartiles differ significantly from the lowest quartile that serves as the reference category. The point estimates generally increase from the second quartile to the top quartile, but differences are not statistically significantly different among estimates for the second, third and top quartile. Holtzapple (2004) and Kane et al. (2011) find a similar pattern for Cincinatti's TES system as well as Jacob and Lefgren (2008) for principal ratings of teachers.

[13] It should be noted that the estimates for reading, spelling and math are not statistically significantly different from each other. Therefore, we should be cautious with interpreting these results as if the strength of the relationship is strongest for spelling and weakest for reading.

[14] A review of value-added estimates of teacher effectiveness in terms of standard deviations in pupil test scores Hanushek and Rivkin (2010) shows that estimated coefficients are larger for math than for reading in every study.

[15] The difference in the average TES-score between teachers in the lowest quartile (i.e. 34 competences shown) and teachers in the highest quartile (i.e. 68 competences shown) comes down to 2.5 standard deviations.

[16] Comparable estimates in Kane et al. (2011) for Cincinatti's Teacher Evaluation System are 0.09 for Math and 0.13 for reading. Kane & Staiger (2012) find estimates in the order of 0.05 to 0.11 standard deviations for four different rubric instruments used in the Measures of Effective Teaching Project.

[17] For instance, the cumulative effects of being in a class with five less pupils for three consecutive years on cognitive skills are estimated to be about 0.15 standard deviations (Frederiksson et al. 2013; Krueger, 1999). Estimates of the effect of a year in school on scores on cognitive tests are in the order of 0.2 standard deviations (e.g. Angrist & Krueger, 1991; Hansen et al., 2004, Webbink & Gerritsen, 2013).

Table 2.5 Relationship between quartiles of teacher evaluation score and pupil math, spelling and reading scores

	Dependent variable:		
	math	spelling	reading
Independent variable:	(1)	(2)	(3)
Indicator for TES-score between 25th and 50th percentile	0.326***	0.399***	0.149
	(0.097)	(0.096)	(0.109)
Indicator for TES-score between 50th and 75th percentile	0.290***	0.337***	0.248**
	(0.107)	(0.104)	(0.120)
Indicator for TES-score between 75th and 100th percentile	0.371***	0.440***	0.236*
	(0.115)	(0.112)	(0.120)
School and grade fixed effects	yes	yes	yes
Previous test scores and other pupil characteristics	yes	yes	yes
Classroom variables	yes	yes	yes
Teacher experience	yes	yes	yes
Observations	2084	2110	2135
Number of classrooms	99	99	99
R-squared	0.453	0.381	0.399

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1. Other pupil characteristics include sex, a dummy variable indicating the education level of the parents (three categories), a dummy indicating whether the pupil lives in a single-parent family, a dummy indicating whether the pupil is from Dutch origin, and age. Classroom variables include class size, multi-grade classroom, the number of teachers teaching in that class, the share of female pupils and the share of pupils with low educated parents.

2.5 Sensitivity analyses

2.5.1 The impact of missing end-of-school-year evaluation scores

As a robustness check we analyze whether our main results are sensitive to missing end-of-year scores. For this purpose we conduct the same analysis on the subset of teachers for which we have both start-of-school-year and end-of-school-year evaluation scores. We exclude those teachers for which no end-of-school-year score on the rubric is available. This reduces our sample of teachers from 125 to 106 and our sample of classrooms from 99 to 88.

Table 2.6 presents a set of estimates that are in line with the specification reported in column (5) of Tables 2.2 to 2.4. We conclude that our estimated coefficients are unlikely to be influenced by imputation of teachers' average evaluation score by their start-of-year evaluation score in case of missing end-of-year scores.

Table 2.6 Relationship between standardized TES-score and math, spelling and reading scores, restricted sample of classrooms with two evaluations per teacher

	Dependent variable:		
	math	spelling	reading
Independent variable:	(1)	(2)	(3)
Standardized TES-score	0.145***	0.153***	0.089
	(0.051)	(0.048)	(0.058)
School and grade fixed effects	yes	yes	yes
Previous test scores and other pupil characteristics	yes	yes	yes
Classroom variables	yes	yes	yes
Teacher experience	yes	yes	yes
Observations	1833	1859	1863
Number of classrooms	88	88	88
R-squared	0.447	0.370	0.391

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1. Other pupil characteristics include sex, a dummy variable indicating the education level of the parents (three categories), a dummy indicating whether the pupil lives in a single-parent family, a dummy indicating whether the pupil is from Dutch origin, and age. Classroom variables include class size, multi-grade classroom, the number of teachers teaching in that class, the share of female pupils and the share of pupils with low educated parents.

2.5.2 Predictive power of one evaluation per teacher instead of two

A relevant question in the light of evaluation costs is how the validity of one observation for predicting pupil achievement gains relates to the validity of the average score of two observations per teacher. Table 2.7 shows the results when using one observation score per teacher, either the start-of-school-year evaluation (panel A) or the end-of-school-year evaluation (panel B). We find that estimation results are rather similar to the results in which the average score of two observations per teacher is used.[18]

This finding, though encouraging, does not in itself promote the use of just one observation per teacher for teacher evaluation purposes. Kane and Staiger (2012) calculate that evaluation reliability increases by about 50 percent when using the average of two classroom observations (of different evaluators) per teacher instead of just one (and even further when using more evaluations).

[18] Estimates in table 8 should be compared with the ones in table 6 on the restricted sample of 88 classrooms where two evaluations per teacher have been carried out.

One might be concerned that there is selective response of teachers to their first evaluation score. For example, we might expect teachers that are confronted with a low start-of-year score to show bigger improvements due to receiving a low score in the first evaluation. If there would be such a selective response, we would expect to find different relationships between the teacher evaluation score and pupils' test scores when looking at the relationship with the end-of-year evaluation scores rather than with the start-of-year evaluation scores. The fact that we do find rather similar estimation results suggests that this concern is not likely to play a role.

Table 2.7 Relationship between standardized score on start-of-year or end-of-year teacher observation and pupil math, spelling and reading scores, restricted sample of class rooms with two evaluations per teacher

	Dependent variable:		
	math	spelling	reading
Independent variable:	(1)	(2)	(3)
Panel A: Evaluation score at start-of-school-year classroom observation	0.157***	0.141**	0.073
	(0.057)	(0.054)	(0.051)
Panel B: Evaluation score at end-of-school-year classroom observation	0.098**	0.125***	0.091*
	(0.044)	(0.041)	(0.055)
School and grade fixed effects	yes	yes	yes
Previous test scores and other pupil characteristics	yes	yes	yes
Classroom variables	yes	yes	yes
Teacher experience	yes	yes	yes
Observations	1833	1859	1863
Number of classrooms	88	88	88

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1. Reported coefficients are from two separate regressions. The first regression includes the start-of year evaluation score and the second regression includes the end-of-year evaluation scores as the relevant independent variable. Other pupil characteristics include sex, a dummy variable indicating the education level of the parents (three categories), a dummy indicating whether the pupil lives in a single-parent family, a dummy indicating whether the pupil is from Dutch origin, and age. Classroom variables include class size, multi-grade classroom, the number of teachers teaching in that class, the share of female pupils and the share of pupils with low educated parents.

2.5.3. Estimates for subsets of classroom practices of the rubric

In this section we investigate to what extent scores on predefined subsets of the classroom practices of the rubric predict pupil achievement gains. The complete set of 75 classroom practices can be divided along two dimensions: by the level of competences (basic versus complex) and by the type of competences (pedagogical, didactical and organizational). The rubric has 45 classroom practices which have been classified as reflecting basic competences that every beginning teacher should master. Thirty classroom practices have been identified as reflecting complex competences that teachers should be able to demonstrate after some years on the job. With respect to the type of competences, the rubric has 15 pedagogical classroom practices, 46 didactical classroom practices and 14 classroom practices reflecting classroom organization competence (see also appendix table A.2). Table 2.8 shows the estimation results for the model with all covariates and school and grade fixed effects. The scores on the three different sub-domains are strongly correlated, see appendix table A.5.[19] We therefore carried out separate regressions for each sub-domain instead of entering the sub-domains jointly in one regression.[20]

We find that the total score on the rubric does a better job in predicting pupil achievement gains than the scores on the subsets of the evaluation rubric. Nevertheless, we find positive coefficients for all five sub-domains. These coefficients are statistically significant for all sub-domains with respect to pupil math scores, and for three out of five sub-domains regarding spelling. In case of reading, none of the coefficients differ statistically significant from zero.

Saving on the number of measured classroom practices seems to have a cost in terms of predictive power of pupil achievement. At the same time evaluating teachers with a rubric with a smaller set of classroom practices is not likely to save a lot in terms of evaluation costs, since evaluators would probably still need a full lesson to score a smaller number of classroom practices. Research on larger samples of teachers and classrooms may identify certain sets of measured classroom practices that are more predictive for pupil achievement than others.

[19] Similar findings have been reported elsewhere for various evaluation rubrics.

[20] We also investigated to what extent we could differentiate between (subsets of) competences by including them in the regressions simultaneously. However, disentangling factors of competence is difficult due to problems of multicollinearity as we work with 99 classrooms and (highly) correlated competences. A principle component analysis reveals that the first component explains about 25 percent of the variance, and that the first 36 components of the 75 items explain about 90 percent. However, we could not give clear interpretations of the identified principle components, which kept us away from using them in our analysis.

Table 2.8 Relationship between standardized score on various sub-domains of the evaluation rubric by level and type of competences and pupil math, spelling and reading scores

	Dependent variable:		
	math	spelling	reading
Independent variable:	(1)	(2)	(3)
Panel A: Level of competence			
1. Standardized score on basic competences (45 items)	0.123**	0.091*	0.060
	(0.049)	(0.053)	(0.056)
2. Standardized score on complex competences (30 items)	0.149***	0.109*	0.046
	(0.053)	(0.059)	(0.062)
Panel B: Type of competence			
3. Standardized score on pedagogical competences (15 items)	0.100**	0.060	0.066
	(0.044)	(0.047)	(0.051)
4. Standardized score on didactical competences (46 items)	0.071*	0.070	0.045
	(0.042)	(0.044)	(0.053)
5. Standardized score on classroom organization competences (14 items)	0.080*	0.073*	0.069
	(0.043)	(0.040)	(0.048)
School and grade fixed effects	yes	yes	yes
Previous test scores and other pupil characteristics	yes	yes	yes
Classroom variables	yes	yes	yes
Teacher experience	yes	yes	yes
Observations	2084	2110	2135
Number of classrooms	99	99	99

Reported coefficients in this table are of five independent regressions carried out separately including one sub-domain of the evaluation rubric at a time. Standard errors clustered on classroom between brackets. *** p<0.01, ** p<0.05, * p<0.1. Other pupil characteristics include sex, a dummy variable indicating the education level of the parents (three categories), a dummy indicating whether the pupil lives in a single-parent family, a dummy indicating whether the pupil is from Dutch origin, and age. Classroom variables include class size, multi-grade classroom, the number of teachers teaching in that class, the share of female pupils and the share of pupils with low educated parents.

2.6 Discussion and conclusion

This research reports the results of a program aimed at measuring teacher competences in primary education in a large city in the Netherlands. We obtain a set of estimates suggesting that teachers with higher evaluation scores on a detailed classroom observation instrument produce greater average gains in pupil achievement. Estimates of these gains range from 0.11 (reading) to 0.19 (spelling) standard deviations if a pupil is assigned to a teacher with a one standard deviation higher evaluation score on the rubric. This finding is consistent with prior work in the United States.[21]

In addition, we find that the rubric is particularly successful in distinguishing weak teachers from other teachers, but less so in differentiating between an average and an excellent teacher. This observation is also consistent with earlier US findings. Predicted pupil achievement gains from being assigned to a teacher in the highest quartile instead of a teacher in the lowest quartile in the evaluation rubric vary between 0.24 (reading) to 0.44 (spelling) standard deviations.

These results suggest that evaluations made by trained experts on a detailed rubric have potential to address the problem of weak teacher quality. One of the advantages of using these rubrics with detailed standards for teacher practices over more subjective ratings by principals or value-added estimates is that the score on the rubric provides signals to teachers and principals as to in what (clusters of) competences or classroom practices improvements can be made. This information may be effectively used in personal development plans to improve teacher quality. Promising in this respect is that Taylor & Tyler (2012) show that repeated evaluations and targeted feedback to (mid-career) teachers by trained experts based on a detailed rubric raise pupil achievement, particularly in the years after evaluation and feedback have been carried out. Particularly encouraging is that they find largest effects on pupil achievement gains for the weakest teachers (i.e. with low ex-ante test-score teacher value-added estimates or with low teacher evaluation scores). The scores on the teacher evaluation rubric may also be used for personnel decisions. Rockoff et al. (2012) show that principals are more likely to retain their effective teachers (and not to retain their weak teachers) when they are provided with estimated teacher effects. All surveyed school principals of the involved schools in our study agree that the rubric is a good instrument to take into account for decisions on promotion or dismissal of their teachers.[22] Similar research on larger samples of teachers and classrooms could give more insight in which particular classroom practices matter most for teacher quality.

[21] Comparable estimates for Cincinatti's TES are 0.09 in math and 0.08 in reading (Kane et al., 2011). Rockoff & Speroni (2011) report a coefficient of 0.05 higher math achievement of a one standard deviation higher rating by mentor teachers.

[22] Furthermore, 86 percent of the principals agree that the rubric is a good instrument to distinguish weak from good teachers. Sixty percent of surveyed teachers are positive about measuring teacher competences by classroom observations, as compared to 33 percent being neutral and 7 percent being negative. Just 13 percent of teachers thinks that classroom observations do not succeed in obtaining a good picture of their competences.

While our results show significant associations between the score on the teacher evaluation rubric and pupil achievement gains, we cannot rule out the possibility that the true causal relationship is different. Although we control for a large set of pupil and classroom characteristics, including baseline test scores, we may not have been able to rule out all biases due to possible non-random matching of teachers to classes on unobservable characteristics. Carrying out the same teacher evaluations in a situation of random assignment of teachers to classes could shed more light on the possible bias in our results.

Another possible source of bias in our estimates would arise when teachers adjust their behavior during the observed lessons because they know they will be evaluated. We cannot exclude the possibility that this happened. However, this would bias our results only if certain subgroups of teachers along the quality distribution adjusted their behavior more than others and if this adjusted behavior affects evaluation scores. The fact that quite some teachers scored very low on the evaluation rubric suggests that it is not very easy to adjust your teaching behavior in such a way that you are able to receive a higher score simply by preparing for the evaluation. There are dozens of classroom practices on which teachers are evaluated, nearly half of which consist of complex teaching practices which are not likely to be mastered simply by preparing for a certain lesson.

A point of attention for any teacher evaluation system is the reliability of the evaluations when carried out by different evaluators. Rockoff & Speroni (2011) find variation in the leniency between evaluators, particularly in the case of evaluations by mentors. Unfortunately we do not know which rater evaluated which teacher. If we would have known this, we could have added rater-fixed effects to the regressions. However, this problem of rater effects is likely to be reduced with independent (external) evaluators who probably have fewer incentives to be lenient. In addition, Kane & Staiger (2012) conclude that it seems possible to constrain tendencies to score too lenient or too harsh when training of evaluators is taken seriously. In our case, all raters were particularly trained for using the rubric with particular attention being given to obtaining consistency in the scores across raters. Both conditions seem to be met in our case. Therefore, we conclude that evaluating teachers by classroom observations by trained and external evaluators using a detailed rubric of teacher practices is likely to be a promising avenue for identifying differences in teacher quality and implementing targeted teacher improvement measures.

Appendix tables

Table A.1 Recent literature on relationship between teacher evaluations and pupil test scores

Study	Where and when	Evaluation tool	Evaluator	N teachers	Findings (a)
Jacob & Lefgren (2008)	School district in west US, grades 2-6, 2002/03	Subjective overall rating on a scale of 1-10 of teacher effectiveness	Principal	162 (reading) and 112 (math)	0.07** (reading) and 0.14** (math). Same year.
Rockoff & Speroni (2011)	New York city, grades 3-8, 2003/4 - 2007/08	Formative evaluations on six competences with each between five and eight items. Every two months.	Trained full-time mentor	1857 (math) and 1879 (reading)	0.02* (math) and 0.01 (reading), same year. 0.03-0.05** (math) and 0.01-0.02** (reading). Next year achievement growth.
Tyler et al. (2010)	Cincinatti public schools, 2000/01-2008-09	Cincinatti's Public Schools' Teacher Evaluation System. 8 standards consisting of 29 classroom practices, scored on a 4-point scale. Two to six evaluations.	Three times by an assigned peer evaluator (high-performing teacher external to the school), once by a local school administrator	100 (math) and 206 (reading)	0.07** (math) and 0.09** (reading). Previous year achievement growth.
Kane et al. (2011)	Cincinatti public schools, grades 3-8, 2003/04-2008/09.	Cincinatti's TES, see under Tyler et al. (2010)	See Tyler et al. (2010)	207 (math) and 365 (reading)	0.08 (reading) and 0.09 (math). Same year. Significance not reported. Difference between top and bottom quartile is 0.09** in math and 0.13** in reading.
Kane & Staiger (2012)	Six districts in the US, grades 4-8, 2008/09 - 2009/10	Four observation instruments: CLASS (3 domains, 11 dimensions, 7-point scale)), FFT (2 domains, 8 components), UTOP (4 sections, 22 subsections, 5-point scale), and MQI (6 elements, 3-point scale). 4-8 evaluations per teacher (video-taped)	Particularly trained teachers for evaluation (17 to 25 hours per rater). Often experienced and high educated teachers. About 70 percent had a degree higher than bachelor, and more than 75 percent had six or more years of experience.	1333	0.05*** (MQI), 0.06*** (FFT), 0.08*** (CLASS) and 0.11*** (UTOP). Prior year math scores. Difference between teachers in top and bottom quartile of the distribution.
Grossmann et al. (2013)	New York City middle school ELA teachers, 2008.	PLATO (10 elements) and CLASS (2 domains and six elements). Six days of instruction observed.	Carefully trained raters.	24 (between 3-6 years of experience)	0.07* (Guided Practice), 0.11** (Explicit Strategy Instruction) on ELA scores. Same year.
Kane et al. (2013)	Six districts in the US, grades 4-8, 2008/09-2009/10	Observation instrument FFT (Framework for Teaching), 2 domains, 8 components.	See Kane & Staiger (2012)	303 (middle school grades) and 392-403 (elementary grades math and ELA)	Elementary grades: 0.11* math and 0.05 ELA. Middle school grades: 0.09** (math) and 0.08*** (ELA). Previous year achievement gain of a 1-point increase at a 4-point FTT scale.
Harris & Sass (2014)	School district in Florida, grades 2-10, 1999/00 - 2007/08.	Subjective overall rating on a 1-9 scale of teacher effectiveness	Principal	237 (math) and 231 (reading)	0.06** (math) and 0.03 (reading). Same year.
Araujo et al. (2016)	Ecuador, kindergarten, 2012-2013	CLASS, 3 domains (emotional support, classroom organization and instructional support). Within each domain a number of dimensions.	A limited group of trained coders scoring teachers on the basis of filmed lessons.	451	0.06** (language), 0.08*** (math) and 0.06** (executive function). Next year achievement growth.

(a) Reported estimates represent the predicted higher pupil achievement expressed in terms of standard deviations corresponding to a one standard deviation higher evaluation score, unless stated otherwise. *** p<0.01, ** p<0.05, * p<0.1. Same year means teacher evaluation has been carried out in the same school year over which pupil achievement gains are measured.

Table A.2 The teacher evaluation rubric "Amsterdamse Kijkwijzer"[23]

Indicator: The teacher.......	Type of compe-tences (a)	Number of sub-items (classroom practices)	Average share of items demonstrated
Clearly sets high expectations	p	4	85
Instruction takes account of relevant differences between pupils	p	4	80
Assimilation of subject matter takes account of relevant differences between pupils	p	4	71
Provides extra instruction and time to learn for weaker pupils	p	3	75
Makes clear how the lesson fits in with earlier lessons	d	4	78
Clearly states the lesson goals at the beginning of the lesson	d	3	78
Provides insight into the organization of the lesson	d	3	75
Clearly explains the lesson material and assignments	d	4	85
Provides feedback to pupils	d	6	69
Checks that lesson goals have been reached	d	5	66
Stimulates reflection via interactive instruction and work methods	d	2	68
Encourages pupils to think out loud	d	2	84
Teaches pupils strategies for thinking and learning	d	6	73
Encourages pupils to reflect on differing solution strategies	d	5	50
Encourages the use of control activities (checks)	d	3	60
Stimulates application of what is learned	d	3	67
Spends the planned time on the lesson goals	o	5	85
Ensures the lesson follows an adequate planning	o	9	77
	Total	75	74

(a) p = pedagogical competence; d = didactical competence; o = organizational competence. (b) This is the average percentage of classroom practices shown by the teachers in our sample over two observations.

The "Amsterdamse Kijkwijzer" rubric has been developed by KPC Groep in cooperation with the school boards and with a program that was set up by the municipality of Amsterdam to improve the quality of primary education in Amsterdam called KBA (Kwaliteitsaanpak Basisonderwijs Amsterdam). In the rubric, the competences identified in the national competence standard for teachers (the so-called SBL-competences) and the most important aspects from the framework used by the Inspectorate of Education have been translated to concrete observable behavior.

[23] A list of the 75 underlying sub-items under the 18 indicators of the rubric is available upon request.

Table A.3 Matrix of pair-wise correlations between classroom variables (n=99). P-values in italics

No..	Description:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Class room TES-score	1,00																
2	Class room teacher experience	-0,13	1,00															
		0,19																
3	Class size	-0,13	-0,03	1,00														
		0,19	*0,74*															
4	Classroom spans multiple grades (%)	-0,23	0,01	0,17	1,00													
		0,02	*0,91*	*0,08*														
5	Classroom has multiple teachers (%)	0,29	-0,15	-0,16	0,02	1,00												
		0,00	*0,13*	*0,12*	*0,86*													
6	Fraction of girls (%)	-0,09	0,00	-0,14	-0,06	-0,14	1,00											
		0,36	*0,99*	*0,18*	*0,57*	*0,17*												
7	Average age	0,22	-0,07	-0,14	-0,24	0,19	-0,09	1,00										
		0,03	*0,50*	*0,17*	*0,02*	*0,06*	*0,40*											
8	Fraction of pupils with low educated parents (%)	0,11	-0,02	0,03	0,02	-0,03	-0,09	0,41	1,00									
		0,27	*0,81*	*0,76*	*0,84*	*0,79*	*0,38*	*0,00*										
9	Fraction of pupils from one parent family (%)	0,16	0,00	-0,16	-0,05	0,14	-0,09	0,27	0,15	1,00								
		0,11	*0,97*	*0,11*	*0,60*	*0,16*	*0,37*	*0,01*	*0,15*									
10	Fraction of pupils with Dutch nationality (%)	-0,03	0,23	-0,09	-0,39	-0,15	-0,03	0,05	-0,09	0,11	1,00							
		0,80	*0,02*	*0,38*	*0,00*	*0,14*	*0,74*	*0,62*	*0,36*	*0,28*								
11	Fraction of pupils that retained (%)*	-0,15	0,01	0,22	0,33	-0,08	0,08	-0,55	-0,17	-0,16	-0,13	1,00						
		0,15	*0,94*	*0,03*	*0,00*	*0,41*	*0,45*	*0,00*	*0,10*	*0,11*	*0,21*							
12	Average test score math	0,27	0,01	0,05	-0,12	-0,08	0,33	-0,08	0,03	-0,09	-0,05	0,14	1,00					
		0,01	*0,96*	*0,61*	*0,25*	*0,44*	*0,00*	*0,45*	*0,77*	*0,35*	*0,60*	*0,17*						
13	Average test score spelling	0,17	0,09	0,10	-0,02	-0,21	0,29	-0,04	0,04	-0,05	0,03	0,18	0,73	1,00				
		0,10	*0,35*	*0,30*	*0,85*	*0,04*	*0,00*	*0,71*	*0,71*	*0,64*	*0,74*	*0,08*	*0,00*					
14	Average test score reading	0,07	0,02	-0,11	0,09	-0,01	0,17	0,01	0,03	0,00	-0,03	0,20	0,43	0,62	1,00			
		0,50	*0,86*	*0,29*	*0,36*	*0,89*	*0,10*	*0,90*	*0,74*	*0,98*	*0,74*	*0,05*	*0,00*	*0,00*				
15	Average previous test score math	0,00	-0,15	0,06	-0,28	-0,12	0,05	0,06	-0,12	-0,10	0,02	-0,18	0,32	0,10	-0,03	1,00		
		0,98	*0,15*	*0,56*	*0,00*	*0,23*	*0,61*	*0,55*	*0,22*	*0,34*	*0,84*	*0,07*	*0,00*	*0,32*	*0,77*			
16	Average previous test score spelling	-0,03	-0,06	0,25	-0,12	-0,29	0,14	0,05	0,15	-0,15	-0,11	-0,08	0,30	0,33	0,10	0,39	1,00	
		0,75	*0,55*	*0,01*	*0,24*	*0,00*	*0,16*	*0,64*	*0,14*	*0,14*	*0,27*	*0,44*	*0,00*	*0,00*	*0,33*	*0,00*		
17	Average previous test score reading	-0,03	-0,13	0,22	0,01	-0,16	0,13	0,09	0,21	-0,01	-0,24	-0,06	0,21	0,21	0,17	0,18	0,78	1,00
		0,77	*0,19*	*0,03*	*0,90*	*0,11*	*0,20*	*0,40*	*0,03*	*0,90*	*0,02*	*0,53*	*0,04*	*0,03*	*0,10*	*0,08*	*0,00*	

* Most pupils are retained in grade 2: they enroll in the school year when they turn four and stay an extra year in kindergarten after two years of kindergarten (grades 1 and 2).

Table A.4 Assignment of teachers to classes: relationship between start-of-year teacher evaluation score and previous year math, spelling and reading score

	Dependent variable: previous year score on:		
	math	spelling	reading
Independent variable:	(1)	(2)	(3)
1. Standardized start-of-year teacher evaluation score	-0.141***	0.038	0.047
	(0.047)	(0.043)	(0.035)
School and grade fixed effects	yes	yes	yes
Observations	2084	2110	2135
Number of classrooms	99	99	99

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1

Table A.5 Matrix of pair-wise correlations among scores on subsets of classroom practices by type and level and the total score on the rubric

No.	Description of type of competences *(number of items in parentheses)*	1	2	3	4	5	6
1	Basic (45)	1.00					
2	Complex (30)	0.90	1.00				
3	Pedagogical (15)	0.89	0.88	1.00			
4	Didactical (46)	0.94	0.98	0.87	1.00		
5	Classroom organization (14)	0.86	0.71	0.64	0.73	1.00	
6	Total score on the rubric (75)	0.98	0.97	0.91	0.99	0.81	1.00

All correlations are significant at a 1-percent significance level.

3.

The effect of schooling vouchers on higher education enrollment and completion of teachers: A regression discontinuity analysis[24]

Abstract

This chapter investigates the effects of schooling vouchers for teachers. We study effects on enrollment and completion of higher education programs, and on the retention of teachers in the education sector. We do this by exploiting a fuzzy regression discontinuity design. The discontinuity in the probability of being assigned a voucher arises due to budget constraints in the first application period. Our estimates suggest that effects of voucher assignment on both higher education enrollment and completion rates are in the order of 10 to 20 percentage points as measured five and a half years after application. Relative to a baseline enrollment rate of 77 percent and a baseline completion rate of 54 percent (i.e. of applicants that were not assigned a voucher), these effect estimates correspond to a 12 to 29 percent higher enrollment and to a 17 to 42 percent higher completion. Effects on enrollment and completion are relatively small for shorter studies (up to one year) and for teachers that had already started at the time of application. The teacher voucher crowds out funding by schools out of their regular professional development budgets as well as own contributions by teachers. Our results suggest small positive effects of voucher assignment on retention in education as measured four years after application.

[24] This is joint work with Roel van Elk; Steeg, M. van der, and R. van Elk, 2015, The effect of schooling vouchers on higher education enrollment and completion of teachers: A regression discontinuity analysis, *CPB Discussion Paper*, no. 305.

3.1 Introduction

This chapter reports about the effects of a public teacher voucher program in which teachers are eligible to receive a voucher to enroll in a bachelor or master degree program. The program was set up by the Dutch government in 2008 to promote participation of teachers in professional development activities that lead to a higher education level or to acquire more skills and knowledge at the same education level. The teacher voucher scheme is targeted at teachers from primary to higher vocational education. The teacher voucher not only consists of compensation for teachers for admission fees and costs of travel and study material, but also of compensation for their employer to arrange a substitute teacher while they are on study leave. The combined value of these two voucher elements may amount to a maximum of 30 thousand euro per voucher application. Nearly 400 million euro has been granted to about 40 thousand teachers and schools over the first five years after the introduction of the voucher scheme (2008-2013).

Raising teacher quality is one of the main concerns of the Dutch government, as it is in many countries. A large literature shows that teacher quality is an important driver of pupil performance. Children assigned to a teacher with a one standard deviation higher quality gain in terms of achievement in the order of 0.10 to 0.25 standard deviations (see e.g. Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Kane & Staiger, 2008; Hanushek & Rivkin, 2010). Moreover, higher teacher quality also seems to positively affect later labour market outcomes of pupils (Chetty et al., 2014). Teacher professional development activities in general and raising the share of teachers with a Master Degree in particular could potentially be one of the channels through which teacher quality and thereby pupil performance can be raised. The literature on the effects of teachers having a Master degree as compared to a bachelor degree on pupil performance shows a mixed picture, however, with some studies finding positive effects, while other studies do not find any effects or even negative effects (see review in Harris & Sas, 2011).[25] There is also a literature on the effects of (providing more money for) professional development activities for teachers on pupil performance, showing mixed evidence (e.g. Angrist & Lavy, 2001; Jacob & Lefgren, 2004; Garet et al., 2008 & 2010; Harris & Sas, 2011). The heterogeneity of the type of professional development activities and specific interventions (e.g. providing schools with money for

[25] This is a predominantly US literature. It is uncertain whether the same results apply in other education systems and with a possibly other variation in value added of master versus bachelor teacher studies.

training of teachers or directly offering specific training programs) as well as that of target groups (e.g. math versus language teachers; teachers at average versus at bad performing schools) prevents us to draw general conclusions from this literature.

In this chapter we investigate the effects of teacher education vouchers on enrollment in and completion of higher education degree programs, as well as on retention of teachers in the profession. We investigate these effects by exploiting a discontinuity in the probability of (ever) having been assigned a voucher that was caused by budget restrictions in the first year of the voucher scheme. A large number of teachers applied for a voucher in a relatively brief period of one-and-a-half month. This led to a situation in which the teachers vouchers have been assigned on a first-come-first-served basis and in which an unexpected cutoff-date was in place after which suddenly no applications for vouchers could be granted anymore. Several validity checks on the regression discontinuity design are carried out in this chapter.

Estimating effects on enrollment and completion is relevant because in order to trigger an effect on teacher productivity one should at least find effects of voucher assignment on enrollment and, even more important, on completion. Large effects of vouchers for adult workers on training or schooling participation are not obvious. Two earlier studies of training vouchers for adult workers found that considerable deadweight loss was involved with these vouchers (Schwerdt et al, 2012; Hidalgo et al., 2014). Deadweight loss arises when training vouchers are being used to finance participation of employees in training that would have been undertaken anyway, that is, in the absence of these vouchers.[26] Both studies are based on randomized experiments. The study by Schwerdt et al. (2012) studies the effects of a Swiss training voucher experiment for adults of all education levels. Hidalgo et al. (2015) investigate effects of a Dutch training voucher experiment for predominantly low-skilled adult workers. Appendix Table A1 gives a comparison of the three voucher schemes and of the main findings of effects on training / higher education participation. Schwerdt et al. (2012) find a deadweight loss of 30 percent, whereas Hidalgo et al. (2014) find a deadweight loss of sixty percent.

[26] Deadweight loss is a serious risk in any public intervention aimed at promoting training participation among adult workers, not particularly only in case of training vouchers for employees. For instance, Abramovsky (2011) find no evidence of effects on qualification-based training of employer-based incentives for low-qualified employees under the Employer Training Pilots undertaken in the UK between 2002 and 2006. Leuven and Oosterbeek (2004) find disappointing effects on training participation of age-related tax deduction for employers for their employees training expenses. They find that these age-related incentives just postpone training participation among workers rather than increasing it.

This study contributes to the small literature on effects of training vouchers for workers in the following way. First, we study the impact of training vouchers for a specific population that consists entirely of highly educated workers, rather than the general population of workers (Schwerdt et al., 2012) or predominantly low educated workers (Hidalgo et al., 2014). Teachers are in particular an interesting target group since they are crucial for human capital production in a country. Second, we study the effect of much larger vouchers in terms of face values as compared to earlier voucher studies. Another distinctive feature of the teacher voucher scheme is the compensation offered to employers for arranging replacement during study leave of their employees. Third, we investigate effects on the probability of completion of higher education programs as well. These effects are relevant to investigate since longer-term degree programs are involved rather than relatively short study courses or training programs. Effects on completion rates may therefore differ from effects on enrollment rates if there are differences in study dropout and delay among voucher receivers and non-receivers. A fourth contribution of our research is that we also investigate effects of vouchers on retention in the profession. Retention seems particularly important in the case of teachers since recent evidence shows that more experienced teachers produce larger achievement gains among their pupils (e.g. Harris & Sas, 2011; Wiswall, 2013; Gerritsen et al., 2014).

Our main findings are as follows. First, estimates of the effects of voucher assignment on both higher education degree program enrollment and completion rates are in the order of 10-20 percentage points, from a base of 77 percent (enrollment) and 54 percent (completion) for teachers who applied for but never received a teacher voucher. These effect estimates point at a substantial degree of crowding out of other means of funding, a phenomenon that is also found in earlier studies on training vouchers for employees. Deadweight loss of the teacher voucher scheme is estimated at about 80 to 90 percent. Second, our results suggest small positive effects of voucher assignment on retention in education, as measured four years after voucher application. This would be a positive side-effect, since recent studies have found that teacher productivity increases with experience. Third, we have indications of heterogeneous effects of voucher assignment across subgroups by applicant and application characteristics. Effects on both enrollment and completion are larger for teachers who had not started their study yet at the time of application. The abolishment of the possibility to apply for a voucher for a study that was already started is expected to have raised effects of voucher assignment on both enrollment and completion by about five percentage points. Effects on enrollment and completion appear much smaller for studies with duration of a year or less, as compared to longer studies. Effects on retention in the teaching profession appear to be concentrated among teachers working in secondary education and teachers above 35 years old.

This chapter proceeds as follows. Section 3.2 describes the teacher education voucher scheme. Section 3.3 presents the data. Section 3.4 presents the empirical strategy and section 3.5 the main estimation results. Section 3.6 discusses heterogeneous treatment effects. Section 3.7 sheds light on the complier population for which effects can be estimated. Section 3.8 discusses substitution patterns in sources of financing of the higher education programs. Section 3.9 concludes and discusses the implication of our findings.

3.2 The teacher voucher scheme

The Dutch teacher voucher scheme (Dutch name "Lerarenbeurs") was introduced in 2008. It aims to stimulate participation in lifelong learning among teachers in primary and secondary education, intermediate and higher vocational education, and special education. Teachers can use the teacher voucher to enroll in a bachelor or master program. Typically four types of programs are involved. The first type is programs targeted at mastery of specific pedagogical and didactical skills. Master Special Educational Needs is an example of this type. Applications for this particular master account for about 30 percent of all applications in the period 2008-2013, with even larger shares among applications of teachers in primary education (42 percent) and in special education (56 percent). The second type is subject-specific programs. These programs are aimed at either acquiring a certification at the same level in another subject or at acquiring certification in the same subject at a higher level (i.e. at master level instead of at bachelor level). This type of programs is most often applied for by teachers in secondary education. The third type is programs targeted at management skills. The fourth type consists of more generic masters such as pedagogy, theory of education and "learning and innovating".

The teacher voucher consists of two subsidies, one for teachers and one for schools. The teacher receives subsidy for tuition costs up to 3500 euro per year and for study materials and travel costs up to 700 euro per year.[27] The school may receive subsidy for giving the teacher study leave and to arrange a substitute teacher while the teacher is on study leave. This

[27] From 2011 onwards, the maximum subsidy for tuition costs has been raised from 3500 to 7000 euro, and for costs of travel and study material from 350 to 700 euro.

subsidy for study leave is maximized at 160 hours per year per teacher (i.e. half a day per week) for a full-time teacher. This amounts to a maximum of 5200 euro per year for schools in primary education to 6700 euro for schools in higher vocational education.

The most important conditions of the teacher voucher scheme are the following:

- The applicant is a certified teacher.
- The applicant is employed at a school or working at a school on a contract with another agency (i.e. not self-employed).
- The applicant is teaching for at least twenty percent of his or her contract.
- The applicant can only apply once in his or her career for a teacher schooling voucher.
- After completion the applicant should continue working in education for at least a year.[28]
- The study program should be completed at most three years after the end of the subsidy period. If not, the subsidy should be paid back according to the share of credits that were not obtained.[29]

Between 2008 and 2013 almost 40 thousand teachers have been assigned a teacher voucher in seven different application periods. About seventy percent of these vouchers were related to applications for bachelor or master degree programs.[30]

In the first application period in the spring of 2008 a little less than 7500 teachers applied for a voucher. Due to a predetermined maximum budget only around two-thirds of these applications could be awarded a teacher schooling voucher. Vouchers have been awarded on a first-come-first-served basis. It is this budget constraint in the first application period that creates a discontinuity in voucher assignment by day of application that we will exploit to determine effects of voucher assignment on enrollment and completion of degree programs. In later years the yearly budget for new applications for the teacher voucher scheme has been increased further. In total 394 million euro of subsidy is involved with the assigned vouchers between 2008 and 2013, of which 174 million euro is targeted to teachers to compensate them for the tuition fees, travel costs and costs of study material. This implies that the

[28] This condition has been abolished as from 2013 onwards.

[29] This condition has been abolished as from 2013 onwards. Instead, a yearly minimum of 15 ECTS credits should be obtained.

[30] As from 2012 onwards, teachers could only apply for registered bachelor or master degree programs. Applications for short courses or other programs not leading to a bachelor or master degree were not allowed anymore. The analyses in this paper are solely focused on applications for bachelor or master programs.

majority of the total teacher voucher subsidy, that is 220 million euro or 56 percent, is directed towards schools to compensate them for the costs of arranging replacement while their teachers are on study leave.

Appendix B provides more facts and figures about the teacher voucher scheme and about professional development of teachers in the Netherlands, as well as about the policy context in which the teacher voucher was introduced.

3.3 Data

3.3.1 Data sources

We use administrative data from three different databases. The first database is called *ABL* and provides data from the administration of the voucher scheme. This database contains information on applications and assignments of vouchers in the first application period and reapplications and assignments in subsequent application years. Applicant characteristics taken from this database are gender, birth date, sector of work and the appointment in FTE. The application characteristics we use are program duration, a dummy indicating whether or not the applicant already started the higher education course at the time of application, and the day of application in the first application period in 2008.

The second source is a national database containing data on higher education enrollment and completion, which is called *BRON HO*. From this database we derive information on whether the applicants actually enrolled in higher education courses after their application and whether they succeeded to complete these courses. We use information regarding the period 2008-2013. Data have been merged to the voucher scheme administration data from *ABL* by a unique personal identifier.

The third source is a national database of teachers. This database contains information on salary and the region of the teacher. We have merged these data with the data from the other two sources by using a unique personal identifier as well.

The data from the three different sources have been supplied by *Dienst Uitvoering Onderwijs (DUO)* that operates the teacher voucher scheme.

3.3.2 Sample

First we have selected all applicants of the first application round. Within this group we make two sub-selections, one on sector of work and one on study type. First, we select applicants working in primary, secondary and special education. These are the sectors for which the budget constraint was binding, that is, sectors where more applications were received than vouchers were available. This implies that we do not consider applications from teachers working in intermediate or higher vocational education, since there is no discontinuity in voucher assignment as in the other sectors. In total 12 percent of all applications in the first round are left out of the estimation sample for this reason.

The second selection is that we only select applications for registered higher education studies. This implies we do not consider applications for (predominantly) brief courses.[31] The reason is that we cannot track enrollment and completion in these courses for all applicants, particularly for the ones that did not receive a voucher. Applications for these brief courses account for about one-third of all applications for these sectors in the first application period. The budget share is lower at an estimate of around 20 percent according to information on assigned amounts of money per applicant. This is due to differences in study length and due to the condition that the voucher subsidy to schools for arranging replacement for teachers on study leave can only be made for higher education courses.

These selections result in an estimation sample of 4,220 teachers out of 7,485 applicants in the first application round. These teachers applied in a relatively brief period of 47 days in the spring of 2008.

3.3.3 Summary statistics

Table 3.1 shows descriptive statistics for the total estimation sample of 4,220 teachers, for the subgroups of applicants on either side of the cut-off date (before: N=3,037, after: N=1,183), and for the voucher recipients (N==3392) versus the ones that never received a voucher (N=828).

[31] It should be noted that, as from 2012 onwards, teacher vouchers could only be assigned for registered higher education studies (i.e. bachelors, masters or premasters), not for brief courses anymore. That is, the type of applications we consider in this paper is the exact same type as the type that is targeted in the current teacher voucher scheme.

Table 3.1 **Descriptive statistics sample of teacher schooling voucher applicants for higher education studies**

	(1)	(2)	(3)	(4)	(5)
Variables	All	Before cut-off date	After cut-off date	Ever received voucher	Never received voucher
Panel A					
Applicant characteristics					
Female	0.75 [0.72]	0.75	0.75	0.75	0.74
Age	37.8 [42.8]	37.6	38.1	37.7	38.1
Working in Randstad region	0.38 [0.41]	0.38	0.40	0.38	0.40
Working in primary education	0.45 [0.56]	0.44	0.46	0.45	0.46
Working in secondary education	0.42 [0.35]	0.43	0.39	0.44	0.36
Working in special education	0.13 [0.10]	0.12	0.15	0.12	0.18
Gross monthly salary (2008) at appointment of 1 FTE	2926 [3213]	2930	2915	2920	2954
Appointment in FTE	0.84 [0.78]	0.84	0.84	0.84	0.84
Application characteristics					
Already started higher education program at time of application	0.22	0.23	0.21	0.19	0.33
Program duration (in years)	2.27	2.30	2.18	2.30	2.13
Panel B					
Treatment variables					
Received voucher in first application period (2008)	0.65	0.90	0.00	0.80	0.00
Received voucher in any of first seven application periods (2008-2013)	0.80	0.94	0.46	1.00	0.00
Panel C					
Outcome variables					
Ever having been enrolled in higher education (2008-2013)	0.91	0.93	0.87	0.95	0.77
Completed higher education program (2008-2013)	0.69	0.72	0.61	0.73	0.54
Proportion completed higher education program of those enrolled (over 2008-2013)	0.75	0.77	0.70	0.77	0.69
Still enrolled in higher education in 2013 but did not complete yet	0.06	0.06	0.09	0.07	0.03
Still working in education in 2012	0.88	0.89	0.86	0.89	0.85
N	4,220	3,037	1,183	3,392	828
Proportion of all applicants	1.00	0.72	0.28	0.80	0.20

Country averages are presented between brackets.

Panel A shows descriptive statistics of the applicants and applications. Applicant characteristics are compared to the total relevant teacher population as well (see population averages between brackets).[32] This comparison shows that voucher applicants are younger than the average teacher population (about five years), and are somewhat more likely to be

[32] Population averages are calculated from a national teacher database provided to us by *Dienst Uitvoering Onderwijs (DUO)*.

female (3 percentage points). Applicants are somewhat more likely to work in schools outside the urbanized Randstad region (3 percentage points), whereas their salary is lower than average (almost 10 percent less) in line with their lower than average age. Their appointment is somewhat larger than average (0.06 FTE). The probability of applying for a teacher voucher for a higher education study is below average for teachers in primary education and above average for teachers in secondary and special education.[33]

Panel B reports group means for the treatment variable, i.e. being assigned a voucher. Ninety percent of all applicants before the cutoff-date were assigned a voucher in the first application period versus zero percent of the applicants after the cut-off date. Due to reapplications and assignment of vouchers in later application periods the difference in the probability of eventual assignment of a voucher has become smaller over time: 94 versus 46 percent. The difference is still sizeable and statistically significant. This results in the discontinuity in voucher assignment around the cut-off date in the first application round that we exploit in this chapter, as also illustrated in Figure 3.1.

Panel C reports group means for our two main outcome variables: ever been enrolled in higher education within the period 2008-2013 and completed a higher education program somewhere in 2008-2013. Completing a higher education program is defined as obtaining either a Bachelor or a Master degree. Both enrollment (93 versus 87 percent) and completion (72 versus 61 percent) rates are higher for the group that applied before the cut-off date than for the group that applied after the cut-off date. The differences are much smaller however than the differences in voucher assignment. On average three quarters of the group that was ever enrolled in higher education during 2008-2013 has succeeded in completing a higher education program within this period. This share is higher in the group before the cut-off date than in the group that applied after the cut-off date (77 versus 70 percent). The next section presents the effect estimates of voucher assignment on higher education enrollment and completion.

[33] A relatively large share of applications from teachers in primary education was made up by applications for (brief) courses not being a bachelor or master course. This possibility ended in 2011.

Appendix Table D1 shows the same descriptive statistics by sector of work. Most notable differences in terms of applicant characteristics are the relative large share of female teachers among applicants in secondary education (58 percent versus 47 percent in the population of secondary school teachers) and the relatively larger appointments in terms of FTE in primary and special education. The proportion of teachers who already started the higher education program at the time of voucher application is markedly larger in secondary education (26 percent) than in special education (16 percent) and program duration in secondary education is also markedly longer than in the other two sectors (0.7 years longer). Whereas higher education enrollment shares are the same in all three sectors, completion shares are markedly larger in primary education than in secondary education (75 versus 62 percent).[34]

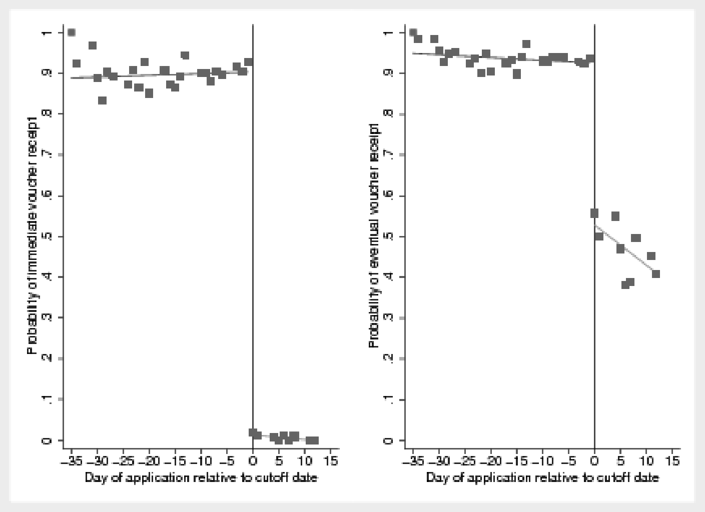
3.4 Empirical strategy and validity checks

3.4.1 Empirical strategy

The main goal is to identify the causal effect of being assigned a teacher schooling voucher on higher education enrollment and completion as well as on retention in the teaching profession. To do so, we have to take into account that there are differences between teachers who did and who did not receive a teacher schooling voucher and that these differences will have separate effects on the outcomes of interest. To identify causal effects, we employ a fuzzy regression discontinuity design (Campbell, 1969; Troachim, 1984; Hahn et al., 2001). We exploit the limited budget for teacher schooling vouchers in the first application round leading to a greater number of applications than could be granted. Vouchers have been assigned on a first-come-first-served basis. This situation results in a clear discontinuity in the probability of immediately being assigned a voucher around a cut-off date as can be seen in the left panel of Figure 3.1.

[34] This may have to do with longer average program duration in secondary education (0.7 years longer). Teachers in secondary education more frequently report serious bottlenecks in terms of study intensity (40 percent), the combination of the study with the private situation (30 percent) and the time that is made available by the school for doing the study (21 percent), see Vink et al. (2012).

Figure 3.1 Relationship between day of application and probability of immediate (left panel) and eventual (right panel) voucher assignment



We would have faced a sharp RD design if all teachers who applied before the cut-off date would have been assigned a voucher and all applicants after the cut-off date would not have been assigned a voucher. There are two reasons however why the discontinuity is not sharp, but fuzzy. The first reason is that a limited share (i.e. less than 10 percent) of the applications before the cut-off date did not meet the conditions of the teacher voucher scheme and was therefore not assigned a voucher. The second reason is that teachers who applied after the cut-off date could reapply for a voucher in later years. While 94 percent of applicants that applied before the cut-off date in the first application rate are assigned a voucher, 46 percent of those that applied after the cut-off date are also awarded a voucher at some point. This results in a drop at the cut-off date in the probability of ever receiving a voucher of approximately 40 percentage points, as can be seen in the right panel of Figure 3.1.

Treatment effects in case of a fuzzy RD can be estimated by two-stage-least-squares, as in an instrumental variables approach (Hahn et al, 2001). This is what we do in this chapter. The following first stage equation is estimated:

(1) $$V_i = \beta_0 + \beta_1 D_i + f(T_i) + \beta_2 X_i + \eta_i$$

where *V* is a dummy indicating voucher assignment in any of the years 2008-2013, *D* is a dummy variable indicating whether the application was received before or after the cut-off date *c* in the first application round (with D = 1 if T >= c and D = 0 if T < c), *f* (T) is a smooth function of the day of application which is allowed to be different at either side of the cutoff, *X* is a vector of predetermined applicant and application characteristics and η is an error term. β_1 is the effect of application after the cutoff-date on the probability of ever having been assigned a voucher over the period 2008-2013.

The second stage equation then uses the predicted values of voucher assignment from the first stage equation to produce the parameter of interest α_1, which is the effect of voucher assignment on the outcomes of interest *Y*, which is either enrollment, completion or retention in the teaching profession.

$$(2) \qquad Y_i = \alpha_0 + \alpha_1 \hat{V_i} + f(T_i) + \alpha_2 X_i + \varepsilon_i$$

Again, *f* (T) is a smooth function of the day of application which is allowed to be different at either side of the cutoff, *X* is a vector of predetermined applicant and application characteristics and ε is an error term.

The effect estimates we present in this chapter are treatment effects on the so-called compliers or local average treatment effects (LATE). A complier is defined in our case by the subset of teachers who are assigned a voucher if they apply before the cut-off date, but are not assigned a voucher if they apply after the cut-off date.[35] We will present an analysis that characterizes the complier population to some extent, that is, showing subgroups according to predetermined applicant and application characteristics that are either more or less likely to be compliers. This characterization of the complier population gives some idea about the external validity of our estimation results.

3.4.2 Assumptions and validity checks

For applying an instrumental variables estimation approach in a regression discontinuity setting a couple of conditions should hold.

[35] This is to distinguish from never-takers and always-takers. These are teachers who would never (always) be assigned a teacher voucher, regardless of applying before or after the cut-off date.

A first condition is that there should be no weak instruments problem. This implies in our case that applying after the cutoff should have an effect on the probability of ever being assigned a voucher that is strong enough. First stage estimates of the effect of application after the cut-off date on voucher assignment are presented for various bandwidth samples in Table 3.2. These estimates indicate that the after cut-off date dummy is a strong instrument for voucher assignment, causing an exogenous drop in voucher assignment of about 40 percentage points. The F-statistics are well above the minimum threshold of 10 suggested by Staiger & Stock (1997) which implies that we do not have a weak instrument problem. Figure 3.1 shows this graphically.

Table 3.2 First stage estimates of effect of application after cut-off date on probability of voucher assignment

	Bandwidth (days around the cut-off date)			
Specification	7	14	21	all
Effect of application after cutoff	-0.368***	-0.383***	-0.399***	-0.395***
	(0.0380)	(0.0304)	(0.0296)	(0.0289)
F-statistic	90.05	159.12	182.11	187.50
Applicant and application controls	Y	Y	Y	Y
Order of polynomial of day of application and interaction term with cut-off date	1	1	1	1
N	1,435	2,468	3,064	4,220

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The estimates are from regressions with a linear control for the day of application and its interaction term with the dummy indicating whether the application was done after the cut-off date. Applicant controls are sex, age category (5 categories), sector of work (three categories), baseline gross monthly salary, appointment size and the region of work (inside or outside Randstad region). Application controls are a dummy indicating whether the applicant had already started and program duration (four categories).

A second condition is that the exclusion restriction assumption should hold. This assumption implies that crossing the cut-off date cannot impact the outcomes of interest except through its effect on voucher receipt. This assumption is not testable. It is not directly clear however why applying (just) after the cut-off date would have a direct effect on the outcomes of interest, other than through its effect on voucher receipt.

A third condition is that the distribution of the baseline covariates should not change discontinuously at the threshold. We check this by both conducting a graphical analysis as well a formal estimation, as suggested by Lee and Lemieux (2010). Figures 3.2 and 3.3 show the distribution of the baseline applicant and application characteristics over the full application period. It can be seen that there are no indications of discontinuities around the cut-off date.

Figure 3.2 Applicant characteristics by day of application in first application round

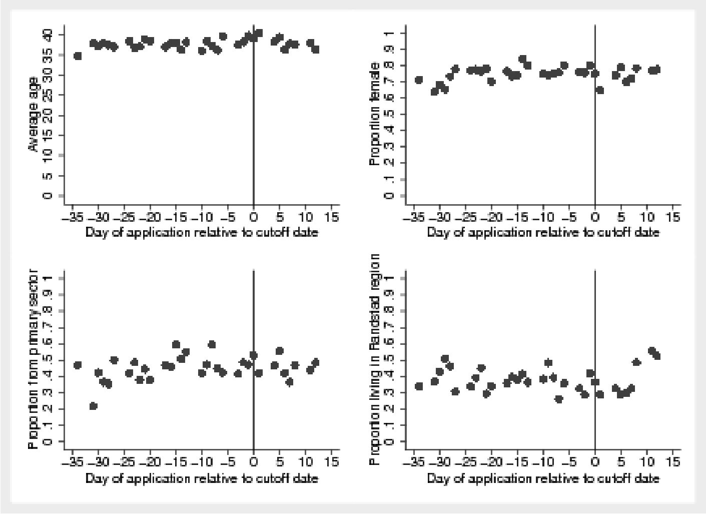
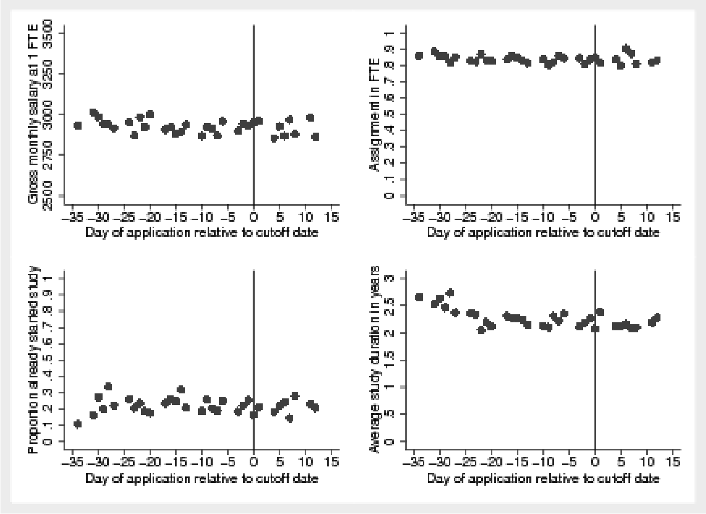


Figure 3.3 Applicant and application characteristics by day of application in first application round



The formal test produces RD estimates for the covariates. These estimates are shown in Table 3.3 for four different estimation samples ranging from seven days around the cut-off date to the full sample of all applicants. The vast majority of the RD estimates are statistically insignificant for the baseline covariates. The most notable estimate is that on program duration in the full sample that is 0.13 years lower of applicants after the cut-off date. This is related to somewhat longer program duration of the very early applicants, which can also be seen in Figure 3.3. The other discontinuity samples show no statistically significant differences in program duration before and after the cut-off date. By carrying out effect analyses on smaller bandwidth samples around the cut-off date we attempt to mirror a situation in which we locally have a randomized experiment. This should make it less likely that any unobserved characteristics are unbalanced between applicants on different sides of the cut-off date.

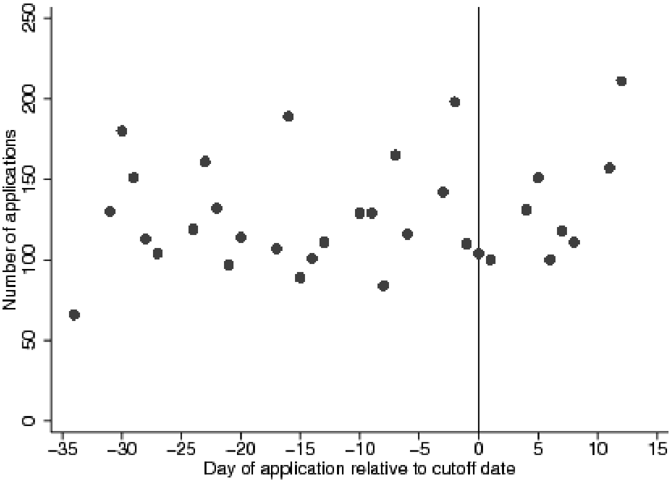
Table 3.3 OLS estimates of application after cut-off date on pre-determined applicant and application characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	female	age	working in primary education	gross salary 2008 (€)	assign-ment in FTE	Randstad region	already started	program duration (years)
1) +/- 7 days around cutoff	-0.0411*	0.491	0.0131	3.368	0.0042	-0.00368	-0.0230	-0.0648
	(0.0228)	(0.548)	(0.0263)	(24.00)	(0.0106)	(0.0240)	(0.0213)	(0.0528)
N	1434	1435	1435	1356	1365	1364	1435	1431
2) +/- 14 days around cutoff	-0.0242	0.291	-0.00745	1.572	0.0007	0.0394**	-0.0152	-0.0221
	(0.0172)	(0.418)	(0.0201)	(18.29)	(0.0084)	(0.0190)	(0.0166)	(0.0396)
N	2466	2468	2468	2316	2332	2331	2467	2461
3) +/- 21 days around cutoff	-0.0147	0.202	-0.00451	-2.652	-0.0035	0.0388**	-0.0157	-0.0346
	(0.0159)	(0.385)	(0.0185)	(16.97)	(0.0076)	(0.0175)	(0.0153)	(0.0360)
N	3062	3064	3064	2886	2906	2905	3063	3057
4) All	0.00142	0.409	0.0227	-13.83	-0.0099	0.0207	-0.0171	-0.126***
	(0.0149)	(0.353)	(0.0170)	(16.06)	(0.0069)	(0.0162)	(0.0142)	(0.0329)
N	4217	4219	4220	3990	4017	4015	4219	4213

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1; No control for day of application.

Another condition for generating a causal effect estimate in regression discontinuity designs is that each individual has imprecise control over the assignment variable, i.e. the cut-off date in our case. We check the plausibility of this assumption by plotting the number of applicants per day against the day of application (as suggested by Lee & Lemieux, 2010; Schochet et al., 2010), see figure 3.4.

Figure 3.4 Number of applications by day of application



If individuals would have had knowledge about the cut-off date, we would expect to see a spike in the number of applications just before the cut-off date. We do not observe such a pattern however. Instead, the number of applications received per day seems to have a rather stable weekly pattern with clear spikes on every Tuesday, probably because teachers have more often finalized their applications in the weekend. A simple test proposed by McCrary (2008) to test whether there is a discontinuity in the density around the cutoff also indicates imperfect control of individuals over applying before or after the cut-off date. Table 3.4 shows the outcomes of a regression of the number of applications on the day of the week the application was received and a dummy variable indicating whether the application was done before or after the cut-off date. This test shows that the number of applications received per day is not significantly lower or higher after the cut-off date, the difference being 3.6 applications per day higher after the cutoff on an average of 130 applications per day.

Table 3.4 Formal test on discontinuity in the density of the assignment variable

Dependent variable	(1) Estimate on number of applications per day	(2) Standard error
Application after cut-off date	3.6	9.7
Day of week (reference = Monday)		
Tuesday	43.4***	12.5
Wednesday	-19.3	13.0
Thursday	-14.6	13.0
Friday	-27.6**	12.5
Constant	129.7***	9.2
N	33	

*** p<0.01, ** p<0.05, * p<0.1.

3.5 Main Results

We use parametric specifications to carry out the instrumental variables analyses. The preferred shape of the smooth function of the day of application turns out to depend somewhat on the size of the bandwidth. The preferred specification is determined by using the Akaike Information Criterion, as suggested by Lee and Lemieux (2010). We report results for a variety of bandwidths, ranging from seven days around the cut-off date to the full sample of all applicants. Outcomes are measured over the period 2008-2013 for enrollment and completion, and for 2012 for retention. The results should be interpreted as estimates of medium-term effects, given that we consider applicants of the first application period in 2008.

3.5.1 Effects on higher education enrollment

Figure 3.5 shows the relationship between the day of application and the actual share of higher education enrollment. The figure also shows fitted lines on either side of the cutoff using a quadratic fit. We observe a small drop in higher education enrollment after the cut-off date. This drop is likely to result from the difference in voucher assignment at the cutoff (see also Figure 3.1). If receiving a voucher had a large impact on higher education enrollment we would have expected higher education enrollment to fall rapidly after the cut-off date.

Figure 3.5 Proportion ever having been enrolled in higher education during 2008-2013 by day of application

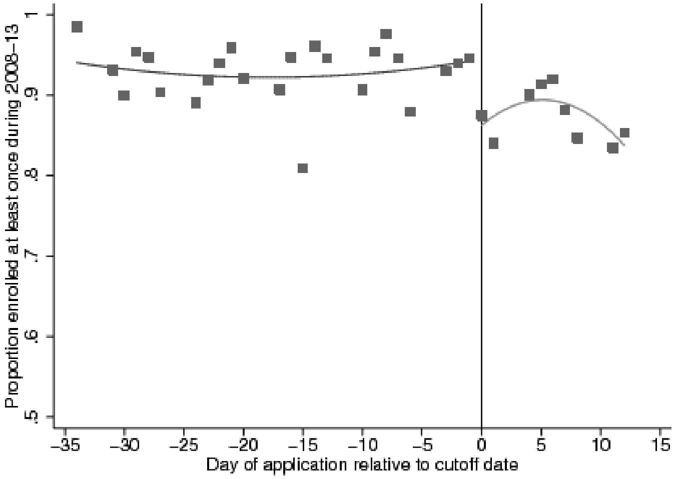


Table 3.5 shows the results from simple OLS estimates of the effect of voucher receipt on higher education enrollment. Effects are shown for four different bandwidths: 7, 14 and 21 days around the cutoff, and the full sample of all applicants. The OLS estimates with all controls (see row 3) suggest that voucher assignment increases higher education enrollment by about 16-21 percentage points. However, selection of voucher assignment on observables raises concerns that selection on unobservable characteristics may still bias the estimates. We estimate the IV model discussed in Section 3.3 to address this concern.

Table 3.5 OLS estimates of effect of voucher assignment on probability of higher education enrollment in period 2008-2013

	Bandwidth (days around the cut-off date)			
Specification	7	14	21	all
(1) No controls	0.127***	0.147***	0.164***	0.173***
	(0.0200)	(0.0158)	(0.0155)	(0.0150)
(2) Adding applicant and application characteristics	0.149***	0.160***	0.176***	0.186***
	(0.0218)	(0.0165)	(0.0160)	(0.0155)
(3) Adding day of application and interaction term with cutoff	0.176***	0.160***	0.198***	0.211***
	(0.0254)	(0.0165)	(0.0190)	(0.0182)
Preferred order of polynomial of day of application and interaction term with cut-off date	1	0	2	2
Control group mean	0.82	0.80	0.79	0.77
N	1435	2468	3064	4220

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The preferred order of the polynomial of day of application and its interaction term with a dummy indicating whether the application was done after the cut-off date is chosen using Akaike's information criterion. Preferred estimates are presented in bold. Applicant controls are sex, age category (5 categories), sector of work (three categories), baseline gross monthly salary, appointment size and the region of work (inside or outside Randstad region). Application controls are a dummy indicating whether the applicant had already started and program duration (four categories).

Table 3.6 shows IV estimation results for a range of bandwidths and smooth functions of the day of application. Throughout this chapter, we show results of different smooth functions of the day of application up to a quadratic polynomial. This follows Gelman and Imbens (2014), who argue that estimates based on higher order polynomials can be misleading. The effect estimates from the preferred specification based on the Akaike Information Criterion are presented in bold. This is a quadratic specification at bandwidths of at least 21 days around the cutoff and a zero order specification at smaller bandwidths. Effect estimates from the preferred specification vary between 9 and 22 percentage points higher enrollment in higher education due to voucher assignment. The estimates of the preferred specification are all statistically significant at the 1 percent significance level. Our preferred IV estimates are roughly in the same range as our OLS estimates. This suggests little bias in OLS effect estimates.

Table 3.6 IV estimates of effect of voucher assignment on probability of higher education enrollment in period 2008-2013

	Bandwidth (days around the cut-off date)			
Specification	7	14	21	all
Polynomial of day of application and interaction term with cut-off date of order:				
Zero	**0.0922*****	**0.137*****	0.125***	0.123***
	(0.0338)	**(0.0250)**	(0.0242)	(0.0229)
One	0.219***	0.0938	0.0991*	0.0905*
	(0.0795)	(0.0594)	(0.0544)	(0.0525)
Two	0.351***	0.215***	**0.221*****	**0.208*****
	(0.121)	(0.0824)	**(0.0804)**	**(0.0728)**
Preferred order of the polynomial of day of application and interaction term with cut-off date	0	0	2	2
Applicant and application controls	Y	Y	Y	Y
N	1,435	2,468	3,064	4,220

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The preferred order of the polynomial of day of application and its interaction term with the dummy indicating whether the application was done after the cut-off date is chosen using Akaike's information criterion. Preferred estimates are presented in bold. Applicant controls are sex, age category (5 categories), sector of work (three categories), baseline gross monthly salary, appointment size and the region of work (inside or outside Randstad region). Application controls are a dummy indicating whether the applicant had already started and program duration (four categories).

Remarkably, these estimates of the voucher effect on higher education enrollment are pretty much in line with self-reports of teachers in a questionnaire that was carried out among voucher applicants in 2011. Thirteen percent of teachers who received a voucher in the first application period report they would not have started the study program if they would not have received a teacher voucher (N=787 respondents).

Deadweight loss

On the basis of these estimation results we calculate a bandwidth for the deadweight loss of the voucher scheme. We do this in a similar way as done by Hidalgo et al. (2014). Instead of using descriptive statistics on enrollment and voucher utilization we use estimation results of a reduced form estimation of the effect of applying after the cutoff on enrollment. Furthermore we use estimation results of the first stage regression of the effect of applying after the cut-off date on the probability of being assigned a voucher. We calculate lower and upper bounds for deadweight loss by using the highest and lowest preferred effect estimate of the effect of applying after the cutoff on higher education enrollment. The calculations are summarized in Table 3.7.

Table 3.7 Deadweight loss calculation of teacher vouchers

	all voucher applicants irrespective of starting status		applicants for studies that have not been started	
	lower bound DWL	upper bound DWL	lower bound DWL	upper bound DWL
(1) Effect on enrollment of application after cutoff (a)	-0.08	-0.04	-0.11	-0.05
(2) Effect on voucher assignment of application after cutoff (b)	-0.37	-0.44	-0.40	-0.43
(3) Crowding out = (1) - (2)	29%	40%	30%	37%
(4) Deadweight loss = (3)/(2)	**78%**	**91%**	**74%**	**87%**
Bandwidth sample (number of days around the cutoff)	21 days	7 days	21 days	7 days
Order of polynomial of control for day of application and interaction term with after cutoff dummy	2	0	2	0

Notes: the smallest and the largest preferred effect estimates are taken to calculate the upper and lower bound for deadweight loss of the teacher voucher scheme.
(a) This is the so-called reduced form estimate.
(b) This is the so-called first stage estimate.

The calculations suggest that the average deadweight loss of the teacher voucher scheme is between 78 and 91 percent. This is larger than the deadweight loss of the Swiss voucher scheme (30 percent, Schwerdt et al., 2012) and of the Dutch training voucher scheme (60 percent, Hidalgo et al., 2014). This difference in deadweight loss could be due to several factors. First, differences in the way the vouchers have been assigned may play a role. The Schwerdt et al. (2012) and Hidalgo et al. (2014) studies involve voucher experiments in which vouchers have been randomly assigned to workers irrespective of their desire to follow training. We observe rather low utilization rates in both studies. Our study involves vouchers for which teachers could apply, and therefore involves workers that are already interested in schooling. It would be interesting to see if deadweight loss of teacher vouchers would decrease if these teacher vouchers would be (randomly) assigned to teachers irrespective of their desire to train, instead of via an application procedure. A second explanation for the higher deadweight loss found for the teacher vouchers could be that schools already had regular yearly budgets for training and schooling of their teaching personnel that exceed training budgets in the two voucher experiments. The yearly budgets of schools amount to over 1 percent of the total wage costs. Moreover, participation in schooling was already subject to tax deduction in the Netherlands for all employees including teachers. A third explanation could be that our voucher scheme is targeted at high educated consistently found that high educated workers more often participate in professional development activities than lower educated workers. This may lower the potential for policy initiatives to increase participation in professional development activities among higher educated workers.

The right hand side of Table 3.7 indicates that the deadweight loss seems somewhat smaller for vouchers that have been assigned to teachers who had not started at the time of application, that is, between 74 and 87 percent. This corresponds to larger than average positive enrollment effects for this subgroup of non-starters, which will be shown in section 3.6 where we discuss heterogeneous effects. This is a relevant finding since the possibility of applying for a study that has been started at the time of application has been abolished in 2011.

3.5.2 Effects on higher education completion

Figure 3.6 shows actual higher education completion rates by day of application and fitted lines on either side of the cutoff again using a quadratic fit. In accordance with the figure on enrollment shares we observe a small drop in completion shares among applicants after the cut-off date.

Figure 3.6 Proportion having completed higher education during 2008-2013 by day of application

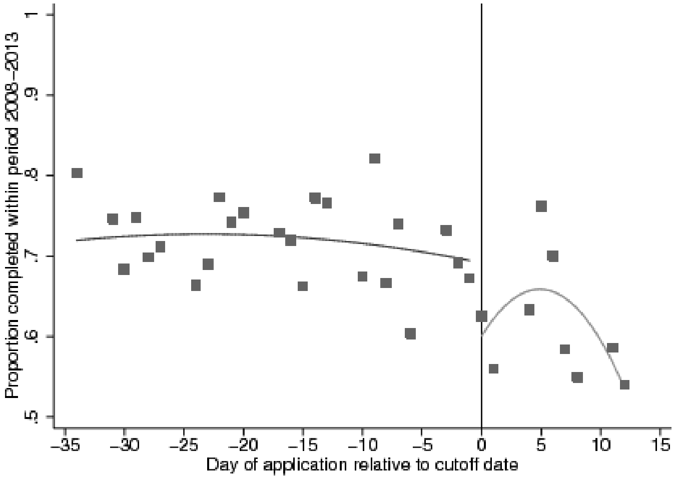


Table 3.8 shows the results from simple OLS estimates of the effect of voucher receipt on higher education completion. The OLS estimates with the full set of controls suggest that

voucher assignment raises the probability of completing a higher education study by a little over 20 percentage points. OLS point estimates on higher education completion are a couple percentage points larger than those on higher education enrollment.

Table 3.8 OLS estimates of effect of voucher assignment on probability of higher education completion in period 2008-2013

	Bandwidth (days around the cut-off date)			
Specification	7	14	21	all
(1) No controls	0.122***	0.162***	0.185***	0.191***
	(0.0280)	(0.0207)	(0.0203)	(0.0190)
(2) Adding applicant and application characteristics	0.191***	0.218***	0.235***	0.244***
	(0.0274)	(0.0207)	(0.0195)	(0.0183)
(3) Adding day of application and interaction term with cutoff	0.227***	0.217***	0.235***	0.243***
	(0.0316)	(0.0249)	(0.0237)	(0.0220)
Preferred order of polynomial of day of application and interaction term with after cutoff dummy	1	2	2	2
Control group mean	0.58	0.54	0.54	0.54
N	1,435	2,468	3,064	4,220

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The preferred order of the polynomial of day of application and its interaction term with the dummy indicating whether the application was done after the cut-off date is chosen using Akaike's information criterion. Preferred estimates are presented in bold. Applicant controls are sex, age category (5 categories), sector of work (three categories), baseline gross monthly salary, appointment size and the region of work (inside or outside Randstad region). Application controls are a dummy indicating whether the applicant had already started and program duration (four categories).

Table 3.9 shows the IV estimates that attempt to address the issue of selection on unobservables. Preferred impact estimates vary between 9 and 23 percentage points higher completion rates due to voucher assignment.[36] These estimates are in the same order of magnitude as the estimates on higher education enrollment. In relative terms effect estimates on completion are larger though, since baseline completion (54 percent) is lower than baseline enrollment (77 percent). The effect estimates point to a 17 to 42 percent increase in completion due to voucher assignment. The precision of the IV completion effect estimates is somewhat lower than of the IV enrollment effect estimates. Our IV point estimates on completion are pretty much in line with our OLS point estimates on completion for the same bandwidths.

[36] If we look at figure 3.6, we observe relatively high completion shares at days five and six after the cutoff date. One might think that our estimation results are driven by these two 'outliers' in combination with a flexible second order polynomial specification for the day of application. However, the preferred effect estimate for the 7 days bandwidth does not involve a polynomial. Moreover, zero order polynomial estimates for the other bandwidth groups are in line with the preferred second order polynomial specification for these bandwidths.

Table 3.9 IV estimates of effect of voucher assignment on probability of higher education completion in period 2008-2013

	Bandwidth (days around the cut-off date)			
Specification	7	14	21	all
Polynomial of day of application and interaction term with cut-off date of order:				
Zero	**0.0919***	0.208***	0.223***	0.241***
	(0.0536)	(0.0384)	(0.0362)	(0.0337)
One	0.174	0.0203	0.0421	0.0468
	(0.124)	(0.0934)	(0.0832)	(0.0783)
Two	0.279	**0.185**	**0.201***	**0.226****
	(0.205)	**(0.129)**	**(0.121)**	**(0.107)**
Preferred order of polynomial of day of application and interaction term with after cutoff dummy	0	2	2	2
Applicant and application controls	Y	Y	Y	Y
N	1,435	2,468	3,064	4,220

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The preferred order of the polynomial of day of application and its interaction term with the dummy indicating whether the application was done after the cut-off date is chosen using Akaike's information criterion. Preferred estimates are presented in bold. Applicant controls are sex, age category (5 categories), sector of work, baseline gross monthly salary, appointment size and the region of work (inside or outside Randstad region). Application controls are a dummy indicating whether the applicant had already started or not and program duration (four categories).

3.5.3 Effects on retention in education

Figure 3.7 shows shares of applicants still working in education four years after (first) application by day of application in the first application period and fitted lines on either side of the cutoff again using a quadratic fit. This figure suggests a small drop in stay rates after the cut-off date.[37]

[37] Retention in education is our preferred outcome of interest here. We could also look at retention in the teaching profession. When we do this, we observe an average difference in retention before and after the cut-off that are slightly smaller (i.e. half a percentage point) than the difference in retention in education.

Figure 3.7 **Probability of still working in education in 2012 by day of application**

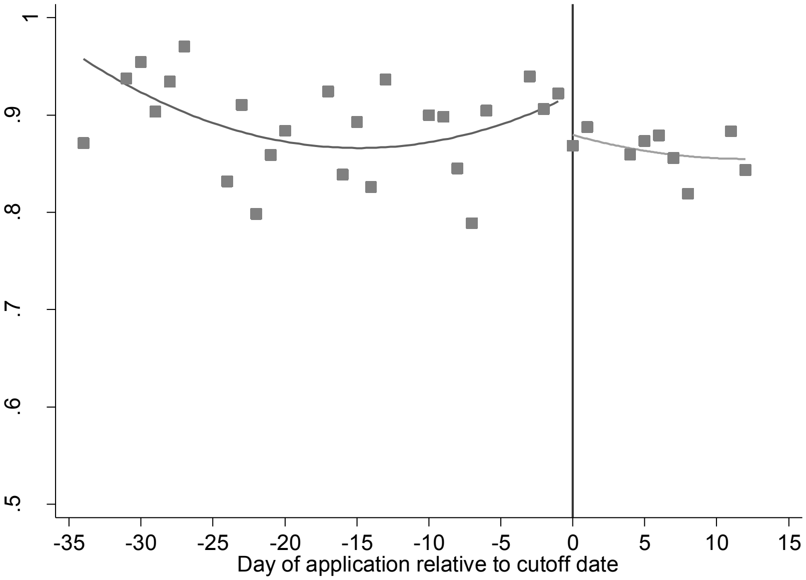


Table 3.10 shows results from simple OLS estimates of the effect of voucher assignment on the probability of still working in education. The OLS estimates with the full set of controls suggest a small positive effect on the probability of staying in education of around 3-5 percentage points.

Table 3.10 **OLS estimates of effect of voucher assignment on probability of still working in education in 2012 (four years after voucher application)**

	Bandwidth (days around the cut-off date)			
Specification	7	14	21	all
Effect of voucher assignment	0.039*	0.047***	0.035**	0.027*
	(0.0235)	(0.0157)	(0.0168)	(0.0155)
Preferred order of polynomial of day of application and interaction term with after cutoff dummy	2	0	0	2
Applicant and application controls	Y	Y	Y	Y
Control group mean	0.85	0.84	0.85	0.85
N	1,365	2,332	2,906	4,017

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The preferred order of the polynomial of day of application and its interaction term with the dummy indicating whether the application was done after the cut-off date is chosen using Akaike's information criterion. Preferred estimates are presented in bold. Applicant controls are sex, age category (5 categories), sector of work (three categories), baseline gross monthly salary, appointment size and the region of work (inside or outside Randstad region). Application controls are a dummy indicating whether the applicant had already started and program duration (four categories).

Table 3.11 shows the IV estimates. Preferred estimates on the basis of the Akaike Information Criterion for model specification range from minus 9 to plus 11 percentage points impact on the probability of still working in education. The negative estimate for the 7 days bandwidth sample with a second order polynomial control for the day of application may well point at an over-specified model for this small bandwidth. The other three preferred estimates range between plus 4 and plus 11 percentage points. These estimates generally lack precision however.

Table 3.11 IV estimates of effect of voucher assignment on probability of still working in education in 2012 (four years after voucher application)

	Bandwidth (days around the cut-off date)			
Specification	7	14	21	all
Polynomial of day of application and interaction term with cut-off date of order:				
Zero	0.045	**0.054***	**0.042**	0.055**
	(0.0393)	**(0.0298)**	**(0.0282)**	(0.0260)
One	0.222**	0.105	0.0774	0.0115
	(0.0925)	(0.0649)	(0.0587)	(0.0548)
Two	**-0.087**	0.180*	0.134	**0.112**
	(0.141)	(0.0972)	(0.0888)	**(0.0781)**
Preferred order of polynomial of day of application and interaction term with cut-off date	2	0	0	2
Applicant and application controls	Y	Y	Y	Y
N	1365	2332	2906	4017

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The preferred order of the polynomial of day of application and its interaction term with the dummy indicating whether the application was done after the cut-off date is chosen using Akaike's information criterion. Applicant controls are sex, age category (5 categories), sector of work, baseline gross monthly salary, appointment size and the region of work (inside or outside Randstad region). Application controls are a dummy indicating whether the applicant had already started or not and program duration (four categories).

All in all, our analyses suggest a small positive effect of voucher assignment on the probability of still working in education four years after first applying for a voucher. This small positive effect may partially occur because of the voucher scheme requirement to keep working in education at least one year after completing the course. A voucher recipient has to pay back the subsidy if he or she leaves education within one year after completion.

3.6 Heterogeneous treatment effects

Table 3.12 shows the effect estimates for various sub-samples by applicant (Panel A) and application characteristics (Panel B).[38] We only focus on the OLS results here due to the reduction in precision in IV estimates when we carry out analyses on subgroups. This approach is in line with Jacob & Lefgren (2011) and Schwerdt et al. (2012), among others.[39]

Panel A shows little differences in effect estimates of voucher assignment on higher education enrollment and completion by sex and sector of work. Effects on completion probabilities (but not on enrollment) seem to increase somewhat by the size of a teacher's appointment in FTE, with point estimates being almost 10 percentage points higher for teachers with an appointment of more than 0.8 FTE as compared to teachers working less than 0.5 FTE. Effects of voucher assignment on both enrollment and assignment seem somewhat smaller for younger teachers (i.e. 15-34 years) than for older teachers (35-64 years), with point estimates being 5-8 percentage points higher for the older group.

It should be noted that completion probabilities gradually decline with age. Teachers in the category 55-64 have a completion probability that is 30 percentage points lower than that of teachers aged 15-24, keeping voucher status and all other applicant and application characteristics constant. Completion probabilities also differ by sex (5 percentage points higher for female teachers), assignment size (about 5 percentage points lower for teachers working more than 0.8 FTE as compared to less than 0.8 FTE) and sector of work (about 8 percentage points lower in secondary education as compared to primary education), again keeping voucher status and all other characteristics constant.[40]

In terms of application characteristics we find some interesting differences in effect estimates as well (see Panel B). First, voucher assignment seems to have relatively small effects (if any) for teachers applying for short higher education studies, that is, with program duration of one year or less. Voucher assignment seems to have largest effects for studies with program duration longer than one year but at most two years, particularly on completion probabilities.

[38] Descriptive statistics of treatment and outcome variables for the various subgroups are presented in Appendix Tables D2 and D3.

[39] Under the assumption that any remaining omitted variable bias in the OLS models does not differ across the sub-populations, the more precise OLS models are informative about the relative size of the effects (Schwerdt et al., 2012).

[40] These differences are statistically significant and robust to different estimations on different bandwidth samples and with different specifications (OLS or IV). These differences are identified in regressions controlling for all applicant and application characteristics at the same time.

Effect estimates on both enrollment and completion probabilities for studies with a program duration of more than two years are somewhere in the middle.

The completion probability declines significantly by program duration. Controlling for all other applicant and application characteristics and voucher assignment, the completion probability is about 20 percentage points lower for studies with duration longer than two years as compared to one year or less.

Another noticeable finding is that enrollment and completion effects of voucher assignment are smaller for teachers applying for a voucher for a study that was already started at the time of application. This is particularly the case for effects on higher education enrollment with the OLS effect estimate being more than twice as big for the group of applicants that did not start their study yet as compared to that for the ones that had already started at the time of application. OLS effect estimates on completion probabilities are about five percentage points higher than average. Appendix Table C2 shows IV effect estimates on higher education enrollment (panel A) and higher education completion (panel B) for the subgroup of applicants that did not start at the time of application. Preferred estimates are four to six percentage points higher than for the total group of voucher applicants including those that had already started at the time of application. These differences in effect estimates by starting status suggest that the deadweight loss of the voucher scheme has been reduced by the abolishment in 2012 of the possibility to apply for a voucher for a study that has already been started at the time of application. Our IV estimates suggest that this may have increased effects of voucher assignment on enrollment to 13-27 (from 9-22) percentage points and on completion to 13-28 (from 9-23) percentage points. Appendix Figures C1 and C2 show visually that differences in both enrollment and completion around the cutoff are marked for the subpopulation of non-starters, but do hardly exist for the subpopulation of starters at the time of application.

Finally we turn to heterogeneous effects of voucher assignment on the probability of still working in education four years after (first) applying for a voucher, as presented in the last two columns of Table 3.12. Regarding applicant characteristics larger than average effect estimates are found for male teachers, teachers working in secondary education, teachers with an appointment of 0.5 to 0.8 FTE, and teachers of 35 years and older at the time of application. Regarding application characteristics, our results suggest larger than average effects on stay rates for teachers who had not started the study yet at the time of application, and for applications for studies lasting between more than one and two years. These patterns of heterogeneous effects on stay rates for subgroups by application characteristics are pretty much in line with patterns found for effects on higher education enrollment and completion.

Table 3.12 Heterogeneity of effects of voucher assignment on probability of higher education enrollment and completion and on working in education four years after voucher application

	Higher education enrollment		Higher education completion		Still working in education in 2012	
Effect on subgroup	+/- 14	all	+/- 14	all	+/- 14	all
Baseline	0.160***	0.211***	0.217***	0.243***	0.047***	0.027*
	(0.0165)	(0.0182)	(0.0249)	(0.0218)	(0.0157)	(0.0155)
Panel A: Applicant characteristics						
Female teachers	0.162***	0.221***	0.206***	0.247***	0.039**	0.024
	(0.0188)	(0.0212)	(0.0290)	(0.0256)	(0.0192)	(0.0186)
Male teachers	0.159***	0.199***	0.235***	0.227***	0.064*	0.031
	(0.0344)	(0.0357)	(0.0495)	(0.0426)	(0.0358)	(0.0328)
Working in primary education	0.133***	0.191***	0.218***	0.252***	0.027	0.010
	(0.0242)	(0.0273)	(0.0376)	(0.0333)	(0.0225)	(0.0235)
Working in secondary education	0.188***	0.226***	0.218***	0.227***	0.080***	0.057**
	(0.0286)	(0.0290)	(0.0399)	(0.0339)	(0.0303)	(0.0261)
Working in special education	0.188***	0.251***	0.221***	0.278***	0.023	0.025
	(0.0376)	(0.0472)	(0.0634)	(0.0562)	(0.0434)	(0.0459)
Appointment <= 0.5 FTE	0.165***	0.203***	0.138	0.167**	-0.003	-0.047
	(0.0528)	(0.0549)	(0.0929)	(0.0766)	(0.0754)	(0.0677)
Appointment > 0.5 & <= 0.8 FTE	0.165***	0.198***	0.194***	0.243***	0.085**	0.055
	(0.0326)	(0.0396)	(0.0522)	(0.0489)	(0.0335)	(0.0372)
Appointment > 0.8 FTE	0.159***	0.218***	0.231***	0.250***	0.046**	0.027
	(0.0205)	(0.0223)	(0.0304)	(0.0260)	(0.0198)	(0.0180)
Age 15-34	0.119***	0.180***	0.176***	0.222***	0.019	0.016
	(0.0228)	(0.0267)	(0.0377)	(0.0331)	(0.0222)	(0.0243)
Age 35-64	0.197***	0.236***	0.255***	0.261***	0.061***	0.040*
	(0.0235)	(0.0248)	(0.0336)	(0.0291)	(0.0213)	(0.047)
Panel B: Application characteristics						
Already started study at time of application	0.081***	0.112***	0.172***	0.204***	0.033	0.009
	(0.0215)	(0.0268)	(0.0498)	(0.0397)	(0.0348)	(0.0311)
Did not start study yet at time of application	0.184***	0.254***	0.221***	0.256***	0.053***	0.036*
	(0.0208)	(0.0232)	(0.0291)	(0.0263)	(0.0197)	(0.0195)
Program duration 0-1 years	0.058**	0.119***	-0.020	0.067	0.049	0.013
	(0.0284)	(0.0462)	(0.0621)	(0.0588)	(0.0315)	(0.0377)
Program duration >1-2 years	0.199***	0.242***	0.315***	0.335***	0.083***	0.059**
	(0.0281)	(0.0302)	(0.0392)	(0.0341)	(0.0270)	(0.0256)
Program duration >2 years	0.166***	0.216***	0.196***	0.206***	0.015	0.004
	(0.0273)	(0.0277)	(0.0388)	(0.0329)	(0.0271)	(0.0263)

Notes: These estimates are based on OLS regressions similar to those in row 3 of Tables 3.4, 3.7 and 3.9. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

3.7 Characterizing the compliers

Regression discontinuity estimates reported in this chapter should be considered as the effect of voucher assignment on the population of so-called compliers. Compliers are teachers who take up a voucher when they apply before the cut-off date, but do not when they apply after the cut-off date. It is not possible to directly distinguish compliers from always-takers (i.e. being assigned a voucher irrespective of the timing of application) and never-takers (i.e. never being assigned a voucher). Angrist and Pischke (2009) however show that it is possible to characterize the complier population by making use of the variation in the first-stage estimates across subgroups. The relative probability that a complier has a certain characteristic is given by the ratio of the first stage estimate for the particular subgroup with that characteristic to the overall first stage estimate. This knowledge about compliers may be important for policy-makers as it shows which groups are either more or less affected in terms of voucher take-up by a budget restriction.

Table 3.13 shows the first-stage ratios for various subgroups by applicant and application characteristics. In terms of applicant characteristics the probability a complier has a certain characteristic is higher than average for males (ratio of 1.21), younger teachers (1.12), teachers in primary (1.09) and special education (1.10) and teachers working more than 0.8 FTE (1.12). Compliance in terms of voucher receipt with the cut-off date is somewhat lower than average for females (0.92), teacher of 35 years and older (0.92), teachers working in secondary education (0.84) and teachers working less hours (appointments smaller than 0.8 FTE).

In terms of application characteristics teachers who have already started at the time of application are more likely to be compliers (1.13). Compliance is particularly larger among applicants for a study with duration of a year or less (1.50), whereas it is lower than average for the group with a program duration longer than two years (0.77). This lower compliance among applicants for longer studies may indicate that schools and/or teachers were less able or willing to finance these longer (and arguably more expensive) studies by means other than the voucher and were more likely to wait for the next application round to obtain a voucher if they were too late to obtain one in the first application period in 2008.

Table 3.13 Characterizing compliers

	First stage estimate	Ratio to overall first-stage	N
All	-0.400*** (0.029)	1.00	4330
Applicant characteristics			
Sex			
Female	-0.368*** (0.033)	0.92	3234
Male	-0.485*** (0.054)	1.21	1093
Age (years)			
15-34	-0.446*** (0.046)	1.12	1948
35-64	-0.368*** (0.037)	0.92	2381
Sector of employment			
Primary education	-0.434*** (0.040)	1.09	1952
Secondary education	-0.335*** (0.048)	0.84	1819
Special education	-0.441*** (0.073)	1.10	559
Appointment in FTE in 2008			
0-0.5 FTE	-0.350*** (0.103)	0.88	334
>0.5-0.8 FTE	-0.331*** (0.055)	0.83	1017
>0.8 FTE	-0.440*** (0.036)	1.12	2866
Application characteristics			
Status of planned study at time of application			
Already started	-0.453*** (0.066)	1.13	953
Did not start	-0.388*** (0.031)	0.97	3376
Program duration of planned study (in years)			
0-1 year	-0.601*** (0.060)	1.50	661
>1-2 years	-0.379*** (0.044)	0.95	1903
>2 years	-0.307*** (0.047)	0.77	1758

*** p<0.01. Robust standard errors in parentheses. First-stage estimates are estimated by regressions using the full set of application and applicant controls and a linear control for day of application and a linear interaction term of day of application with a dummy indicating whether the application was done after the cut-off date. The ratio in the last column indicates the relative probability compliers have the particular applicant or application characteristic indicated in each row.

3.8 Crowding out of other types of funding

We have observed that a considerable share of the teachers who did not receive a voucher was still enrolled in higher education studies and managed to complete these studies. This suggests that the voucher substitutes for other sources of funding of these studies. Table 3.14 gives an indication of what type of funding the voucher substitutes for. Data are from questionnaires among voucher applicants. These data suggest that the voucher substitutes for school funding and for funding by own means of the teacher. Indications that the voucher substitutes for school funding may not be surprising since schools have yearly budgets reserved for professional development of their personnel.[41] Co-funding by both schools and teachers happens as well. The shares of these funding means are almost equal on average. Program duration seems to matter: longer studies are more often completely financed by teachers whereas shorter studies are more often completely financed by schools.

Table 3.14 Funding means of studies by teachers who applied for but did not receive a teacher voucher

	<= 1 year	1-2 years	> 2 years	All studies
Share financed by				
school	0.71	0.48	0.16	0.38
school and teacher	0.05	0.28	0.29	0.24
teacher	0.10	0.33	0.55	0.39
other means	0.14	0.02	0	0.03
Number of respondents	21	54	56	131

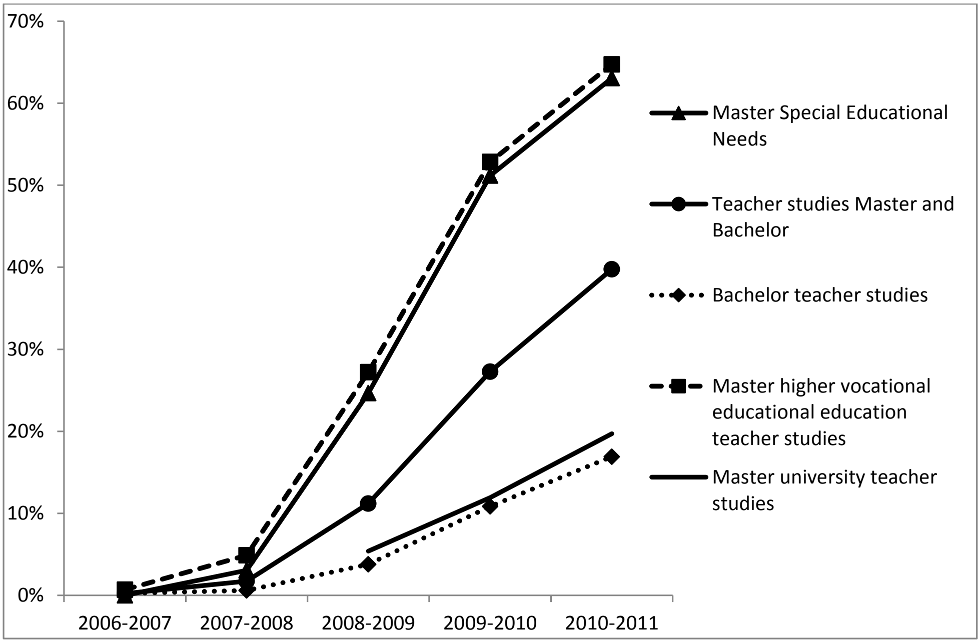
Source: own calculations based on questionnaire data among voucher applicants by IVA Onderwijs together with CPB Netherlands Bureau for Economic Policy Analysis that was carried out in 2009.

Figure 3.8 shows the evolution of the share of enrolled teachers who is financed with a voucher in total enrollment by teachers for a couple of main categories of teacher studies. The figure shows a strong increase in enrollment by teachers with a voucher at the expense of enrollment by teachers without a voucher since the introduction of the voucher scheme. This is particularly the case for the master Special Educational Needs (i.e. the single program most applied for) and for teacher studies at master level at higher vocational education institutes, where shares of enrollment with a teacher voucher have steadily increased from zero to about 65 percent in three years time since the introduction of the teacher voucher scheme.

Further indications of crowding out come from questionnaires among teachers. Thirty percent of surveyed teachers in 2011 (three years after the introduction of the teacher voucher) agrees that the teacher voucher scheme has led to less available money at their schools for individual professional development, which is up from 19 percent of teachers in 2009 (Vink et al., 2012).

[41] Expenses of schools on continuing education of their personnel are about 1.4 percent in primary education and 1.1 percent in secondary education of the total wage costs, as reported by school directors (see Vink, 2012). Thirty percent of teachers in 2011 agree that the teacher voucher has led to less school means for individual continuing education. This share is increasing since 2009 (a year after the introduction of the teacher voucher scheme), when 19 percent of teachers agreed that school budgets for continuing education had been reduced due to the teacher voucher.

Figure 3.8 Share of enrollment by teachers in teacher studies that is financed with a teacher voucher (source: own calculations based on tables in Vink et al., 2012)



3.9 Conclusions and discussion

In this chapter we have exploited a discontinuity in the probability of receiving a teacher voucher that was caused by budget restrictions to detect effects of teacher voucher assignment. We find positive effects of voucher assignment in the order of 10-20 percentage points on the probability of both higher education enrollment and completion among teachers. Higher education enrollment among voucher applicants that never received a voucher is 77 percent, whereas higher education completion among this group equals 54 percent, both measured over a period of five years since voucher application. This lower completion base rate implies that relative effects of voucher assignment on completion are larger than on enrollment. The deadweight loss of the voucher scheme in terms of enrollment in higher education degree programs is estimated at about 80 to 90 percent.

The teacher voucher scheme appears to crowd out both school-financed and teacher-financed participation in continuing education by teachers. This phenomenon of substantial crowding out of other sources of financing by public vouchers has also been found in earlier studies on training vouchers for workers (e.g. Schwerdt et al., 2012; Hidalgo et al., 2014). Vouchers for relatively short studies (i.e. up to one year) more often appears to crowd out funding by schools out of regular budgets, whereas vouchers for longer studies (more than two years) more often appears to crowd out funding by teachers themselves. Regular yearly school budgets for professional development of teachers amount to a little over 1 percent of the total wage costs.

We find heterogeneous treatment effects of voucher assignment by program duration and starting status at the time of application. Our results suggest largest effects of applications for studies with duration of between one and two years and for applications for studies that have not been started yet at the time of application. The voucher seems to trigger least additional enrollment and completion for studies with a duration of one year or less. The possibility to apply for a voucher for a study that has been started yet at the time of application has been abolished in 2011. Our estimates suggest that this may have led to an increase in the effects of voucher assignment on both higher education enrollment and completion by about five percentage points.[42] The deadweight loss is expected to have fallen to the same extent. Our estimation results also suggest that effects on enrollment and completion are somewhat smaller for teachers aged under 35 than for older teachers. These effects do not seem to differ strongly by sex, size of appointment and education sector.

Teachers who were more reliant on voucher funding may have been more likely to have reapplied for a voucher in later application periods when they did not receive a voucher in the first application period. An indication of this is the larger reapplication probability for longer studies. Arguably more costs are involved in these longer studies, that is, both in terms of direct study costs and the costs for schools for arranging replacement while teachers are on study leave. The enrollment and completion experience of these re-applicants that received a voucher in later rounds is not reflected in our estimated local average treatment effects. The possibly larger reliance of re-applicants (and their schools) on voucher funding would imply

[42] There may well be some other differences in teacher and application characteristics and in contextual factors (e.g. the financial position of the schools, the need for certified teachers, the promotion opportunities for teachers) between the first application period and later application periods that may cause differences in voucher effects in more recent application periods. It is difficult to investigate these effects in more recent application periods in the same manner since there was no discontinuity anymore in voucher assignment that we could exploit in later application periods. Survey results among voucher recipients of subsequent application periods that are not yet available may shed some light on the evolution of voucher effects over time. These indications would be based on stated preferences rather than revealed preferences however.

that our *local* average treatment effect estimates are somewhat lower than the average treatment effects of voucher assignment on enrollment and completion.

Our results suggest small positive effects of voucher assignment on the probability to stay working in education as measured four years after voucher application. These effects seem concentrated among teachers working in secondary education, teachers aged 35-64 years, teachers who did not start their study yet at the time of application, and teachers applying for a study with a duration of between one and two years. Positive effects on teacher retention would be a positive side effect of the teacher voucher scheme since recent evidence shows that teacher value added improves with experience. It would be interesting to monitor whether these small positive retention effects persist over a longer-term.

The teacher voucher instrument could have had effects on other policy relevant outcomes that have not been studied here. One may think of effects on alleviating shortages of certified teachers in certain subjects or regions, on the attractiveness of teaching as a profession and on the professional culture in schools. Another interesting question for further research would be if and to what extent voucher utilization for participation in higher education degree programs crowds out participation in other professional development activities.[43]

[43] A first indication of the occurrence of some effects is that fifteen percent of surveyed voucher applicants state that they are not allowed anymore to participate in other continuing education activities due to the teacher voucher. Twenty percent of teachers state that they didn’t have time anymore to participate in training activities related to maintenance of their teacher competences (Vink et al., 2012).

Appendix A Teacher voucher scheme versus two other adult training voucher schemes

Table A.1 Comparison of teacher voucher with two other training voucher schemes

	Schwerdt et al. (2012)	Hidalgo et al. (2014)	Van der Steeg en Van Elk (2014)
Country	Switserland	The Netherlands	The Netherlands
Year	2006	2006	2008
Type of education/training	All kinds of study courses / training sessions	All kinds of study courses / training sessions	Degree programs (Bachelor or Master)
Eligibility	Employed and unemployed with varying educational attainment	Employees in four sectors in the Netherlands; mainly low educated workers	Employed teachers; high educated
Voucher value	250/750/1500 Swiss Francs	1000 euro	Max 4200 euro per year for teachers (a) and max 6700 euro per year for employers to arrange replacement.
Redemption period	Within six Months	Within two years	Within one year
Redemption rate	18 percent	41 percent	95 percent
Type of data	Labour Force Survey panel data	Survey data collected particularly for evaluation	Administrative data
Sample size	10,521	1,266	4,220
Empirical approach	Randomized experiment	Randomized experiment	Fuzzy RD
Effects investigated over period of	One year	Two years	Five years after first voucher application
Effect size of voucher receipt on participation / enrollment	+13 percentage points	+20 percentage points	+9 to +22 percentage points (b)
Control group mean participation	33 percent	45 percent	77 percent
Relative effect (= effect size in percentage points / control group mean)	+39 percent	+44 percent	Between +12 and +29 percent on average (b)
Deadweight loss	30 percent	59 percent	Between 78 and 91 percent (b)
Other noticeable findings	- Smaller effects for vouchers with lowest face value - Significant crowd-out of firm-financed education	- Positive impact on future training plans - No effects on job mobility	- Larger effects for applications for studies that had not been started yet at the time of application - Smallest effects for applications for short term studies (a year or less)

(a) This maximum amount has been raised to 7700 euro per year from 2011 onwards. (b) Depending on bandwidth and functional form. Deadweight loss is estimated at 74 to 87 percent for studies that had not been started yet when applying.

Appendix B Facts and figures of the teacher voucher scheme, teacher professional development in the Netherlands and policy measures and goals

The teacher voucher scheme

Table B.1 shows the evolution of the number of requested and assigned teacher vouchers over the first seven application periods. These numbers include applications for generally brief courses and applications for acknowledged bachelor or master degree programs. As from 2012 onwards, vouchers could only be requested for bachelor or master degree programs.

Table B.1 Number of requested and assigned teacher vouchers and amount of money involved

Appli-cation period	Year	Number of applications	Number of vouchers assigned	Share of applicants assigned a voucher	Subsidy assigned to teachers (million euro)	Subsidy assigned to schools (million euro)	Subsidy assigned to teachers and schools (million euro)	Average total subsidy per assigned voucher (euro)
1	2008	7,501	4,866	65%	14.2	16.6	30.8	6,324
2	2009	4,128	3,497	85%	10.6	14.4	25.0	7,135
3	2009	5,679	5,169	91%	17.5	24.4	41.9	8,114
4	2010	8,304	7,087	85%	26.6	36.6	63.2	8,918
5 (a)	2011	8,747	8,227	94%	36.0	38.6	74.5	9,061
6 (b)	2012	5,221	4,722	90%	29.0	40.6	69.6	14,739
7 (c)	2013	6,916	6,188	87%	40.4	48.7	89.1	14,399
Total	**2008-13**	**46,496**	**39,609**	**85%**	**174.2**	**219.9**	**394.1**	**9,950**

(a) The maximum yearly subsidy that could be assigned to teachers has been raised to 7700 euro in 2011. Between 2008 and 2010 it was 4200 euro.

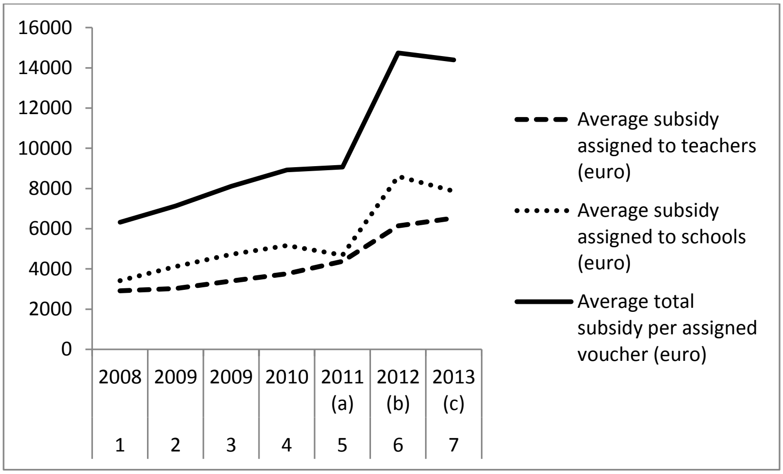
(b)The possibility to apply for brief courses (i.e. being not bachelor or master degree programs) has been abolished in 2012.

(c) Vouchers have been assigned for one year only since 2013. If a study program lasts for more than a year, the teacher has to reapply for a voucher in the next year(s). The figures shown for 2013 are predicted figures on the basis of the ratio of assigned subsidies for the total study period relative to those for the first study year, taken from the preceding year 2012.

Source: own calculations on figures provided by *Dienst Uitvoering Onderwijs (DUO)*.

Almost 40 thousand teacher vouchers have been assigned over the period 2008-13 in seven application periods. Almost 400 million euro of subsidies is involved with these vouchers, of which 174 million goes to teachers as compensation for study fees and costs of study materials and travel costs, and 220 million goes to schools to give them the opportunity to provide study leave and arrange a replacement teacher. The average total subsidy per assigned voucher has more than doubled over time, that is, from 6.3k euro in 2008 to 14.7k euro in 2012. This is due to a number of factors. First, the maximum yearly subsidy for teachers has been raised from 4,200 to 7,700 euro in 2011. Second, vouchers could only be assigned for registered bachelor or master degree programs as from 2012 onwards. Vouchers could not be assigned for other brief courses or training programs anymore. This has raised the share of applications for bachelor or master degree programs in one year by 40 percentage points. These bachelor and master degree programs are often more expensive in terms of total study fees because of longer study duration. Moreover, compensation to schools for study leave is only possible for applications for bachelor or master degree programs. Figure B1 shows the evolution of average total subsidy costs per voucher and the average subsidy provided to schools and to teachers.

Figure B.1 Evolution of average costs per voucher



See notes for years 2011, 2012 and 2013 under table B1.

The abolishment of the opportunity to apply for a voucher for other courses or training programs than bachelor or master degree programs has contributed to the strong decline in the total number of applications in 2012. The extension of the voucher eligibility to teachers with a flexible contract or replacement teachers in 2013 has contributed to the increase in the number of applications to a small extent, according to figures provided to us by DUO.

On average 2.9 percent of all teachers in the eligible education sectors have applied yearly for a teacher voucher over the period 2008-2013. On average 2.0 percent of all teachers have applied yearly for a voucher for a bachelor or master degree program, which amounts to nearly 70 percent of all applications. The share of teachers applying yearly for a bachelor or master degree program ranges from 1.6 percent in intermediate post-secondary vocational education (MBO) to 2.5 percent in secondary education. Shares in primary education (1.8 percent) and special education (2.2 percent) are in between. Nearly one third of all applications have been for Master Special Educational Needs, a degree program in which teachers learn to cope better with pupils with special educational needs. This share is largest in special education (56 percent) and primary education (42 percent). In secondary education and intermediate post-secondary vocational education relatively larger shares of applications have been for subject-specific degree programs.

Teacher professional development in the Netherlands: concerns and figures

Raising teacher quality is high on the policy agenda in the Netherlands. This stems from concerns about teacher quality that have been expressed by policymakers, The Inspectorate of Education, school leaders and teachers themselves. Results from the TALIS survey among teachers in 32 countries show that a large share of over 70 percent of teachers in the Netherlands thinks that good education is hindered by a shortage of qualified and/or good performing teachers (OECD, 2014). PISA (2012) figures show that the Netherlands have the highest share of uncertified teachers in lower secondary education of all OECD countries (Kordes et al., 2013). Berndsen et al. (2013) show that on average 17 percent of all lessons in secondary education in 2011 were given by teachers who are not certified for the subject (or not at the required level) with even larger shares in certain shortage subjects and in the more urbanized regions. The Dutch Inspectorate of Education has found that two-thirds to three-quarters of all teachers does not succeeds in differentiating their lessons according to differences in level and speed of their pupils in secondary education, intermediate post-secondary vocational education and in special education. In primary education this share is between 40 and 50 percent (Inspectorate of Education, 2014).

TALIS survey results show that though the degree of participation in professional development activities among Dutch teachers is somewhat larger than average, the intensity of these activities in terms of number of days involved is lower than average (OECD, 2014). Participation of teachers in qualification programs (e.g. a degree program) is relatively low compared to participation in brief courses or workshops in the countries participating in the TALIS survey (18 versus 71 percent of teachers in the last twelve months). This contrasts with the opinion of teachers that these more intensive and longer professional development activities are the more effective professional development activities (Inspectorate of

Education, 2012). PISA 2012 figures show that professionalization activities of Dutch teachers in math stay behind those of teachers in other OECD countries, particular among math teachers (Kordes et al, 2013).

Both TALIS and a large Dutch survey among teachers offer insights into impeding factors for participation in professional development activities. TALIS finds that the most mentioned impediments for participation are in descending order that there is no relevant professional development offered (39 percent), that professional development conflicts with the work schedule (38 percent), that there are no incentives to participate in such programs (30 percent), a lack of employer support (27 percent), that professional development is too expensive/unaffordable (26 percent), and that they do not have the prerequisites (8 percent). Results from a large Dutch teacher survey show that about sixty to seventy percent of Dutch teachers state that their professional development is seriously hindered because they are too busy with their daily work (Berndsen et al., 2014). Other less frequently mentioned limiting factors mentioned by teachers in this survey are that professional development is impeded by their work schedule (34-46 percent), no time because of family affairs (16-32 percent), that the employer does not give enough support (17-26 percent), that it is too expensive (15-25 percent), and that it is not stimulated by their managers (14-23 percent).

The Dutch Inspectorate of Education mentions that in Dutch primary education professionalization activities are often team activities. This causes a lack of tailored activities to the professionalization needs of individual teachers (Inspectorate of Education, 2012). In a more recent publication the Inspectorate concludes that the room for professionalization that teachers have is certainly not used by all teachers, particularly not by the weakest teachers. High work pressure experienced by teachers and limitations within the school organization to reserve time are most mentioned impediments for teacher professionalization (Inspectorate of Education, 2013). The Inspectorate also concludes in this publication that professionalization activities by teachers often have too little focus and are too often not targeted at specific goals to improve own teaching practices.

The share of teachers who had to pay for none of the professional development activities undertaken is above average of the TALIS countries (i.e. 78 versus 66 percent). This share is lower than in North-Western Europe however (Van der Boom and Stuivenberg, 2014). Non-monetary support for Dutch teachers in terms of for instance study leave is about average (13.5 versus 14.1 percent). Unfortunately data are lacking to compare the evolution of monetary and non-monetary support over time in an international perspective. The introduction of the teacher voucher in the Netherlands may have affected both types of support.

Policy measures and goals

The most recent policy program of the Ministry of Education is the Teachers Program 2013-2020. This program was developed in collaboration with teachers, principals, school boards and educators. The program has seven broad areas of attention, for which specific targets and policy measures have been formulated.[44] One of the specific goals formulated in this program is to raise the share of teachers with a master degree in 2020 to 50 percent in secondary education and to 30 percent in primary education. This is up from a current share of 37 percent and 20 percent, respectively.[45] Apart from this general master goal that makes no distinction between vocational and academic masters, there is a specific goal to raise the share of academic master teachers in upper secondary education from 60 to 80-85 percent by 2020. The number of uncertified teachers should gradually fall to zero by 2020, down from 17 percent in 2011.[46]

One of the important policy tools to achieve these goals of more master teachers and less uncertified teachers is the teacher voucher scheme. Budgets for the teacher voucher scheme have been raised every year since the start of the scheme. Other recently announced policy measures that may contribute to this goal is promoting alternative and more flexible routes to teaching for talented master educated young people. These measures are part of a policy package *Landelijke inpuls leraren tekortvakken* in which in total 100 million euro will be spent over the years 2013-2016 (Ministry of Education, 2013).

A larger policy package named *Actieplan Leerkracht van Nederland* targeted at raising teacher quality and quantity has been launched in 2007 (Ministry of Education, 2007). Over

[44] These areas of attention are: better students in teacher training programs, better teacher training programs, attractive and flexible development pathways, starting as a teacher, schools as learning organisations, all teachers skilled and qualified, and a strong professional organization.

[45] See http://www.trendsinbeeld.minocw.nl/vervolg.php?h_id=5&s_id=29&v_id=60&d_id=38&titel=Master/academici

[46] See http://www.trendsinbeeld.minocw.nl/vervolg.php?h_id=5&s_id=29&v_id=60&d_id=37&titel=Gekwalificeerde_leraren

eighty percent of this more than one billion euro package was directed towards improvement in teacher compensation. These salary measures had two major components. The first component was gradually providing extra money to schools to enable them to place a larger share of their teachers in higher pay scales. An underlying goal of this measure was to create more variety in teacher salaries, which should trigger teachers to keep on investing in their skills and careers in order to increase their chances of being promoted. The second component was a gradual reduction in the number of years in which a teacher reaches the maximum of his or her salary scale. The launch of the teacher voucher scheme was part of the policy package *Actieplan Leerkracht van Nederland* as well.

Appendix C Outcomes of regression analyses

Table C1 IV effect estimates on higher education enrollment and higher education completion for subgroup of applicants that had not started yet at the time of application

	Bandwidth (days around the cut-off date)			
Specification	7	14	21	all
Panel A: effect on higher education enrollment				
Polynomial of day of application and interaction term with cut-off date of order:				
Zero	**0.128*****	**0.169*****	0.156***	0.161***
	(0.0416)	**(0.0308)**	(0.0300)	(0.0282)
One	0.271***	0.160**	0.160**	0.136**
	(0.889)	(0.1347)	(0.0640)	(0.0630)
Two	0.366***	0.259*	**0.265*****	**0.251*****
	(0.1347)	(0.143)	**(0.0878)**	**(0.0816)**
Preferred order of the polynomial of day of application and interaction term with cut-off date	0	0	2	2
Applicant and application controls	Y	Y	Y	Y
Control group mean	0.78	0.77	0.74	0.73
Panel B: effect on higher education completion				
Polynomial of day of application and interaction term with cut-off date of order:				
Zero	**0.131****	0.247***	0.266***	0.161***
	(0.0648)	(0.0468)	(0.0442)	(0.0282)
One	0.247***	0.086	0.104	0.081
	(0.1347)	(0.1080)	(0.0962)	(0.0924)
Two	0.322	**0.259***	**0.253***	**0.282****
	(0.2306)	**(0.1403)**	**(0.1303)**	**(0.1174)**
Preferred order of the polynomial of day of application and interaction term with cut-off date	0	2	2	2
Applicant and application controls	Y	Y	Y	Y
Control group mean	0.51	0.48	0.47	0.46
N	1,141	1,934	2,396	3,289

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The preferred order of the polynomial of day of application and its interaction term with the dummy indicating whether the application was done after the cut-off date is chosen using Akaike's information criterion. Applicant controls are sex, age category (5 categories), sector of work (three categories), baseline gross monthly salary, appointment size and the region of work (inside or outside Randstad region). Application controls are a dummy indicating whether the applicant had already started and program duration (four categories).

Figure C1 Proportion ever having been enrolled (left panel) and ever having completed (right panel) higher education during 2008-2013 by day of application, for subgroup of non-starters at time of application

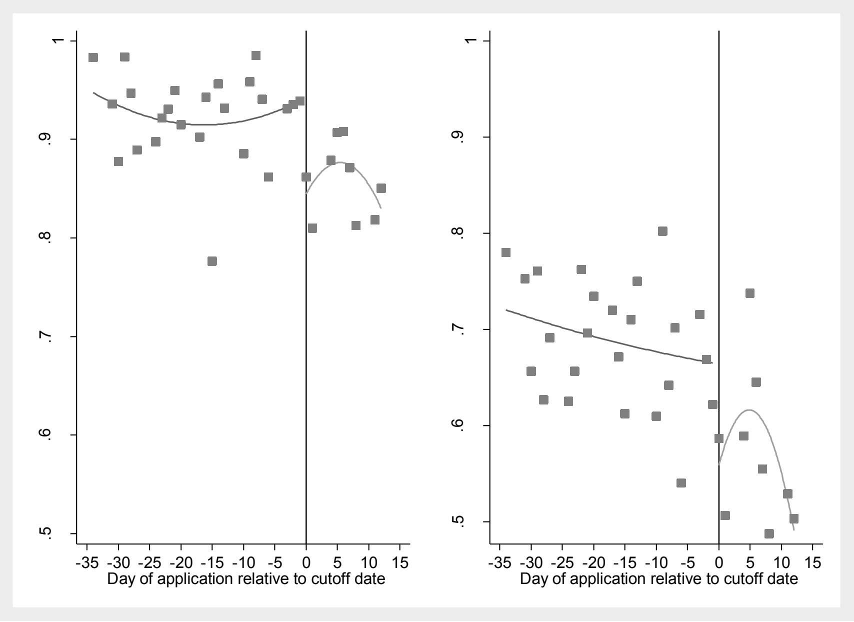
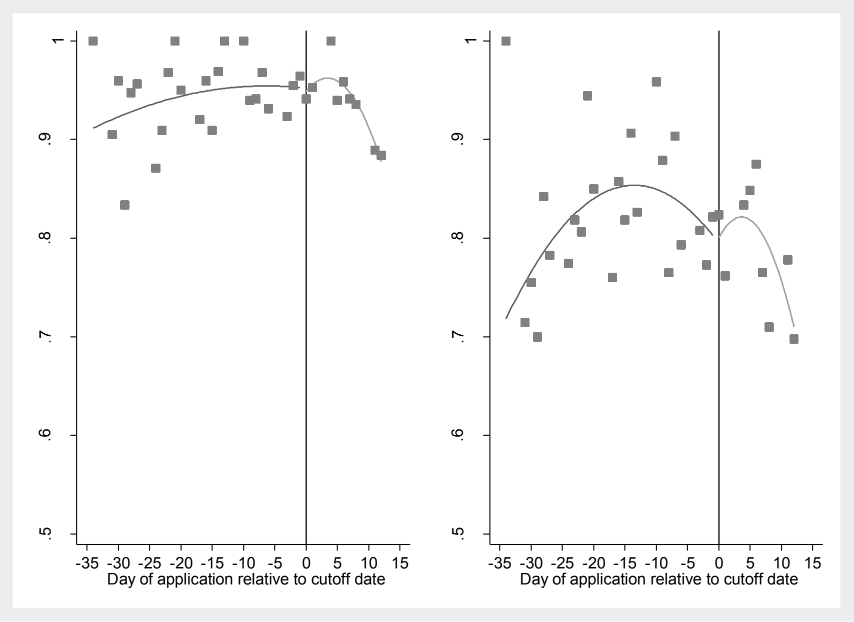


Figure C2 Proportion ever having been enrolled (left panel) and ever having completed (right panel) higher education during 2008-2013 by day of application, for subgroup of applicants that had already started at time of application



Appendix D Descriptive tables for subgroups by applicant and application characteristics

Table D1 Descriptive statistics sample of voucher applicants for bachelor of master degree programs by sector of work

	(1)	(2)	(3)
Variables	Primary education	Secondary education	Special Education
Panel A			
Applicant characteristics			
Female	0.89 [0.86]	0.58 [0.47]	0.78 [0.75]
Age	37.2 [41.9]	38.4 [44.3]	37.7 [42.5]
Living in Randstad region	0.39 [0.40]	0.37 [0.41]	0.39 [0.39]
Gross monthly salary (2008) at appointment of 1 FTE	2829 [3099]	3030 [3383]	2938 [3275]
Appointment in FTE	0.82 [0.74]	0.85 [0.83]	0.88 [0.81]
Application characteristics			
Already started higher education program at time of application	0.20	0.26	0.16
Program duration (in years)	1.97	2.68	1.96
Panel B			
Treatment variables			
Received voucher in first application period (2008)	0.64	0.66	0.60
Received voucher in any of first seven application periods (2008-2013)	0.80	0.83	0.73
Panel C			
Outcome variables (2008-2013)			
Ever having been enrolled in higher education	0.91	0.91	0.91
Completed higher education program	0.75	0.62	0.68
Proportion completed higher education program of those enrolled	0.82	0.68	0.75
Still enrolled in higher education in 2013 but did not complete a program during 2008-13	0.03	0.11	0.03
Still in education in 2012	0.89	0.81	0.89
N	1,893	1,776	551
Proportion of all applicants	0.45 [0.56]	0.42 [0.35]	0.13 [0.10]

Country averages are presented between brackets.

Table D2 Treatment and outcome variables before and after the cut-off date by sector of work and age category

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Variables	Primary education		Secondary education		Special education		15-34 years		35-64 years	
	before	after	before	after	before	after	before	after	before	after
Panel A										
Treatment variables										
Received voucher in first application period (2008)	0.91	0.00	0.88	0.00	0.89	0.00	0.89	0.00	0.90	0.00
Received voucher in any of first seven application periods (2008-2013)	0.93	0.44	0.93	0.54	0.92	0.33	0.94	0.44	0.93	0.48
Panel B										
Outcome variables (2008-2013)										
Ever having been enrolled in higher education	0.90	0.88	0.93	0.86	0.93	0.86	0.94	0.90	0.92	0.85
Completed higher education program	0.77	0.68	0.65	0.53	0.70	0.63	0.76	0.65	0.68	0.59
Proportion completed higher education program of those enrolled	0.85	0.77	0.70	0.61	0.75	0.73	0.81	0.74	0.73	0.69
Still enrolled in higher education in 2013 but did not complete a program during 2008-13	0.03	0.05	0.10	0.15	0.02	0.03	0.05	0.09	0.06	0.08
Still in education in 2012	0.90	0.88	0.91	0.85	0.82	0.82	0.89	0.86	0.89	0.86
N	1,382	570	1,352	467	381	178	1,377	523	1,660	660

Table D3 Treatment and outcome variables before and after the cut-off date for subgroups by program duration and starting status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Variables	Program duration 0-1 year		Program duration >1-2 years		Program duration > 2 years		Not started at time of application		Already started at time of application	
	before	after	before	after	before	after	before	after	before	after
Panel A *Treatment variables*										
Received voucher in first application period (2008)	0.92	0.00	0.92	0.00	0.87	0.00	0.93	0.00	0.78	0.00
Received voucher in any of first seven application periods (2008-2013)	0.94	0.26	0.95	0.49	0.92	0.55	0.96	0.51	0.85	0.30
Panel B *Outcome variables (2008-2013)*										
Ever having been enrolled in higher education	0.93	0.88	0.93	0.86	0.93	0.88	0.92	0.86	0.94	0.93
Completed higher education program	0.81	0.65	0.77	0.65	0.64	0.55	0.69	0.57	0.82	0.78
Proportion completed higher education program of those enrolled	0.88	0.74	0.83	0.76	0.68	0.62	0.74	0.66	0.87	0.84
Still enrolled in higher education in 2013 but did not complete yet	0.02	0.06	0.03	0.04	0.09	0.15	0.06	0.09	0.05	0.07
Still in education in 2012	0.88	0.86	0.90	0.88	0.88	0.85	0.89	0.87	0.89	0.84
N	387	261	1,376	481	1,274	441	2,353	936	684	246

4.

The effects of higher teacher pay on teacher retention: Evidence from regional variation in teacher salaries[47]

Abstract

This chapter investigates the effects of higher teacher pay for secondary school teachers on their teacher retention decision and enrollment in additional schooling. We exploit regional variation in teacher pay induced by the introduction of a new teacher remuneration policy. This policy provided schools in an urbanized region with extra funds to place a larger share of teachers in a higher salary scale. We exploit this policy in an instrumental variable setup to estimate the effects of higher teacher pay on our outcomes. The main finding is that we find no effects of higher teacher pay on the probability to stay in the teaching profession. The policy however succeeded in keeping a slightly larger share of teachers in the targeted region. In addition, our findings suggest that the policy increased teachers' enrollment in bachelor or master degree programs from 2.3% to 3.2%. This finding is consistent with the setup of the policy in which one of the criteria for placement in a higher salary scale is that teachers would obtain extra qualifications or gain extra expertise.

[47] This is joint work with Sander Gerritsen and Sonny Kuijpers: Steeg, M. van der, S. Gerritsen & S. Kuijpers, 2015, The effects of higher teacher pay on teacher retention: Evidence from regional variation in teacher salaries, *CPB Discussion Paper*, no. 316.

4.1 Introduction

Many countries face teacher shortages, especially in regions where there are better outside options for teachers, higher costs of living and higher shares of low-SES pupils. (e.g. Clotfelter et al., 2008; Greaves & Sibieta, 2014). Policymakers respond to these shortages with various policies to attract more people into the teaching profession or to retain a higher share of teachers. Among these policies, a higher teacher pay is one of the most widely used.[48] It is not immediately clear whether higher teacher pay increases teacher retention rates. The teacher retention choice is often motivated by factors other than salary. Studies by Hanushek et al. (2004) and Clotfelter et al. (2011) suggest that the effects of teacher pay on teacher retention are very modest compared to the effect of pupil characteristics. Teachers prefer not to work in schools with high shares of disadvantaged children. Moreover, while there is a large literature that suggests that higher teacher pay increases teacher retention (e.g. Murnane et al., 1989; Dolton & Van der Klaauw, 1995; Hanushek et al., 1999; Imazeki, 2005; Reed et al., 2006; Gilpin, 2011), very few of these studies exploit (plausibly) exogenous variation in teacher pay. Most existing studies of effects of teacher pay on teacher retention exploit across-district variation in teacher salaries.[49] Estimates derived from these studies will be biased for instance if districts with unobserved positive (negative) attributes of teachers offer higher (lower) salaries.

The studies of Hendricks (2014) and Clotfelter et al (2008) are two notable exceptions. Hendricks (2014) uses detailed panel data from the state of Texas to estimate the effects of higher teacher pay on teacher retention in a differences-in-differences type of setting. Controlling for changes in district and local labour market characteristics, he finds that a 1% increase in teacher pay reduces the turnover rate by 1.4% and that this effect is largest for inexperienced teachers. Clotfelter et al. (2008) exploit both within- and between-school variation in teacher pay caused by the introduction of subject-specific retention bonuses for teachers in public secondary schools with either high-poverty rates or low test scores. Controlling for time-varying school, district and labour market characteristics, they find that 1800 USD retention bonuses led to a relative reduction of turnover rates at targeted schools by 17%.

[48] Other policies are writing-off student loans in exchange for a commitment to teach, subsidies for housing and the expansion of alternative certifications (Hanushek et al., 2004).

[49] This also holds for studies looking at effects on entry decisions into teaching (e.g. Manski, 1987; Dolton, 1990; Wolter & Denzler, 2003; Chevalier et al., 2007) or on pupil test scores (e.g. Dolton & Marcenaro-Gutierrez, 2011).

We contribute to this small literature by exploiting regional variation in teacher pay induced by the introduction of a Dutch teacher pay policy in 2009. The policy provided schools in an urbanized region with relatively large shares of disadvantaged pupils with additional funds to place a larger share of their teachers in a higher salary scale. Nearly 20% of all teachers in the targeted region were given the perspective of a 17% salary increase through placement in a higher salary scale. We use this policy as an instrument for higher teacher pay in an IV-strategy to estimate its effects on retention as a teacher.

Our research differs from that of earlier teacher pay and retention papers by Hendricks (2014) and Clotfelter et al. (2008) in an important way. Whereas these studies look at the effects of salary increases on teacher turnover at targeted schools (Clotfelter et al.) or regions (Hendricks), we focus on effects on retention in the teaching profession. That is, we investigate whether salary increases affects teachers' decisions to stay in the teaching profession.[50] Our national data allows us to track all Dutch teachers in the Netherlands from 1995-2014, such that we can reliably determine whether a teacher leaves the teaching profession or not. In addition, we look at the effects of higher teacher pay on teachers' decisions to complete more schooling as this was one of the criteria for being placed in a higher salary scale. To our knowledge, effects of higher teacher pay on schooling activity of teachers have not been studied before.

Our main findings are as follows. First, we find no effects of being placed in a higher salary scale on the probability to stay in the teaching profession. Second, we find that teachers switch somewhat less from treatment to control regions because of the new remuneration policy, but that this does not affect our results found for retention in the teaching profession. Hence, the policy succeeded in keeping a slightly larger share of teachers in the targeted region. These positive effects are however small relative to the costs of the policy.[51] Third, we find that the policy has a positive impact on teachers' enrollment in degree programs. Our estimates suggest an increase in this probability from 2.3% to 3.2%. This finding is consistent with the setup of the policy in which one of the criteria for placement in a higher salary scale was that teachers would obtain extra qualifications or expertise.

[50] We use a dummy that equals 1 if the teacher is in the teaching profession and 0 if she is out. Clotfelter et al. and Hendricks use a dummy that equals 1 if the teacher works at a school or region and 0 if she exits this school or region. In that case the teacher can still teach but at another school or region not covered in the dataset.

[51] The policy cost on average 50 million euro per year and an estimated 0.4% of teachers per year decided not to switch from treatment region to control region because of the policy, see also section 7.

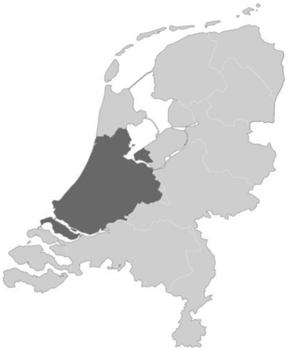
Taken our findings and those of Clotfelter et al. and Hendricks together, we conclude that a higher teacher pay might not be effective in increasing retention rates in the teaching profession, but might be an appropriate policy tool to decrease turnover rates in specific schools or regions, especially in schools or regions with high shares of disadvantaged pupils. In addition, it can be used as a financial incentive to increase participation in follow-up teacher training.

This chapter proceeds as follows. Section 4.2 gives details of the regional teacher pay policy and context. Section 4.3 discusses our data. Section 4.4 describes the empirical strategy. Main results are presented in section 4.5. Section 4.6 deals with heterogeneous treatment effects. Section 4.7 concludes and provides a discussion of the results.

4.2 Institutional background and the regional teacher pay policy

The regional teacher pay policy was introduced in 2009. Secondary schools in the Dutch Randstad region received additional funds from the government to place a substantially larger share of their teachers in a higher salary scale. Figure 4.1 shows the Randstad region in the Netherlands in dark grey. The Randstad region covers around 40% of all schools and pupils in the Netherlands, indicating a relatively densely populated area. It is a relatively urbanized area that contains the four biggest cities of the Netherlands.

Figure 4.1 Randstad region within the Netherlands



The motivation for the Randstad policy was to reduce the relatively large wage differential between teaching and other jobs in the Randstad region in order to mitigate (future) teacher shortages. In addition, the policy had to compensate for more difficult working conditions in Randstad schools due to more disadvantaged pupil populations. Heyma et al. (2006) show that the regional wage differential to jobs outside teaching for female teachers is -30% in Amsterdam (situated in the Randstad), whereas in the rural province of Friesland (not situated in the Randstad) this is -5%. In addition, the share of lessons not given by a certified teacher for the subject is about a quarter higher in the Randstad region, i.e. 28% versus 22%.[52] The share of pupils that lives in so-called 'poverty problem accumulation areas'[53] is more than twice as large in the Randstad schools (i.e. 24% versus 10%, see Table 4.1 in the next section).

There are three salary scales for secondary school teachers in the Netherlands: LB (low), LC (middle), and LD (high) with maximum (i.e. end-of-scale) gross monthly salaries of 3784, 4413 and 5022 euro, respectively. Starting salaries are roughly equal at around 2500 euro. Most of the teachers are in the low salary scale. Being placed in the mid salary scale instead of the low salary scale gives the perspective of a 17% higher salary end-of-scale, which is equal to 7200 euro in gross terms per year.

The goal of the policy was to place 39% of the 'LB'- teachers in the Randstad in the LC scale by 2014. Schools outside the Randstad also received some additional funding, but much less than the schools inside the Randstad. They could only place 10% of the ‘LB’- teachers in the mid salary scale LC. This implies a difference of 29 percentage points in the growth of the share to be placed in the mid salary scale at the cost of the low salary scale between the Randstad and the non-Randstad region.

In total, 290 million euro was made available to the Randstad schools over the period 2009-2014 to achieve this goal, that is, a little less than 50 million euro per year on average.[54] By 2014, however, the realized difference between Randstad and non-Randstad schools was 18 percentage points and not the originally targeted 29 percentage points (see also figure 4.2 in section 4.4). The Ministry of Education stated in a letter to parliament that the budget turned out to be insufficient to cover the structural extra wage costs of the Randstad policy (Ministry of Education, 2015). In addition, due to concerns about how the additional funding

[52] Based on pre-treatment year 2008. Being certified for a certain lesson depends on two things. First, the teacher should have a teacher education degree in the subject of that lesson. Second, the teacher should have a master degree in the subject of that lesson when the lesson is taught in the upper years of secondary education.

[53] This is a zip code area that meets the following three criteria: 1) the share of people with low incomes exceeds 15%, 2) the share of people being welfare recipient exceeds 13% and 3) the share of non-western immigrants exceeds 7%.

[54] The goal was publicly monitored by a website. National goals have been translated into goals per school. As an additional incentive for schools to spend the additional funding on placement of teachers in higher salary scales the extra funding for the second half of the policy period only became available upon reaching intermediate targets for 2011.

was spent, it was confirmed that the extra budget was used for placing more teachers in higher salary scales and not for other purposes. We have investigated this in Section 4.5.3, and have found no evidence that the funding has been spent in ways other than in placing teachers in higher salary scales.

The additional funds were given to the schools in addition to the regular lump-sum funding that schools receive from the government. Teacher salaries are paid out of this lump-sum funding and take up the largest share of expenditures. The lump-sum also covers salaries of non-teaching personnel such as management and supporting staff, material costs and maintenance costs of school buildings. School leaders decide in which salary scale teachers are placed. This is often based on teacher qualifications (i.e. whether the teacher has a master degree or not) and specific expertise. In the Randstad policy, one of the additional criteria for being placed in a higher teacher salary scale was that teachers would complete extra schooling. This could be i) extra training or expertise in a pedagogical-didactical area, ii) an additional qualification that allows a teacher to teach in two subjects or more or iii) a master degree in the particular subject being taught. Since 2008, teachers are stimulated to obtain this additional schooling by applying for publicly financed schooling vouchers that allow them to follow additional education. These vouchers consist of a financial contribution to teachers to cover tuition costs of a bachelor or master degree program and a contribution to their schools to finance a substitute teacher while the teacher is on study leave.[55]

4.3 Data

4.3.1 Data sources and variables

The data for our analysis come from various sources. Information on teacher retention and teacher salary come from two files: the *Mirror* and *Functiemix* datafiles. *Mirror* has been produced and provided to us by an executive agency of the Dutch Ministry of Education (i.e. *DUO, Dienst Uitvoering Onderwijs*). This dataset contains information on the working status of all Dutch teachers in the period 1995-2013. It indicates for instance whether a teacher works in a particular year and, if so, at which school she works. *Functiemix* has been provided by *CenterData* and is similar to *Mirror*, except that it contains more information on teachers and spans a shorter time period, 2006-2014. For our analysis we use both datasets. For checks on our identification strategy we use *Mirror* as this covers a longer time period.

[55] We refer to Van der Steeg and Van Elk (2015) for more details on this teacher schooling voucher scheme.

For estimation of the effects we use *Functiemix* as it contains more detailed information and hence more covariates. Throughout the chapter we refer to *Mirror* as our long sample and *Functiemix* as our estimation sample. All information in these files is measured in October of a particular year.

Information on teacher's schooling decisions comes from another data source of DUO, the *Teacher Schooling Voucher* file. This file gives information at the school level on the share of teachers that applied for a teacher schooling voucher in a particular year. Data are available for the years 2008-2013. This means we have no data on the pre-treatment period. This is because the teacher schooling voucher was not introduced until 2008. From DUO we also obtained information on additional school characteristics such as the share of pupils with a disadvantaged background and school size. This file is called the *school-pupil characteristics file*. From Statistics Netherlands we obtained some additional information on local labor market conditions, that is, unemployment rates in Randstad and non-Randstad regions.

To obtain one main estimation sample, we merged the latter two files with the *Functiemix* file at the school level. We also carried out a few steps to get rid of noise in our data. We refer to Appendix A.1 for a detailed explanation of this procedure. Our estimation sample finally consists of 480,600 observations for which we have full information on teacher retention. The observational unit is a teacher by year.

Main outcomes

In our main analysis we use two outcomes: a dummy for teacher retention and the share of teachers that applied for a schooling voucher. The first is given at the individual teacher level, the latter at the school level (see above).

Teacher retention equals 1 at time *t* if a teacher is observed working at *t* and *t-1.* It equals 0 at time *t* if a teacher is observed working at *t-1* but not at *t*. Teacher retention is reported missing at time *t* if the teacher is not observed working at *t-1*. In that case we cannot identify the retention status of the teacher. Note that by defining this variable as such, we lose the first year of our data.

The share of teachers that applied for a schooling voucher is given for every school *s* and year *t*.

Main independent variable

The main independent variable is a dummy that equals 1 if the teacher is observed in the mid or high salary scales LC or LD and 0 if the teacher is observed in the low salary scale LB. In the next section we describe our empirical instrumental variable strategy to identify the effect of being placed in a higher salary scale on teacher retention.

It is important to note that we will look at the effects of being placed in a higher salary scale rather than at the effect of having a higher salary since the latter would not capture the full treatment given to teachers. Teachers are promoted by the Randstad policy rather than being given a higher salary in itself. Hence, a higher salary alone does not capture the future career prospects of being promoted to a higher salary scale. Taking salary as independent variable would invalidate the exclusion restriction as we cannot distinguish salary effects from the effects of possible future career prospects of being promoted.

Covariates

As covariates we use a number of school and teachers characteristics. Teacher covariates include age, teaching load, and gender. School covariates include school size, pupil population growth, and the share of pupils from disadvantaged neighborhoods.[56]

Other outcomes

In our analysis we also use other outcomes necessary to support the assumptions underlying our identification strategy. They include the unemployment rate, the pupil-teacher ratio, school board finances, the number of new teachers, the share of lessons given by a certified teacher and a dummy that indicates whether a teacher switches between Randstad and non-Randstad schools. These outcomes will be discussed when they are exploited.

4.3.2 Construction of main estimation sample

We have a full sample of 480,600 observations for which we have full information on teacher retention. If we use this sample in our analysis, then we compare the whole Randstad region with the rest of the Netherlands. These two regions might not be very similar. The setup of the Randstad policy allows us to create more similar regions. This can be done by selecting schools around the geographical cutoff that separates the two regions. The idea is that

[56] A pupil from a disadvantaged neighborhood is defined here as a pupil living in a so-called poverty problem accumulation area. See footnote 4 for a description of the criteria used for identifying these areas.

(teachers in) schools will become more similar if we select schools closer to the border of the Randstad. To create such a sample, we have selected 53 municipalities. They comprise the first two rings of municipalities around the geographical cutoff. Taking these two rings is based on the consideration that i) the municipalities are close to the Randstad border and ii) the treatment and control group would comprise a large enough sample size to estimate effects.[57] Note that the biggest four cities in the Randstad (Amsterdam, Rotterdam, Den Haag and Utrecht) are not included in this sample as they do not lie at the border. This local sample will be our main estimation sample. Table A.1 in the Appendix gives the list of the selected municipalities and figure A.1 provides a map.

4.3.3 Descriptive statistics

Table 4.1 shows descriptive statistics for our local and full sample. Panel A shows statistics for all years pooled together, panel B shows statistics for the pre-treatment year 2008. For each sample, statistics are given for the treatment group (Randstad) and control group (Outside Randstad). We observe similar patterns in panels A and B, except that in 2008 there are no significant salary differences between our control and treatment group. In the full sample there are statistically significant differences between the groups in terms of teacher and school characteristics. Randstad teachers are a bit younger, are more likely to be female, and have a somewhat smaller assignment size compared to non-Randstad teachers. In addition, Randstad schools are smaller in size and have more disadvantaged children than non-Randstad schools. In our local sample these differences disappear. This is what we would expect if we select schools in regions closer to the border of the Randstad; they should become more similar.

[57] Taking three rings would mean including big cities like Amsterdam and Utrecht in the treatment group but not in the control group. This would create less similar comparison groups. Taking only one ring would decrease the sample size substantially.

Table 4.1 Descriptive statistics for local and full sample, all years pooled (panel A) and pre-treatment year (panel B)

	Local sample			Full sample		
Variable	Outside Randstad (Control)	Randstad (Treatment)	P-value	Outside Randstad (Control)	Randstad (Treatment)	P-value
A: all years pooled						
Main outcome variables						
Retention as a teacher (a)	0.934	0.930	0.176	0.936	0.925	0.000
Teacher applies for schooling voucher (b)	0.030	0.029	0.483	0.027	0.025	0.094
Main independent variable						
Teacher in mid or high salary scale (a)	0.393	0.494	0.000	0.404	0.516	0.000
Covariates						
Teacher’s age in years	45.29	45.09	0.609	45.85	45.06	0.000
Teacher’s assignment size in FTE	0.822	0.823	0.912	0.826	0.809	0.000
Female teacher	0.484	0.479	0.686	0.479	0.501	0.000
School size (b)	1,524	1,512	0.950	1,54	1,399	0.105
Yearly school population growth (b)	0.016	0.011	0.507	0.007	0.011	0.351
Pupils from disadvantaged neighborhood (b)	0.083	0.091	0.679	0.098	0.238	0.000
Number of observations	61,611	58,882	120,493	279,149	201,451	480,600
Number of schools	78	74	152	350	287	637
B: pre-treatment year (2008)						
Main outcome variables						
Retention as a teacher (a)	0.916	0.915	0.895	0.926	0.910	0.000
Teacher applies for schooling voucher (b)	0.032	0.023	0.077	0.027	0.022	0.010
Main independent variable						
Teacher in mid or high salary scale (a)	0.357	0.340	0.344	0.357	0.361	0.686
Covariates						
Teacher’s age in years	44.24	44.11	0.770	44.70	44.06	0.002
Teacher’s assignment size in FTE	0.828	0.830	0.785	0.833	0.813	0.000
Female teacher	0.469	0.466	0.802	0.463	0.489	0.000
School size (b)	1,476	1,506	0.879	1,530	1,340	0.032
Yearly school population growth (b)	-.005	.003	0.315	-.005	-.004	0.867
Pupils from disadvantaged neighborhood (b)	0.094	0.097	0.906	0.111	0.245	0.000
Number of observations	7,736	7,645	15,381	33,721	25,002	58,723
Number of schools	76	72	148	346	270	616

Notes: a) weighted by the assignment size of teachers in FTE’s, b) school averages.

4.4 Empirical strategy

4.4.1 Instrumental Variables Framework

We are interested in the effect of the treatment, i.e. being placed in a higher teacher salary scale, on our outcomes. To estimate the treatment effect one could use the following specification:

$$(1)\ Y_{ist} = \alpha_0 + \alpha_1 HS_{ist} + \alpha_2 \boldsymbol{X_{ist}} + \theta_{ist}$$

in which Y_{ist} represents the outcome of teacher *i* in school *s* in year *t*, HS_{ist} represents a dummy that equals 1 if teacher *i* in school *s* at time *t* is in scale LC or higher (and 0 if in LB), $\boldsymbol{X_{ist}}$ is a set of controls, e.g. school fixed effects and teacher characteristics, and θ_{ist} is the error term which captures unobservable determinants of the outcome. For our outcome teacher retention we relate the retention decision at time t, which is measured with respect to t-1, to the conditions they were exposed to at t-1.[58] The parameter of interest is α_1, which represents the effect of being placed in a higher teacher salary scale on the outcome.

Using cross sectional data and estimating this specification with OLS will probably yield a biased estimate of α_1 because of the endogeneity of HS_{ist}. Salaries are not randomly assigned to teachers. On the contrary, there are a lot of reasons why some teachers end up earning more than others. Teachers and their salaries often differ from each other in ways not observed by the researcher. For instance, better teachers with unobserved qualities could have been placed in higher salary scales by the school board in order to keep them in the teaching profession. In that case their unobserved qualities influence both their salary and their retention decision, causing any OLS-estimate to be biased.

We therefore use a two stage least squares (IV-)approach to address this endogeneity problem. We exploit the Randstad bonus as an instrument for being placed in a higher teacher salary scale. This bonus affected and benefited the teachers in the schools in the Randstad after 2008, while the teachers in the schools in the other regions were unaffected by this bonus. The first stage in this framework is

[58] In that case the index of the right-hand side variables is t-1. For instance, the retention decision of a teacher in 2014 with respect to 2013 is related to the salary scale and school characteristics she is exposed to in 2013. Because for our main analysis we pool our data over 2007-2014, the choice of using yearly retention rates as our outcome measure may give rise to selection effects over time. We address these issues in sections 5.2 and 5.3.

$$(2)\ HS_{ist} = \beta_0 + \beta_1 RS_s * POST_t + \beta_2 RS_s + \beta_3 \boldsymbol{X_{ist}} + \tau_t + \varepsilon_{ist}$$

in which RS_s represents a dummy that equals one if school *s* resides in the Randstad (RS) region; $POST_t$ represents a dummy that equals 1 if the outcome is observed post treatment, i.e. in 2009 or thereafter (*t≥2009*) and 0 if the outcome is observed pre-treatment, i.e. in 2007 or 2008, and τ_t are year fixed effects. The parameter of interest is β_1, which represents the effect of the Randstad bonus on the probability of being placed in a higher teacher salary scale. Note that this first-stage equation is a basic differences-in-differences model in which HS_{ist} is the outcome. The second stage is

$$(3)\ Y_{ist} = \gamma_0 + \gamma_1 \widehat{HS_{ist}} + \gamma_2 RS_s + \gamma_3 \boldsymbol{X_{ist}} + \tau_t + \vartheta_{ist}$$

where $\widehat{HS_{ist}}$ is the predicted probability of equation (2). Estimates of parameter $\gamma_1$yield the causal effect of the treatment on the outcome if the regular IV-conditions apply (see below). The corresponding reduced form of equation (3) is

$$(4)\ Y_{ist} = \delta_0 + \delta_1 RS_s * POST_t + \delta_2 RS_s + \delta_3 \boldsymbol{X_{ist}} + \tau_t + \theta_{ist}$$

in which δ_1 represents the impact of the Randstad policy on the outcome. This can be considered as an intention-to-treat effect. We use this specification to look at the impact of the Randstad policy on the share of teachers that applied for a schooling voucher. We cannot use an IV-approach for this outcome variable because applying for a schooling voucher precedes being placed in a higher salary scale. As noted in section 4.2, one of the criteria for being placed in a higher salary scale is that teachers would obtain additional schooling.

4.4.2 Validity of the IV-approach

To apply this IV-setup three conditions should hold. First, the Randstad policy should have a significant impact on the probability of being placed in a higher salary scale, which means that the model should not suffer from a weak instrument problem. Second, receiving the Randstad policy treatment bonus should be independent of the error term (second stage independence). Third, the Randstad policy should *only* have an effect on teacher retention via the increased probability of being placed in a higher salary scale (second stage exclusion restriction). We address the first two conditions below. After presenting our main results, we discuss possible selection effects like switching behavior of teachers that may bias our IV-estimates. Also, we discuss the validity of the exclusion restriction.

First stage relevance

The Randstad teacher pay policy, introduced in 2009, should have a significant impact on teacher salary in the Randstad region as compared to regions outside the Randstad. Figures 4.2a (full sample) and 4.2b (local sample) show that this is the case. The share of teachers in the mid or high salary scale (i.e. not in the low salary scale) increases substantially more in the treatment group than in the control group after the introduction of the Randstad policy. The difference increases to 18 percentage points in 2014.[59] The F-statistics of the first-stage regressions are above 200, largely exceeding the rule of thumb of 10 (Staiger & Stock, 1997). This shows that we do not suffer from a weak instrument problem; that is, the Randstad policy significantly increases the probability of being placed in a higher salary scale for teachers in the Randstad region. In addition, the figures show that the pre-treatment trends are rather similar across treatment and control regions. This means that salaries were not that different between the regions before introduction of the policy. From the figures it also becomes clear that the nearly 20 percentage points difference between the Randstad and non-Randstad does not coincide with the policy goal of a 29 percentage points difference. According to the Ministry of Education the 290 million euro turned out to be insufficient to achieve the policy goal, see section 4.2.

[59] The impact on average gross salary of all teachers is 2.2 percentage points by 2014, which is also highly significant. This implies that the 18% of teachers that were additionally placed in a higher salary scale due to the policy received approximately 13% more salary (=2.2/0.18). The perspective of being placed in a higher salary scale was a 17% higher salary.

Figure 4.2a Development of share of teachers in mid or higher salary scales for control and treatment group, full sample

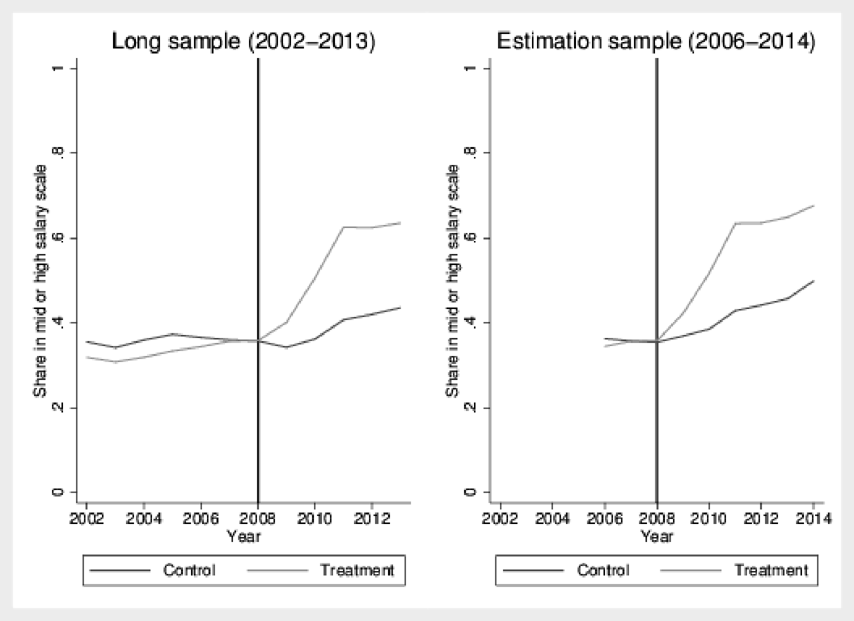
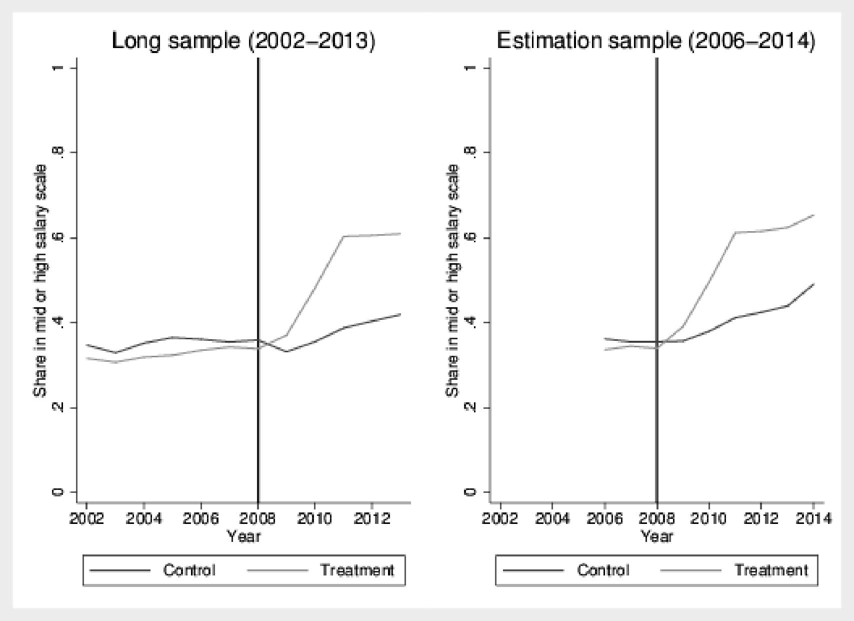


Figure 4.2b Development of share of teachers in mid or higher salary scales for control and treatment group, local sample



Second stage independence

Second we address the second stage independence. This means that receiving the Randstad bonus should not be correlated with unobserved time-varying characteristics of teachers and their outcomes. As we compare Randstad teachers with teachers outside the Randstad over time, this comes down to a common-trend assumption. That is, we assume that the trend in the outcome in the treatment group would have followed the same trend as that of the control group in absence of the treatment (i.e. receiving the Randstad policy). Although this assumption cannot be tested directly, its credibility can be strengthened by showing pre-treatment trends in the outcome variables for the treatment and control group. These trends should be similar and should not diverge until the introduction of the Randstad policy in 2009.[60] Our data allows us to investigate this for teacher retention but not for enrolment in schooling vouchers. We can test this with two datasets: a long dataset spanning the period 1995-2013 (*Mirror*) and a main estimation sample which spans the period 2007-2014 and has more covariates (*Functiemix).* With the long sample we can test whether trends are similar before introduction of the policy as we have data that go back as far as 1995. The main estimation sample gives us only one year before the intervention as it only goes back to 2007. We will use both datasets to check whether trends are similar, but only use the main estimation sample to estimate the effects of the treatment in the next section.

We perform three analyses. First, we start with a graphical analysis. Figure 4.3a shows the trends in teacher retention rates for the control and treatment group for our full sample. The left (right) figure exploits the long (estimation) sample. A vertical is drawn at 2008, the last year before introduction of the policy. The figure shows that retention rates vary between 90 and 95%, and that the pre-trends between control and treatment group are rather similar. As the trends continue to be similar after introduction of the policy, they also suggest that there are no direct effects of the Randstad policy. Figure 3b does the same for our local sample. Hence this figure shows the trends when the dataset is limited to schools in the 53 border municipalities. The idea is that trends will become more similar if we select schools closer to the Randstad border. This seems to be confirmed as the lines lie closer to each other when compared to figure 3a.

[60] A reason why pre-trends could differ is when early announcement of the program would cause teachers to select themselves in Randstad schools before the start of the Randstad teacher pay policy. We think this is unlikely to be the case as information about the policy was not made public until spring of 2009.

Figure 4.3a Development of retention rates for control and treatment group, full sample

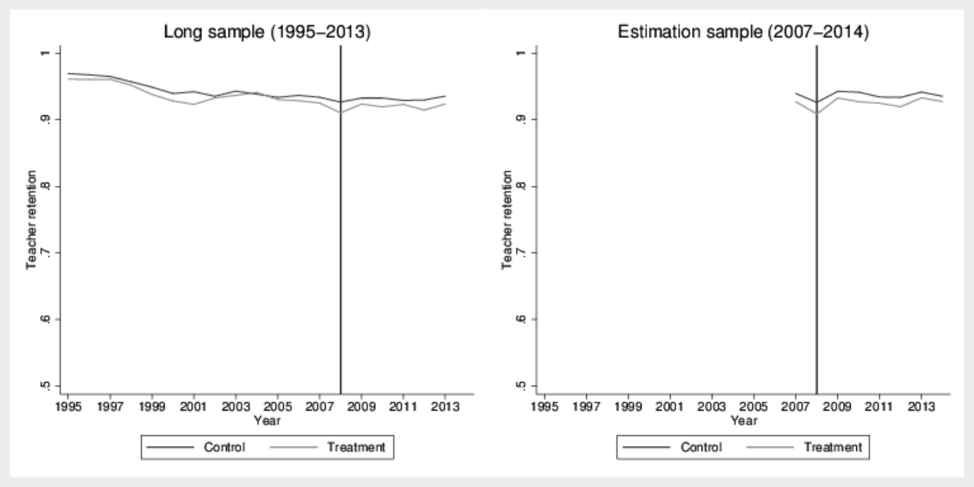
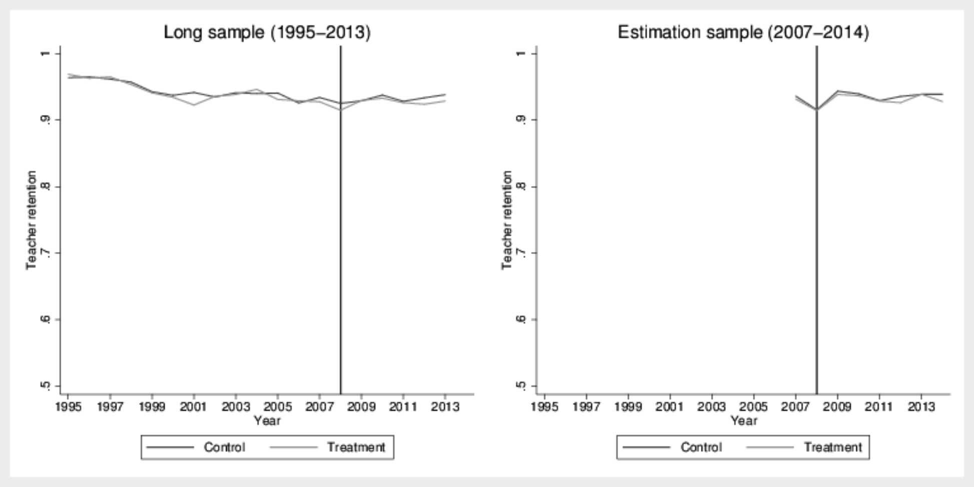


Figure 4.3b Development of retention rates for control and treatment group, local sample



However, from this graphical analysis we cannot be certain yet that pre-trends are similar. To shed more light on the similarity of the pre-trends we perform a second analysis in which we statistically test whether trends are similar. We use the long sample and run two types of regressions.

First, we select the observations before 2008 and regress teacher retention on a constant, the set of available control variables, a linear time trend, a dummy for Randstad, and the interaction of the time trend with the Randstad dummy. If pre-trends are similar, then the estimated coefficient for the interaction should be close to zero.

Second, we run a similar regression except that, instead of a linear time trend, we include dummies for the years and its interactions with the Randstad dummy. Hence, in this model we are more flexible and allow the time trends to deviate from each other in a non-linear way. If pre-trends are similar, then the estimated coefficients of the interactions should be close to zero.

Table 4.2 presents results of these regressions for our local and full sample. Panel A shows results of the model with a linear time trend, and panel B gives results of the model with a non-linear trend.

Table 4.2 Test on similarity of pre-treatment trends for teacher retention

	Local sample		Full sample	
	Estimate (1)	Standard error (2)	Estimate (3)	Standard error (4)
Panel A: Linear trend				
year*treatment region	-0.000	0.000	0.000	0.000
Panel B: Non-linear trend				
1995*treatment region	0.016**	0.008	0.007*	0.004
1996*treatment region	0.005	0.009	0.006	0.004
1997*treatment region	0.011	0.008	0.010**	0.004
1998*treatment region	0.005	0.007	0.010**	0.004
1999*treatment region	0.006	0.008	0.004	0.004
2000*treatment region	0.009	0.008	0.004	0.005
2001*treatment region	-0.012	0.007	-0.005	0.005
2002*treatment region	0.012	0.009	0.012**	0.005
2003*treatment region	0.004	0.008	0.008*	0.004
2004*treatment region	0.016**	0.008	0.016***	0.004
2005*treatment region	-0.001	0.009	0.010**	0.005
2006*treatment region	0.014	0.012	0.007	0.005
2007*treatment region	0.004	0.007	0.006	0.004
Number of observations	206,654		907,950	

Notes: Every pair of columns in each panel represents the results of an OLS-regression. The odd columns give the estimates and the even columns give the standard errors. We control for the covariates as presented in table 4.1. The estimates in panel B represent deviations from the trend with respect to 2008 (2008=omitted category). Standard errors are adjusted for clustering at the school level. Significance levels: *** p<1%, ** p<5%, * p<10%.

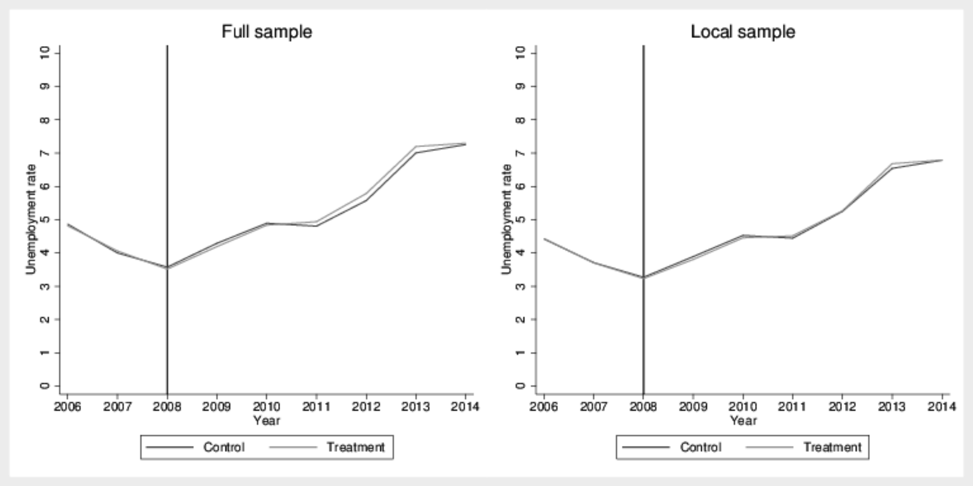
Panel A shows that the pre-trends in the teacher retention rates between control and treatment group do not deviate from each other when using a linear time trend. The estimated coefficients for the interaction term are close to zero and insignificant in both columns. Panel B shows that there are small deviations in some years when we allow the trend to be non-linear. In the full sample we find that for some years the estimated coefficient of the interaction is significant, suggesting that the pre-trend of the treatment group for these years deviate from that of the treatment group. However, when we limit the full sample to our border sample, the estimated coefficients are no longer significant anymore for most of the years.[61]

For our third test we investigate whether the labor market conditions are different between Randstad teachers and non-Randstad teachers. Although we are quite confident that the previous tests support the common trend assumption in the local sample, which suggests that unobserved differences between teachers develop similarly between regions, we consider this third test as an extra check on the independence assumption. If labor market conditions develop more favorably for teachers in the Randstad than for those outside the Randstad, for instance if Randstad teachers have more outside options during recessions than non-Randstad teachers, then this might affect their retention decision differently. Figure 4.4 shows unemployment rates for Randstad and non-Randstad regions for our full and local sample. The figure suggests that there are no large differences between the regions with respect to employment: both the level and the development of unemployment look similar, both pre- and post-treatment.[62] When we perform statistical tests as in table 4.2, the null hypothesis of the similarity of the trends is not rejected. This suggests that labor market conditions do not develop differently, which may give extra support to the common trend assumption. In the next section we present our main results. Thereafter we continue with a discussion on possible selection effects and the exclusion restriction.

[61] We stress that with so many years and hence estimates, it is not unlikely that some estimated coefficients pop up significant at conventional significance levels when testing a true null hypothesis of no effect.

[62] The unemployment rates have been weighted by the size of the labor force by municipality.

Figure 4.4 Trends in unemployment rates for control and treatment group



4.5 Main results

4.5.1 Effects of placement in higher salary scale on teacher retention

Table 4.3 contains first stage, reduced form, OLS- and IV-estimates. The OLS and IV results show estimates of the effect of being placed in a higher (i.e. non-low) salary scale on teacher retention. The reduced-form (RF) estimates show (intention-to-treat) estimates of the effect of the regional teacher policy on teacher retention. The first stage, also in the rows, represents the effect of the policy on the probability of being in a higher salary scale. The first four columns exploit our local sample of treatment border municipalities, whereas the last four columns exploit the full sample. Odd columns include no controls except year-fixed effects. Even columns include school-fixed effects, teacher's age, gender, teaching load, school size, population growth, and the share of pupils from a disadvantaged neighborhood. Standard errors have been clustered at the school level.

The first stage estimates mirror the picture in figure 4.2. They are highly significant and around 0.16. They suggest that the Randstad policy led to a 15-17 percentage points increase in the probability of being in a higher salary scale.

The OLS-estimates in column (1) and (2) vary between 0.02 and 0.03 and are significant at the 1% level. They suggest that being placed in a higher salary scale leads to a 2-3 percentage points higher probability of being retained. These estimates cannot be interpreted causally, because of the endogeneity of being placed in a higher salary scale.

Teachers in higher pay scales differ from teachers in lower pay scales in ways not observed by the researcher.

The IV-results control for this endogeneity and show negative and insignificant estimates. The point estimate is -0.015 in column (3). Including controls in column (4) hardly changes the estimate. This is what one would expect when treatment and control groups are similar. Hence, based on these IV-results, we find no effect of being placed in a higher salary scale on teacher retention.

We continue by investigating whether the results found with our local sample can be replicated with the full sample in columns (5) to (8). In these regressions we are less confident about the validity of the second stage independence assumption (see previous chapter). The OLS-estimates are similar to those in columns (1) and (2). The IV-estimates are insignificant. The IV-estimate in column (7) is 0.022 but drops to 0.008 when including controls in column (8). This may reflect the fact that treatment and control regions are less similar in the full sample. The result in column (8) seems to replicate the result found with the local sample. We conclude that we find no effect of being placed in a higher salary scale on retention as a teacher.

Table 4.3 Estimates of the effect of a higher salary scale (OLS and IV) and of teacher pay policy (reduced form) on retention as a teacher

	Local sample				Full sample			
	OLS	OLS	IV	IV	OLS	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Effect on retention	0.028***	0.021***	-0.015	-0.018	0.026***	0.020***	0.022	0.008
	(0.003)	(0.003)	(0.031)	(0.029)	(0.002)	(0.002)	(0.017)	(0.017)
Reduced form			-0.002	-0.003			0.003	0.001
			(0.005)	(0.005)			(0.003)	(0.003)
First stage			0.154***	0.160***			0.155***	0.159***
			(0.013)	(0.011)			(0.008)	(0.007)
Number of observations	120,493	120,493	120,493	120,493	480,600	480,600	480,600	480,600
Controls								
teacher characteristics	no	yes	no	yes	no	yes	no	yes
school characteristics and school-fixed effects	no	yes	no	yes	no	yes	no	yes

Notes: In the even columns we include a set of controls. Included teacher covariates are: gender, age category and assignment size category in FTE's. School covariates include school size category, the share of disadvantaged pupils and school population growth. Standard errors in parentheses adjusted for clustering at the school level. All regressions include year-fixed effects. Significance levels: *** p<1%, ** p<5%, * p<10%.

4.5.2 Selection effects

In our identification strategy we compare teachers in Randstad regions with teachers in non-Randstad regions over time. As we pool our data over 2007-2014 and use yearly retention rates as our outcome measure, this may give rise to selection effects that might bias our estimates. Two things can happen, which both change the composition of the teacher population in the Randstad relatively to that of the non-Randstad regions.

First, the Randstad policy can lead to unobserved different inflow and outflow of teachers. After each year teachers enter and exit the teaching profession, causing the teacher population to change over time. If the Randstad policy causes other types of teachers to enter or exit the teaching profession in the Randstad regions than in non-Randstad regions, this may bias our estimates of the effect of the treatment on teacher retention. This can happen, for instance, if newly entering teachers who select themselves in schools in the Randstad because of the higher salary are more likely to leave the teaching profession.

Second, the Randstad policy might lead to switching behavior of existing teachers. Teachers outside the Randstad who favor higher salaries might leave their schools and move to schools inside the Randstad. In addition, Randstad teachers may become less willing to switch to schools outside the Randstad because of the higher salary. In the same way as above, this may bias our estimates if switchers differ in unobserved ways from non-switchers.

To address these two issues we perform four tests. The first three tests relate to the first issue, the fourth relates to the second.

First, we look at the effects of the Randstad policy on the number of new teachers per school. If the policy increases the attractiveness of the teaching job and hence the number of teachers, then this might be an indication of a changing teaching population. Panel A in table 4.3 shows reduced form estimates of the effect of the policy on the number of new teachers per school. We find no evidence in favor of an increased influx.[63]

Second, we investigate whether the policy changes the composition of the teacher population in the Randstad relative to the non-Randstad. Panel B investigates this issue by showing estimates of the effect of the policy on (observable) background characteristics of teachers. We do not find evidence in favor of a changing distribution in terms of age, gender

[63] This finding is supported by recent research among bachelor and master students in teacher training programs. It was found that these students seriously underestimate both starting and maximum wages for teachers, that is, by 15% and 40% respectively (Researchned, 2015). It thus seems that prospective teachers are unaware of improved career prospects for teachers.

or assignment size in the local sample; the estimated coefficients are insignificant. We also do not reject the null hypothesis of similarity of the pre-trends in these variables (not shown in table).

Third, we investigate whether the quality of teachers has changed because of the policy. Changes in the quality of teachers could hint at composition effects and may lead to biased estimates of effects on teacher retention if teacher quality (certification) is correlated with teacher retention. We use the share of lessons given by a certified teacher as a proxy for teacher quality. It has been found that being certified for the subject is positively correlated with pupil outcomes (Goldhaber&Brewer, 2000; Clotfelter et al., 2010). Panel C of table 4.3 shows estimates of the effect of the Randstad policy on the share of lessons given by a certified teacher. The statistically insignificant and close-to-zero estimates do not hint at composition effects in terms of teacher quality.

Fourth, we look at the effects of the Randstad policy on switching behavior of teachers. Although less than 1% of the teachers switch annually between control and treatment regions, we will investigate to what extent switching behavior has changed due to the Randstad policy. This analysis also sheds light on the question whether the Randstad bonus succeeds in keeping more teachers in the targeted region. One of the goals of the policy is to retain teachers in the Randstad region. If teachers switch less from treatment to control group because of the policy, this could be considered a success. In panel D of table 4.3 we show reduced form estimates of the effect of the policy on a dummy that equals 1 if a teacher switches from Randstad region to non-Randstad region or vice versa. We find a small albeit statistically significant effect of -0.4 percentage points ($p<0.05$). This shows that switching behavior of teachers decreased a bit due to the policy.[64] The estimate suggests that those who would have switched from Randstad regions to regions outside the Randstad in absence of the policy now stick to the Randstad because of the higher salary. In the next section we show that switchers are more likely to exit the teaching profession. In our IV-setup we would then estimate a lower bound, because after introduction of the policy the treatment group will consist of a larger share of teachers with a higher probability of leaving the teaching profession. In the next section, we therefore investigate the sensitivity of our estimates with respect to switching behavior. In addition, we investigate the sensitivity of our estimates with respect to teacher's entries and exits, although the tests provided in this section do not hint at

[64] In this differences-in-differences setting the estimate would also have been negative if teachers in the control group would switch more often to the treatment group. Graphs of switching behavior, however, show that the effect is driven by the treatment group. It can be seen that switching behavior decreases in the treatment group relatively to that of the control group. Graphs are available upon request.

a changing distribution of teachers. The exclusion restriction of our IV-strategy will be discussed thereafter in Section 5.4.

Table 4.4 Reduced form estimates of impact of policy on various variables, check for composition effects

	Local sample		Total sample	
	(1)	(2)	(3)	(4)
A: Number of new teachers per school	-0.927	-0.517	-0.765	-0.702
	(1.521)	(1.089)	(0.722)	(0.537)
Number of observations	1,142	1,142	4,729	4,729
B: Background characteristics of teachers				
Age	0.167		-0.012	
	(0.242)		(0.126)	
Female	-0.009		-0.008***	
	(0.007)		(0.003)	
Assignment size	0.002		0.005**	
	(0.004)		(0.002)	
Number of observations	161,662		651,667	
C: Share of lessons given by a certified teacher	0.008	0.001	0.003	0.001
	(0.009)	(0.008)	(0.005)	(0.004)
Number of observations	62,154	62,154	247,368	247,368
D: Region switching	-0.004**	-0.004**	-0.004***	-0.004***
	(0.002)	(0.002)	(0.001)	(0.001)
Number of observations	111,082	111,082	442,893	442,893
Controls:				
teacher characteristics	No	Yes	No	Yes
school characteristics	No	Yes	No	Yes

Notes: Each cell is an OLS-regression. Standard errors are adjusted for clustering at the school level in panels B to D. In panel A robust standard errors are used. School-year average teacher and school covariates are used in panel A. The share of lessons given by a certified teacher is weighted by total number of lessons given. All regressions include year-fixed effects. Significance levels: *** p<1%, ** p<5%, * p<10%.

A potential issue that we cannot address with our data may arise if the teacher pay treatment causes spillovers in terms of decreased motivation of teachers because of missing out on a promotion to a higher pay scale. If this decreased motivation affects the decision of these teachers to leave the teaching profession our finding of a lacking effect on teacher retention might be a combination of a positive effect on retention of promoted teachers and a negative effect on retention of non-promoted teachers. In that case our reduced-form estimates of the zero average effects of the teacher pay policy on retention profession are still relevant from a policy perspective. It should be noted though that a separate effect on non-promoted teachers may only occur when decreased motivation of non-promoted teachers, if any, translates into increased exits out of the teaching profession. One could wonder whether such a big decision to leave the teaching profession would be taken due to disappointment of missing out on a promotion. In addition, it should be noted that a loss in motivation due to missing out on a promotion, if any, is then expected to occur in both regions, since in both regions a higher share of teachers was placed in a higher salary scale (but an even larger share in the treatment region). It is therefore not clear though, whether disappointment effects, if any, would be significantly different across treatment and control regions.[65]

4.5.3 Sensitivity analysis

In this paragraph we test to what extent our estimates are sensitive to possible selection effects. First we address region switching, as this has been shown to be an issue. Thereafter we will address selection effects that might occur because of unobserved changes in the teacher population composition. Although the tests in the previous paragraph do not hint at a changing teacher population composition, we cannot be fully sure that unobserved characteristics of entering and exiting teachers develop differently over time in the Randstad than in the non-Randstad regions because of the policy. For all our sensitivity analyses we use our local sample and include the full set of controls.

First, we investigate the robustness of our results with respect to switching behavior of teachers between control and treatment regions. Columns (1) and (2) in table 4.5 show results of this sensitivity analysis. In column (1) we select the observations for which we have full information on teacher retention and switching, and run our IV-regression while controlling for (an indicator of) switching. By doing so we lose another year of our data as switching behavior is measured with respect to the previous period. For example, we investigate whether a teacher who switched in 2013 with respect to 2012 is retained in 2014 with respect to 2013. The estimate in column (1) is similar to that in column (4) of table 4.3. Including the switching variable hardly changes the IV-estimate. Switching in itself, however, seems not to be trivial. A teacher who switches between treatment and control region in a particular year has a 9 percentage points higher probability to drop out in the next year. We therefore also run an IV-regression in which we exclude all switchers from our

[65] If anything, we would expect disappointment effects to be larger in the control region because a larger share of teachers in the low salary scale did not receive a promotion in the control region. This would imply that our (zero) estimates on retention are an upper bound of the true effects of a higher teacher pay on teacher retention.

estimation sample in column (2). This estimate is in the same order of magnitude as the previous IV-estimates. Hence our results seem to be robust to switching behavior.
We proceed by addressing possible unobserved changes in the teacher population composition. For our main estimation results we pool the data from 2007-2014, hence we do not distinguish between short and long term effects. However, short-run effects on teacher retention would hint at a changing distribution of teachers, such that estimates of medium-run effects could be biased. Pooling our data as we do in our main specification by including a post-treatment by Randstad interaction dummy would then render biased effect estimates. In columns (3) and (4) we therefore distinguish between short- and medium-run effects by running our IV-regression for years 2010-2011 (short run) and 2012-2014 (medium run) separately. Both short-run and medium-run effects are statistically insignificant and point estimates are (slightly) negative. The absence of short-run effects indicates that our main estimation results do not suffer from possible selection effects.

We continue with a final test on possible selection effects. We take the teacher population from the pre-treatment year 2008 for our local sample of border municipalities and follow this cohort over time. As such, we rule out the risk of selection effects due to a changing teacher population as we keep the estimation sample fixed. We investigate to what extent these 2008-teachers exit the teacher profession for post-treatment years 2009-2014. In columns (5)-(10) we show reduced form estimates of the effect of the Randstad policy on a dummy that equals 1 if the teacher is observed working in the teaching profession in a particular year and 0 otherwise. That is, we look at their retention rate with respect to 2008.[66]

Again, we find no effects of the Randstad policy on this outcome variable. When we perform the same analysis for teacher cohorts from pre-treatment years 2003-2007, the results are similar and never significantly different from zero. Figure 4.5 shows these results graphically. In these graphs we show the development of retention rates for control and treatment group for these cohorts over time, i.e. for the 2003-cohort, 2004-cohort, etc. An advantage of taking a number of years before treatment is that we can investigate whether control and treatment group have the same pre-trend for this outcome variable. It can be observed that the retention rates of control and treatment group almost lie on top of each other and develop rather similarly over time before introduction of the policy. We have also

[66] Hence we look at the probability that a 2008-teacher is in the educational labor market after 1 year (in 2009) , 2 years (in 2010), 3 years (in 2011) etc. Note that this retention rate is different from a survival rate because we allow teachers to reenter the system after a drop out. Note also that we cannot use our IV-strategy when using this outcome variable, as we do not have information on teacher's salary after a teacher drops out.

empirically tested the similarity of the pretends and have found no evidence in favor of deviating trends.[67]

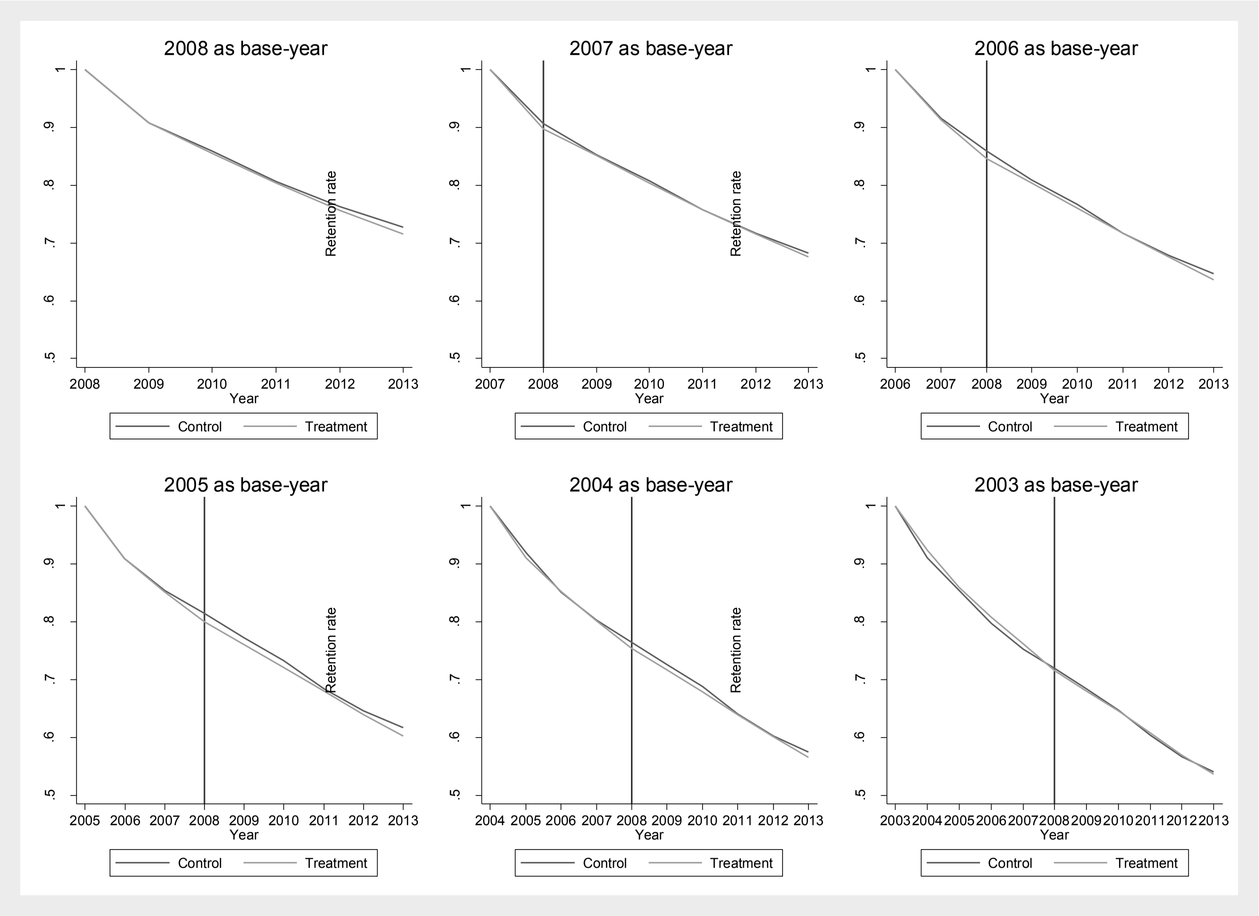
Table 4.5 Estimates of the effect of a higher salary scale (IV) and of teacher pay policy (RF) on retention as a teacher, local sample

	Taking into account region switching		Short term (10-11)	Medium term (12-14)	1 - 6-year retention rate					
	IV	IV	IV	IV	RF	RF	RF	RF	RF	RF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					2009	2010	2011	2012	2013	2014
Effect on retention	-0.019	-0.014	-0.005	-0.021	-0.008	-0.013*	-0.012	-0.014	-0.011	-0.020
	(0.048)	(0.030)	(0.064)	(0.025)	(0.006)	(0.008)	(0.009)	(0.010)	(0.011)	(0.013)
Region switch	-.091***									
	(0.015)									
Reduced form	-0.003	-0.002	-0.000	-0.004						
	(0.008)	(0.005)	(0.006)	(0.005)						
First stage	0.165***	0.162***	0.090***	0.209***						
	(0.011)	(0.011)	(0.012)	(0.013)						
Number of observations	95,379	115,425	67,188	84,010	15,717	15,717	15,717	15,717	15,717	15,717
Controls:										
teacher characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
school characteristics and school-fixed effects	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

Notes: Included teacher covariates are: gender, age category and assignment size category in FTE's. School covariates include school size category, the share of disadvantaged pupils and school population growth. Regressions in columns (1) to (4) include year-fixed effects. Standard errors in parentheses adjusted for clustering at the school level. Significance levels: *** p<1%, ** p<5%, * p<10%.

[67] We performed tests like in table 2. Results are available upon request.

Figure 4.5 Retention rates for different cohorts of teachers



4.5.4 Second stage exclusion restriction

In this section we address the second stage exclusion restriction of our IV-strategy. The increase in the probability of being placed in a higher salary scale should be the only channel through which the Randstad policy may have an impact on teacher retention. Is this the case? Although the Randstad policy is meant to increase the salaries of the teachers, it might be possible that the policy has not been fully used for this purpose. The extra funds have been given to autonomous schools that are in principle free to choose how to spend this additional money. Hence, instead of increasing teacher salaries, schools might spend it (partly) on other activities such as reducing the pupil-teacher ratio. This may be a threat to the exclusion restriction. For example, if the additional funds are used to hire new teachers to reduce the pupil-teacher ratio instead of promoting teachers, then this channel might affect the teacher retention decision. Reductions in class size may cause teachers to stay in the profession as these reductions might render the teaching profession more attractive. In that case the Randstad policy affects the retention decision via class size reductions. The second stage exclusion then fails. To address these types of issues we estimate the effect of the policy on a number of variables that can be considered channels through which the policy might affect the outcome. We estimate the effect of the policy on the pupil-teacher ratio (just discussed), the share of non-teaching personnel, the amount of money saved by the school board (i.e. yield of school board) and the share of school board expenditures not spent on personnel. The last three outcomes may be relevant if the extra funds are not given to teachers but saved or given to non-teaching personnel. Table 4.6 shows results of this analysis. All estimated reduced form effects are close to zero and insignificant. This gives support to the second stage exclusion restriction. Hence, we have no indications that the schools have spent the additional funds to destinations other than placement of teachers in higher salary scales.

Table 4.6 Test on the exclusion restriction: reduced form estimates of the effect of the Randstad policy on various outcomes

	Local sample		Full sample	
Effect Randstad policy on:	(1)	(2)	(3)	(4)
A: Pupil-teacher ratio	0.107	0.160	0.122	0.133
	(0.162)	(0.157)	(0.086)	(0.087)
Number of observations (school-year)	1,274	1,274	5,223	5,223
B: Share of non-teaching personnel	0.003	-0.002	-0.001	0.003
	(0.006)	(0.006)	(0.003)	(0.003)
Number of observations (employee-year)	224,745	224,745	904,902	904,902
Controls				
school-fixed effects	no	yes	no	yes
teacher and school covariates	no	yes	no	yes
C: Yield of school board (in %-points)	-0.750	0.338	-0.601	-0.225
	(-1.360)	(0.748)	(0.620)	(0.521)
Number of observations (board-year)	597	597	2,190	2,190
D: Share of expenses of the school board not spend on personnel	-0.001	-0.000	0.003	0.004
	(0.007)	(0.007)	(0.004)	(0.004)
Number of observations (board-year)	597	597	2,190	2,190
Controls				
School board-fixed effects	no	yes	no	yes

Notes: All regressions include year-fixed effects. Standard errors in parentheses are clustered at the school (panel A and B) or board (panel C and D) level. For pupil-teacher ratio (panel A) and share of non-teaching personnel (panel B) we include the same personnel and school covariates as in table 4.3.

4.5.5 Effects of the policy on teacher's schooling decisions

One of the criteria for being placed in a higher salary scale is that a teacher would obtain extra schooling. The policy should therefore lead to a higher share of teachers being enrolled in additional education. We do not have a direct measure for this outcome. We use the share of teachers that applied for a schooling voucher as a proxy for actual enrollment in degree programs among teachers. This seems to be a reliable proxy since Van der Steeg and Van Elk (2015) show that nine out of ten applicants actually start with the study they applied for. Table 4.7 contains reduced form estimates of the effects of the Randstad policy on this outcome. Columns (1) and (2) show the results for the full sample, columns (3) and (4) for the local sample. The results in the full sample suggest that the Randstad policy increased the probability of applying for a teacher schooling voucher by 0.5 percentage points. The estimates are statistically significant at the 5% level. In the local sample the point estimates are higher (0.9 percentage points), but marginally insignificant (p=0.11). With 2.3 % having applied for a schooling voucher before introduction of the policy, the estimated effect comes down to a 39% (=0.009/0.023) increase in the probability of applying for a schooling voucher, and hence, in enrollment in a bachelor or master study of teacher education. This finding is consistent with the setup of the policy in which one of the criteria for placement in a higher salary scale was that teachers would gain extra qualifications or expertise.

It should be noted though that for this analysis we could not check the plausibility of the second stage independence, i.e. common trend, assumption as we have no data on enrollment in teacher schooling vouchers before 2008. This is because the teacher schooling voucher was not introduced until 2008.

Table 4.7 Reduced-form estimates of the effect of the Randstad policy on the share of teachers that applied for a schooling voucher

	Local sample		Full sample	
	(1)	(2)	(3)	(4)
Effect of Randstad policy	0.009*	0.009	0.005**	0.005**
	(0.006)	(0.006)	(0.002)	(0.002)
Number of observations (school-year combinations)	885	885	3649	3649
Controls				
school-fixed effects	no	yes	no	yes
teacher and school covariates	no	yes	no	yes

Notes: Each column is an OLS-regression. The even columns include the same set of control variables as in table 4.3, except for the fact that teacher covariates have been aggregated at the school level. Standard errors in parentheses adjusted for clustering at the school level. All regressions include year-fixed effects. Significance levels: *** p<1%, ** p<5%, * p<10%.

4.6 Heterogeneous treatment effects

In his section we investigate whether treatment effects for teacher retention differ by teacher's age and gender, and by school's population composition. We study age and gender effects because earlier literature suggests that young teachers (Gilpin, 2011; Hendricks, 2014; Hendricks, 2015) and male teachers (Dolton, 2006) are more sensitive to higher salary with respect to their retention decisions.[68] We study effects by school composition as it has been consistently shown that teachers in schools with a higher share of low-SES (or disadvantaged) pupils are less likely to be retained (e.g. Boyd et al., 2002; Hanushek et al, 2004; Bonhomme et al, 2015). A higher teacher pay might therefore affect these teachers differently than teachers in schools with lower shares of low-SES pupils.

[68] Hendricks (2014) however finds no differences in the sensitivity to higher wages across males and females.

Table 4.8 shows the results of this heterogeneous treatment effects analysis.[69] It shows estimates for three different age categories, for males and females, and for two groups of schools: one with more than 10% pupils from high-poverty areas and one with less than 10% pupils from high-poverty areas. All estimated effects are statistically insignificant and do not significantly differ from each other. Hence, we find no evidence for retention effects for these subgroups.

Table 4.8 Heterogeneous treatment effects of higher salary scale (IV) and Randstad policy (RF) on teacher retention

	Age			Sex		% of pupils from high-poverty areas	
	18-34	35-54	>55	Male	Female	<=10%	>10%
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IV (a)	0.005	0.017	-0.061	0.001	0.017	0.013	0.023
	(0.039)	(0.017)	(0.039)	(0.023)	(0.020)	(0.021)	(0.030)
First stage	0.126***	0.182***	0.140***	0.139***	0.174***	0.162***	0.156***
	(0.011)	(0.009)	(0.009)	(0.007)	(0.009)	(0.010)	(0.010)
RF (b)	0.001	0.003	-0.008	0.000	0.003	0.002	0.004
	(0.005)	(0.003)	(0.005)	(0.003)	(0.003)	(0.003)	(0.005)
Number of observations	113,902	226,294	140,404	245,981	234,619	291,593	189,007

Notes: All models include the same set of controls as in the even columns in table 4.3. Standard errors in parentheses adjusted for clustering at the school level. Significance levels: *** p<1%, ** p<5%, * p<10%.
a) IV indicates effects of being placed in a non-low salary scale.
b) Reduced form indicates estimates of the effect of the teacher pay policy on the outcome of interest.

4.7 Conclusion and discussion

In this chapter we have investigated the effects of higher teacher pay on teacher retention and teacher's schooling decisions in secondary education. We exploited variation in teacher pay induced by the introduction of a new remuneration policy. The policy provided schools in a targeted urbanized region with extra funds to place a higher share of teachers in a higher salary scale. The salaries of these extra promoted teachers increased by approximately 13% in the targeted regions as a consequence of the policy. We used this regional variation in teacher pay in an instrumental variables setup to estimate the effects of being placed in a higher salary scale on our outcomes of interest. The setup of the new remuneration policy allowed us

[69] We use the total sample for this heterogeneity analysis to increase power. Results are similar when we use our local sample but estimates are much less precise due to the smaller sample size when splitting up the local sample by teachers' age sex and SES.

to create similar treatment and control groups by selecting (teachers in) schools around the geographical cutoff that separate treatment and control regions.

Our main findings are as follows. First, and most importantly, we find no effects of higher teacher pay on teacher retention. That is, we do not find that placement in a higher salary scale leads to a higher probability to stay in the teacher profession. Second, we find that the policy led to a small reduction in annual switching from treatment to control regions, but that this does not affect our results found for teacher retention. Hence, the policy succeeded in keeping a slightly larger share of teachers in the targeted region. However, these positive effects are small relative to the costs of the policy. The policy cost on average about 50 mln euro per year. Around 0.4% of teachers per year decided not to switch from the treatment to the control region because of the policy. This would imply a cost of about 400k euro to prevent one teacher from switching from the treatment to the control region.[70] Third, we find that the policy has a positive impact on teachers' enrollment in additional schooling. Our estimates suggest that the policy increased teachers' enrollment in bachelor or master degree programs from 2.3% to 3.2%. This finding is consistent with the setup of the policy in which one of the criteria for placement in a higher salary scale is that teachers would complete extra schooling.

What do we learn from these results? First we discuss why we do not find effects for teacher retention. Although the realized salary increase of approximately 13% (and a prospect of a 17% increase end-of-scale) is by no means small, it may not have been large enough to increase teacher retention rates. This suggests that the retention decision is often motivated by factors other than salary. Studies by Hanushek et al. (2004) and Clotfelter et al. (2011) show that effects of teacher pay on retention are very modest compared to the effect of pupil characteristics. Teachers prefer not to work in schools with high shares of low-SES children (e.g. Boyd et al., 2002; Hanushek et al, 2004; Bonhomme et al., 2015). This suggests that salaries need to be increased substantially in order to increase retention rates of teachers in schools or regions with relatively high shares of low-SES children. It would be interesting to investigate what other policies could be more (cost-) effective. Policies one could think of are better guidance of starting teachers (i.e. induction programs) or opening up and investing in

[70] Calculation is available upon request.

alternative routes to teaching to recruit highly talented people in hard-to-staff schools or regions, such as Teach for America in the US.

Second, we discuss the effects of the new remuneration policy on teachers' enrollment in additional schooling programs and switching behavior. Our findings suggest that offering opportunities to be placed in a higher salary scale can induce the existing teacher workforce to participate in additional schooling, and hence can be used as an incentive to get a better qualified teaching workforce. Furthermore, our results suggest that a higher teacher pay can be used to reduce switching out of a shortage region. This is consistent with the study of Hendrickx (2014) that also finds positive effects of higher teacher pay on retention rates at the regional level.

Taken these findings together, we conclude that a higher teacher pay may not be effective in increasing retention rates in the teaching profession, but might be effective in decreasing turnover rates in specific schools or regions, especially in schools or regions with relatively high shares of disadvantaged pupils.

Appendix

A.1 Data preparation

We took four steps to prepare the data for our analyses. First, we removed teachers that are employed in a Randstad school and in a non-Randstad school at the same time (499 observations, i.e. teacher-year combinations). For these teachers we cannot determine whether they belong to the control or treatment group. Second, we removed teachers for whom we have missing data on age and gender, or whose reported age is lower than 18 (1,882 observations). Imputing missing values on these covariates and including them in the estimation sample does not change results. Third, we removed data on (teachers in) schools in a particular year if more than 50% of the teachers drops out of school in that year (18,236 observations). In that case the school did not (correctly) provide the personnel data to the Ministry of Education (DUO). The Ministry of Education applies this criterion as well before using the data for calculating statistics. Fourth, we imputed missing values for school size and the share of disadvantaged pupils (300 observations).

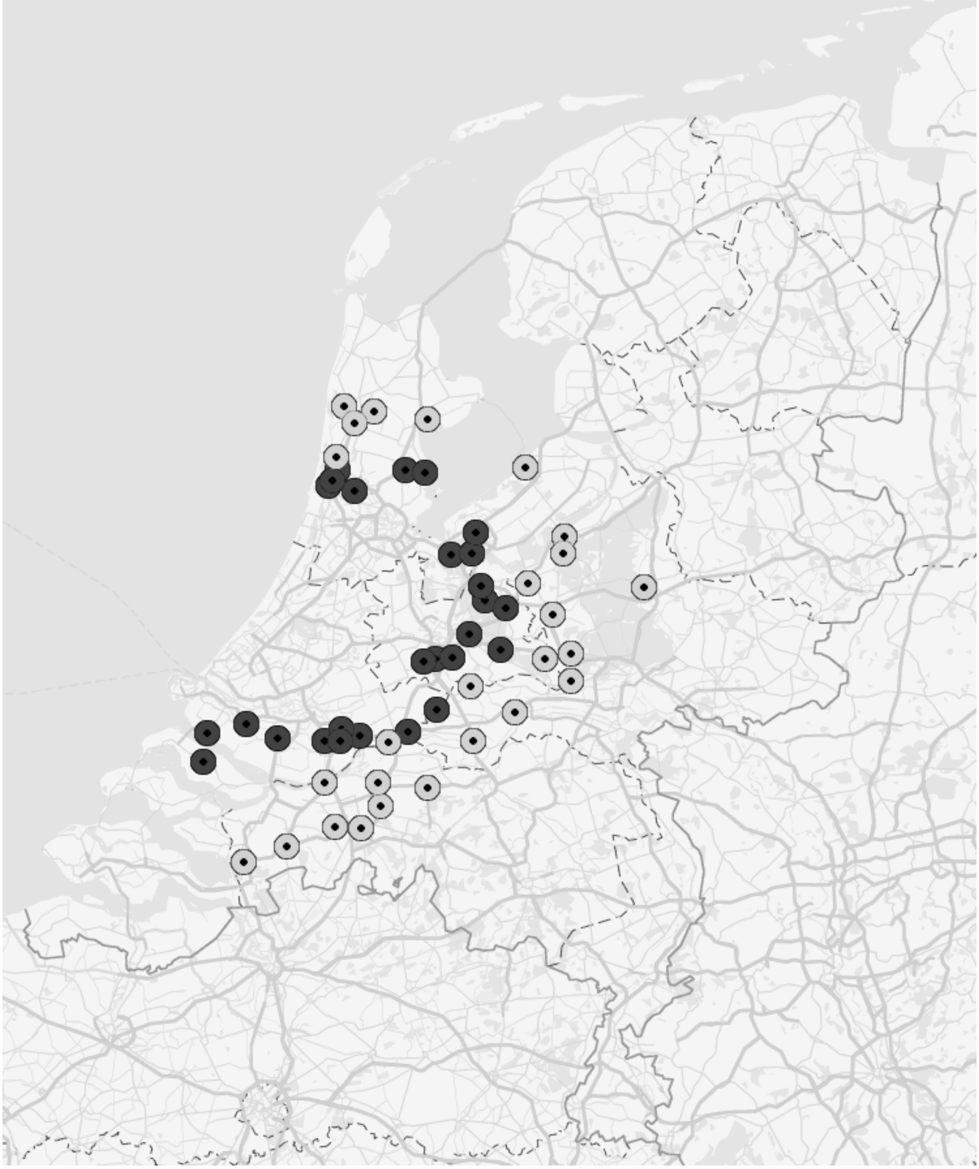
A.2 Construction of local sample

We selected 53 municipalities at the border of the Randstad for our local estimation sample. Table A.1 gives the list of the selected municipalities and figure A.1 provides a map. We selected the first two rings of municipalities around the Randstad border. Taking these two rings was based on the consideration that these municipalities are close to the border and that they would comprise a sample size large enough to estimate effects. Note that the biggest four cities in the Randstad (Amsterdam, Rotterdam, Den Haag and Utrecht) have not been included in the sample as they do not lie at the border.

Table A.1: 53 Municipalities in local sample

Control group	Treatment group
Alkmaar	Almere
Apeldoorn	Amersfoort
Barneveld	Baarn
Bergen (NH.)	Beverwijk
Bergen op Zoom	Dordrecht
Breda	Edam-Volendam
Castricum	Goeree-Overflakkee
Culemborg	Gorinchem
Ede	Heemskerk
Ermelo	Hellevoetsluis
Etten-Leur	Houten
Geertruidenberg	Huizen
Harderwijk	Ijsselstein
Heerhugowaard	Leerdam
Hoorn	Naarden
Lelystad	Nieuwegein
Moerdijk	Nissewaard
Nijkerk	Oud-Beijerland
Oosterhout	Papendrecht
Roosendaal	Purmerend
Tiel	Sliedrecht
Veenendaal	Soest
Waalwijk	Utrechtse Heuvelrug
Wageningen	Velsen
Werkendam	Zaanstad
Zaltbommel	Zeist
	Zwijndrecht

Figure A.1 Selection of 53 border municipalities for the local sample. Bullets in dark (light) grey are treated (control) municipalities.



5.

Does intensive coaching reduce school dropout? Evidence from a randomized experiment[71]

Abstract

School dropout is an important social and economic problem. This chapter investigates the effect of an intensive coaching program aimed at reducing school dropout rates among students aged 16 to 20. Within the coaching program students were offered fulltime support and guidance with their study activities, personal problems and internships in firms. The coaching program lasted one or two years. Students were randomly assigned to classes and the coaching program was randomly assigned to classes as well. We find that one year of coaching reduced school dropout rates by more than 40 percent from 17 to 10 percentage points. The second year of coaching further reduced school dropout by one percentage point. The program is most effective for students with a high ex-ante probability of dropping out, such as students no longer obliged to be in formal education, male students, and students not living with both parents. Cost-benefit analysis suggests that one year of coaching is likely to yield a net social gain.

[71] This is joint work with Roel van Elk and Dinand Webbink: Steeg, M. van der, R. van Elk, and D. Webbink, 2015, Does intensive coaching reduce school dropout? Evidence from a randomized experiment, *Economics of Education Review*, vol. 48, October 2015, pp. 184-197.

5.1 Introduction

Dropping out of school is an important social and economic problem in many countries. A large literature documents the benefits of education, for instance higher wages (Card, 1999; Harmon et al., 2003; Heckman et al., 2006), better health (Oreopoulos, 2007; Lleras-Muney, 2005), less participation in crime (Lochner & Moretti, 2004; Machin et al., 2012;), and a higher intergenerational transfer of human capital (Oreopoulos et al., 2006). However, in many countries the proportion of students that do not finish their education remains high, in particular their secondary education (OECD, 2012). Not completing their education will reduce the future prospects of students, especially for students with a low level of completed education, and might induce costs for society at large. The problem of school dropout is not new; schools and policy makers have long been concerned with high dropout rates and have actively searched for interventions or programs to increase graduation rates. In the recent literature two approaches aimed at reducing school dropout seem most promising. First, financial incentives for students (e.g. Dearden et al., 2009) or conditional cash transfers (e.g. Schultz, 2004; Attanasio et al., 2010) have been shown to reduce school dropout or to increase enrolment. The second approach, which is the focus of this chapter, is to use coaches that give intensive personal attention and support to students at risk.

Intensive coaching or mentoring programs appear to be able to reduce school dropout rates and/or improve educational progress and attainment among adolescents. For instance, positive results have been reported from the Big Brothers/Big Sisters program (Grossman and Tierney, 1998; Herrera et al., 2007), Sponsor-A-Scholar program (Johnson, 1999), the Check-and-Connect program (Sinclair et al., 1998; Sinclair et al., 2005), the Quantum Opportunities Program (Schirm et al., 2006; Rodriguez-Planaz, 2012a[72]) and the Pathways to Education program (Oreopoulos et al., 2014). In addition, an evaluation of twenty dropout prevention programs in the United States showed promising results of programs characterized by an intensive and personal approach in smaller groups (Dynarski et al. 1998). Carneiro & Heckman (2003) review a number of evaluations of dropout prevention programs in the United States. They conclude that sustained interventions targeted at adolescents still enrolled in school can positively impact learning and subsequent employment and earnings, but that interventions targeted at dropouts seem less successful. The National Guard Youth

[72] Rodríguez-Planas (2012a) found modest average long-term effects of the Quantum Opportunity Program on educational outcomes, with shorter-term effects being more impressive.

ChalleNGe Program, which includes a mentoring program, also appears to be effective (Millenky et al., 2010). Bettinger and Baker (2013) find positive effects of the InsideTrack coaching program for college students on the probability of staying in college.

This chapter focuses on an intensive coaching program aimed at reducing school dropout of students aged between 16 and 20 in secondary (vocational) education in the Netherlands. The coaching program included a range of preventive activities such as working on study skills (e.g., planning and organizing), counseling in case of personal problems and contacts with parents. The coaches had extensive educational experience and were highly trained. They monitored the students closely through intake sessions, home visits, observations of behavior and attendance in class and visits during internships. Students received support and guidance with their study activities, with internships in firms, and with personal problems. On average one fulltime coach was assigned to a class of twenty students. Students within five vocational courses were randomly assigned to classes that received the coaching program and to classes that received care as usual. The random assignment of students enables us to identify the causal effect of the program. Our study focuses on two cohorts of students. The first cohort received two years of coaching whereas the second cohort received one year of coaching.

Our main finding is that the intensive coaching program has a large effect on school dropout. One year of coaching reduces the school dropout rate by more than forty percent, that is, from 17 to 10 percent. The estimated effect after two years of coaching is slightly larger. We find larger effects for students with a higher ex-ante probability of school dropout: male students, students not living with both parents, and students above the compulsory school-leaving age. Tentative cost-benefit calculations suggest that one year of intensive coaching yields a net social gain whereas two years of coaching probably does not. The internal rate of return of one year of coaching is calculated at 6.9 percent, whereas that of two years of coaching is calculated at 3.7 percent. Targeting the program towards student with a high ex-ante probability of dropping out and towards the first year of the vocational course is expected to improve the cost-effectiveness of the program.

Our study contributes to the literature on school dropout prevention interventions in secondary education by adding rigorous evidence about a high quality intervention that seems widely applicable. The coaching program investigated in this study shares several elements with mentoring programs studied in the literature, such as assignment of a coach/mentor with a strong personal and supportive approach, and a focus on student-coach interactions and activities for students still enrolled in school. However, the high quality and intensity of the program, as indicated by the educational experience and level of educational attainment of the coaches, the student/coach ratio, the full-time availability of a coach, and the broad range of interventions, seem different from previous rigorous courses. In addition, the context, timing and target group of this program is also different. While previous courses mainly studied interventions at middle or high school level, this program focused on students with an average age of 18 years starting in intermediate post-secondary vocational education. These students had just made a transition towards a new vocational course. The target group of students was the general population of students , whereas previous courses mostly focused on students with disadvantaged or lower socioeconomic backgrounds.[73] The target group of students in the Dutch program includes students both under and above the statutory school-leaving age. This enables a comparison of program effects by compulsory schooling status. In addition, most courses have investigated US programs whereas this study has a European context.

This chapter proceeds as follows. Section 5.2 gives a description of the coaching program. Section 5.3 presents the setup of the experiment. Section 5.4 presents the empirical strategy, whilst Section 5.5 describes the data. Section 5.6 shows the effects of one year in the program on school dropout, followed in section 5.7 by the effects of two year of coaching on school dropout and degree completion. Section 5.8 presents the tentative cost-benefit analyses of one and two years of intensive coaching. Section 5.9 concludes and gives a brief discussion of the main results. Appendix A provides further information about the Dutch context and the background of the experiment. Appendix B gives summary statistics for the first of two cohorts, whereas Appendix C provides more details on the cost-benefit analyses.

[73] The average school dropout rates are lower in our experiment than in previous mentoring courses in the ‘care-as-usual’ situation. For instance, the school dropout rates in the US Quantum Opportunities Program were about 50 percent and in the Education Maintenance Allowance control areas 36 percent, whereas the dropout rate in the Dutch coaching experiment was less than 20 percent.

5.2 The coaching program

The coaching program consisted of various types of interventions, both preventive and after students dropped out from a particular vocational program (i.e. ‘curative’ interventions). The following preventive interventions were part of the coaching program:

- Intake sessions with all students aimed at getting to know each other, detecting personal and/or educational problems and to make follow-up arrangements for various tracks. Different guidance tracks were initiated, for example with respect to dyslexia, fear of failure, social skills, self-confidence, or study skills. Coaches also gave guidance in case of financial problems or problems with housing.
- A home visit in the first month of the new educational program in order to get to know the parents or guardians and reduce the ‘social distance’ between home and school. Later on contacts with parents or guardians were possible if needed.
- Instruction on and help with study planning and organization with a focus on stimulating self-reliance.
- The coaches regularly attended lessons to observe the students and to give them study support if needed after the lessons. The coaches informed other school teams regularly in formal team meetings and helped other coaches by sharing succesful initiatives.
- The coaches visited the students at their internship/apprenticeship with the aim of observing problems with work or social skills, and if needed initiate extra training for improving these skills. The coaches also played an active role in obtaining a good match between the student and the company at which the internship took place;
- In case of absence from school the coach immediately contacted the student and/or parents to discuss the reasons for not attending classes. If needed, the coaches implemented action plans to prevent further absence from school.

The ‘curative’ interventions were used when it was likely that the student would dropout from the particular vocational course. These actions aimed to guide the student to another vocational course by setting up an intensive track to help them choose the right course. This track consisted of talks, testing, guidance to another vocational course and checking whether the student had been accepted and actually started in the new vocational course. All the above interventions were carried out by two part-time coaches per class, adding up to one full-time equivalent available per class. This is equal to 40 hours per week. Only one experimental group had one full-time coach instead of two part-time coaches. The coaches had on average 18 years of experience in education, of which 8 years at the school where the experiment took place. All coaches except one had obtained a higher education degree. Almost 60 percent of the coaches were teachers before they started their job as a coach in the experiment.

A local project coordinator had the task of implementing the assignment of students to classes and of communicating the 'rules' of the experiment in the participating courses, of organizing data collection and delivery, and of monitoring the experiment. This coordinator also organized regular meetings in which coaches could discuss cases with each other, and in which particular themes were addressed aimed at improving the expertise of the coaches. These meetings ensured that the coaches worked with the same vision and set of interventions across the different vocational courses.

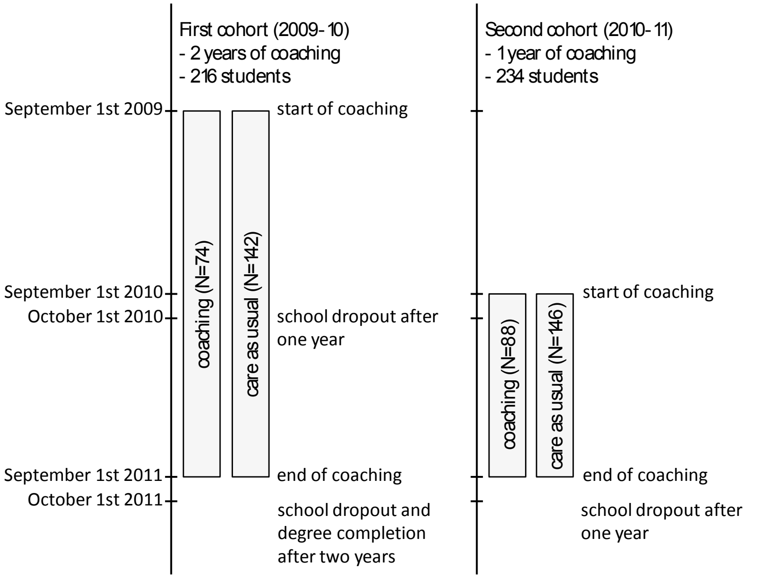
The coaching experiment was funded by the Dutch Ministry of Education at a total intervention cost of € 720,000. These costs consisted of € 60,000 per full-time equivalent of coaching per class per year, or € 3,000 per student per year.

5.3 Design of the coaching experiment

The experiment focused on students in Dutch intermediate post-secondary vocational education aged 16 to 20. Figure 5.1 shows the timing and design of the coaching experiment. The experiment took place in a school for intermediate post-secondary vocational education in Arnhem, a medium-sized city in the Netherlands. The experiment started in the school year 2009-2010 with a first cohort of students receiving two years of coaching. The nominal duration of the educational program was also two years. The first cohort was followed by a second cohort receiving one year of coaching. In total 450 students participated in the experiment. The experiment was implemented at level 2 of intermediate vocational education, which is equal to ISCED level 3. Appendix A provides more information on the context of the experiment.

Students were randomly assigned to an experimental or a control group. The experimental classes were offered the intensive coaching program provided by a full-time equivalent of coaching per class (in most cases two part-time coaches per class), whereas the control classes received 'care as usual'. It should be noted that the program was not compulsory; students could received the support of the coach but participation in the various activities was not compulsory. The 'care as usual' consisted mainly of a dropout desk that advises students after they have dropped out. Five vocational courses participated in the experiment: health care, hairdressing, cooking, security and sales. Average degree completion rates of the participating courses were similar to national average completion rates of these subjects.

Figure 5.1 Design of coaching experiment, outcome measures and measurement moments



The complete list of applicants of the five vocational courses was used for the random assignment of individual students within each course. In total 23 classes participated in the experiment: eight classes received the experimental treatment and fifteen classes received ‘care as usual’. Applicants were randomly assigned to classes within each of the five course types. One class within every vocational course was then randomly assigned intensive coaching, the other class or classes (depending on the total number of classes) received care as usual. The selection of one treatment class per vocational program per cohort was due to budget constraints. The randomization was carried out just before the start of the school year. At that time students were not informed about the project and the assignment to classes. The program was announced to students in the treatment classes and parents just after the start of the school year. The students in the control classes were not informed about the coaching program. The timing of the announcement implies that the coaching program could not have affected the choice of course. This was also the case for the second cohort, since the decision to extend the experiment with a second cohort has been made very late (i.e. in the summer of 2010). This implies that it is unlikely that the introduction of the coaching program for a second cohort could have been affected students or their parents in their choices. Students in both cohorts have not been moved between classes after the randomization took place, which was in line with our instructions.

5.4 Empirical strategy

For investigating the impact of the coaching program on school dropout rates we estimate OLS regressions of the following form:

$$(1) \quad DROPOUT_{ics} = \beta_o + \beta_1 Coaching_{cs} + \beta_2 X_{ics} + \beta_3 Cohort + \alpha_s + \varepsilon_{ics}$$

where $DROPOUTi_{cs}$ is a binary variable which takes the value 1 if student *i* from class *c* of vocational course *s* dropped out and 0 otherwise, $Coaching_{cs}$ is a dummy variable indicating whether class *c* in vocational course *s* received the offer of the coaching program, X_{ics} is a vector of observable characteristics of the student[74], Cohort is a dummy indicating the student cohort, α_s is a fixed effect for the vocational course and ε_{ics} is an error term. The coefficient β_1 can be interpreted as the causal effect of the program on school dropout because the treatment was randomly assigned among the students.

As in any experimental or quasi-experimental design there are deviations from the ideal experimental design that might bias the estimated effects. A first concern is that not all applicants that were assigned to a treatment or control group actually started the vocational course they applied for. This could bias our estimates if the decision to actually start could be affected by the treatment status. However, due to the design of the experiment this is not possible. Students were informed about the coaching program only after the start of the school year so the decision not to start in the course they had applied for was made before the announcement of the coaching program. The non-starting of students can be considered as an example of sample selection within an experimental setting. Lee (2002, 2009) has introduced an approach for obtaining sharp lower and upper bounds for average treatment effects in the presence of sampling bias within such a setting. This approach is based on the assumption

[74] The sample mean by cohort has been imputed for the few students with missing values on certain covariates and a dummy 'missing' has been set to 1.

that all individuals in the control group would also be observed in the treatment group if their treatment status would change (the monotonicity assumption). Hence, the treatment assignment can only affect sample selection in one direction. If this assumption holds we might observe more individuals in the treatment group but sharp lower and upper bounds can be obtained by trimming the treatment group with the proportion of excess individuals. In our setting this seems a weak assumption because the starting decision was made before the information about the treatment assignment became available. Hence, it is highly likely that the individuals in the control group would also have started their study if they had been assigned to the treatment group. Lee (2009) shows that in the case of continuous outcomes upper bounds are obtained after trimming the lower tail of the outcome distribution; lower bounds are obtained by trimming the upper tail of the outcome distribution. Lee (2002) derives bounds in the case of binary outcomes without covariates. To assess the possible bias due to non-starting we have calculated lower and upper bound estimates of the treatment effect following Lee (2002). For obtaining estimates for the models with covariates we use the conditional means for the treatment and control group instead of the unconditional means as in Lee (2002). The standard errors are based on the analytic standard errors provided in Lee (2009). It should be noted that this issue of non-starting differs from the more standard issue of non-compliance. Individuals that did not start in the course they applied for did not receive the offer of the coaching treatment and did not refuse this offer, as is the case of non-compliance. As a result we do not observe their performance in their vocational course during this experiment. Therefore, we cannot address the issue of non-starting with the usual instrumental variable approach that estimates intention-to-treat or treatment-on-the-treated effects.

The second concern in our experimental design is spillover effects from students in the experimental groups to students in the control groups. We expect that these spillovers are unlikely because experimental and control groups had their own schedule and interactions among students took place mostly within their class and not across classes. In addition, we expect that spillovers, if any, from the experimental to the control group students would probably reduce the probability of dropout for students in the control group (i.e. positive spillovers). In that case our estimates should be considered as lower bounds of the true effects.

Our data consists of two cohorts of students that have received the coaching program or not. The first cohort received two years of coaching, the second cohort received one year of coaching. We will separately analyze the effects of one year of coaching and the effects of two years of coaching.

5.5 Data description

The data come from four sources. Data on school dropout and previous highest attained education have been collected from a national database called BRON that includes information on the school careers of all Dutch students. Data on dropout from a particular vocational course and on certain student background characteristics are collected from the school's central administration. In addition, data was used from intake tests among applicants taken before the start of the school year. We also used data from a student survey that was carried out just after the start of the school year, for instance on the degree of self-reported personal problems in several domains.

Dependent variables

The main dependent variables are school dropout, switching to another vocational course and having obtained a so-called 'start qualification'. School dropout, our main dependent variable, is defined as having left education without having obtained a so-called start qualification. A start qualification is comparable to having finished a degree at ISCED level 3, and is considered to be the minimum necessary qualification level for successful entry to the labor market in the Netherlands. The national student database BRON registers whether a student is in education every year on the first of October. The database contains relevant information about the school, the study level and particular vocational course of the student. . School dropout is a dummy variable that has been taken from this database. This variable is also used in the national, regional and school statistics on school dropout that are produced by the Ministry of Education. School dropout is only measured once a year by comparing the situation at the first of October of a given year with the situation of the same student one year before. This implies that we do not know the exact timing of school dropout during the school year. Switching to another vocational course is defined as having left the education program in which the student started without having graduated for that course and subsequently being enrolled in another vocational course at the time of measurement (first of October). Having obtained a 'start qualification' is measured two years after the start. This time span corresponds to the nominal duration of two years at level 2 of intermediate post-secondary vocational education, which is the level at which the experiment took place.

Covariates

As covariates in our regression analyses we employ a rich set of student background characteristics and information about the personal situation and cognitive level of the students at the start of the experiment. Student background characteristics are gender, a dummy stating whether the student was born in the Netherlands or not, highest previous attained education (containing six categories) and age at the start of the experiment. From the age variable we also derive a dummy variable indicating whether the student is legally obliged to be in formal education until the end of the first year of the experiment. Every student under the age of 18 that has not yet obtained a start qualification has to go to school and be enrolled in a study that leads to a start qualification. Information about the cognitive level of the student is obtained from intake tests from the start of the school year. These intake tests consist of tests in numerical and verbal skills. Both types of skills are measured on a scale of 1 to 5. Two indicators provide relevant information about the personal situation of the students. The first is a dummy variable indicating whether the student lives with both parents or not. The other indicator is a dummy indicating whether the student has personal problems to some degree in at least one of the following four areas that may hinder them in their educational program: financial situation, contacts with police and/or justice, housing, and family and friends. This information is self-reported from a survey that was carried out just after the start of the experiment among all participating students. This student survey also yields a dummy variable indicating whether the student decided early or late (before or after 1 July) to enroll in the particular vocational course.

Table 5.1 shows the number of students that participated in the experiment by treatment status for the pooled sample and, separately, by cohort. The total list of applicants of the five participating courses in the experiment consisted of 503 students. Approximately 10 percent of all students did not start the vocational course they applied for. Most of these ‘non-starters’ chose a different vocational course within the same school or at another school. The proportion of starters in the participating courses is somewhat larger in the treated groups than in the control groups but the difference is statistically not significant. We will address the possible bias due to non-starting in Section 5.4.

Table 5.1 **Applicants and starters in participating courses by assignment status.**

	Control Group	Treatment Group	Total
A) Pooled sample (2 cohorts)			
Applicants	327	176	503
Starters	288	162	450
(% of applicants)	(88%)	(92%)	(89%)
B) First cohort (2009-10 cohort)			
Applicants	166	81	247
Starters	142	74	216
(% of applicants)	(86%)	(91%)	(87%)
C) Second cohort (2010-11 cohort)			
Applicants	161	95	256
Starters	146	88	234
(% of applicants)	(91%)	(93%)	(91%)

Table 5.2 presents sample means by treatment status for the five vocational courses that participate in the experiment. The sample consists of students that actually started the vocational courses they applied for. The table shows that in all five courses the treatment and control groups are quite similar on a broad range of student characteristics. Only two out of sixty differences in average characteristics within the five participating courses are statistically significant. Hence, the randomization produced similar groups within each vocational course and cohort. For the pooled sample we find no statistically significant differences is student characteristics after controlling for vocational course and cohort (within which randomization took place).

The lower part of table 2 gives a first impression of the effect of the treatment on school dropout and switching to other courses after one year. We observe that 17 percent of the students in the control group dropped out of school whereas 7 percent of the students in the treatment group dropped out. The difference in the proportion of students that switched to another vocational course is quite small (21 versus 19 percent).

Table 5.2 Sample statistics of treatment and control groups by vocational course and for pooled sample.

	1) Health care (nursing)			2) Hair dressing		
	Control	Treated	p-value[a]	Control	Treated	p-value[a]
Background characteristics						
1. Male	0.09	0.08	*0.72*	0.13	0.02	*0.05*
2. Age (in years)	18.7	18.8	*0.88*	17.9	17.7	*0.64*
3. Obliged to be in formal education[b]	0.33	0.37	*0.70*	0.50	0.61	*0.25*
4. Born in the Netherlands	0.90	0.90	*0.88*	0.92	0.89	*0.72*
5. Living with both parents	0.50	0.42	*0.43*	0.50	0.64	*0.15*
6. Having problems in at least one of the following areas: finance, police and justice, family and friends, or living/housing situation	0.45	0.30	*0.13*	0.37	0.41	*0.92*
7. Score on verbal skills at intake test (1-5)	3.2	3.2	*0.98*	3.2	3.5	*0.12*
8. Score on numeric skills at intake test (1-5)	2.8	2.7	*0.44*	3.1	2.9	*0.26*
9. Highest previous attained degree (1-6)	2.4	2.4	*1.00*	2.4	2.4	*0.84*
10. Already obtained a start qualification before the start	0.02	0.03	*0.68*	0.08	0.04	*0.35*
11. Late study choice (July or later)	0.28	0.29	*0.91*	0.18	0.27	*0.21*
12. Average class size (of started students)	19.2	19.0	*0.61*	19.8	23.0	*0.27*
Outcome variables (after one year)						
13.a. School dropout[e]	0.18	0.08	*0.08*	0.16	0.07	*0.14*
13.b. Switch to another study	0.22	0.18	*0.66*	0.25	0.22	*0.57*
13.c. Still in same study	0.60	0.74	*0.17*	0.59	0.72	*0.16*
Number of classes	6	2	8	5	2	7
Number of observations	115	38	153	99	46	145

	3) Cook and catering			4) Security[c]		
	Control	Treated	p-value[a]	Control	Treated	p-value
Student characteristics						
1. Male	0.86	0.87	*0.95*	0.81	0.79	*0.88*
2. Age (in years)	17.8	17.6	*0.55*	18.6	18.0	*0.23*
3. Obliged to be in formal education[b]	0.57	0.50	*0.49*	0.33	0.46	*0.40*
4. Born in the Netherlands	1.00	0.95	*0.17*	0.95	0.96	*0.93*
5. Living with both parents	0.83	0.71	*0.17*	0.38	0.54	*0.30*
6. Having problems in at least one of the following areas: finance, police and justice, family and friends, or living/housing situation	0.31	0.50	*0.12*	0.67	0.48	*0.22*
7. Score on verbal skills at intake test (1-5)	3.4	3.8	*0.14*	4.0	3.8	*0.21*
8. Score on numeric skills at intake test (1-5)	3.9	3.9	*0.68*	3.6	3.6	*0.97*
9. Highest previous attained degree (1-6)	2.5	2.3	*0.49*	2.6	2.6	*0.96*
10. Already obtained a start qualification before the start	0.08	0.03	*0.33*	0.05	0.00	*0.33*
11. Late study choice (July or later)	0.25	0.18	*0.52*	0.24	0.26	*0.87*
12. Average class size (of started students)	18.5	19.0	*0.33*	21.0	24.0	
Outcome variables (after one year)						
13.a. School dropout[e]	0.11	0.03	*0.15*	0.33	0.13	*0.10*
13.b. Switch to another study	0.16	0.21	*0.63*	0.00	0.08	*0.18*
13.c. Still in same study	0.73	0.76	*0.87*	0.67	0.79	*0.83*
Number of classes	2	2	4	1	1	2
Observations	37	38	75	21	24	45

Table 5.2 (continued)

	5) Sales[c]			All courses pooled		
	Control	Treated	p-value[a]	Control	Treated	p-value[d]
Student characteristics						
1. Male	0.75	0.75	*1.00*	0.29	0.42	*0.23*
2. Age (in years)	17.7	17.5	*0.54*	18.3	18.0	*0.39*
3. Obliged to be in formal education[b]	0.56	0.81	*0.14*	0.43	0.52	*0.17*
4. Born in the Netherlands	0.73	0.94	*0.14*	0.91	0.92	*0.91*
5. Living with both parents	0.75	0.50	*0.15*	0.55	0.58	*0.90*
6. Having problems in at least one of the following areas: finance, police and justice, family and friends, or living/housing situation	0.27	0.19	*0.61*	0.41	0.39	*0.63*
7. Score on verbal skills at intake test (1-5)	2.7	2.8	*0.83*	3.3	3.5	*0.19*
8. Score on numeric skills at intake test (1-5)	3.1	2.9	*0.42*	3.1	3.2	*0.39*
9. Highest previous attained degree (1-6)	1.7	2.4	*0.05*	2.4	2.4	*0.54*
10. Already obtained a start qualification before the start of the experiment	0.00	0.07	*0.33*	0.05	0.03	*0.31*
11. Late study choice (July or later)	0.40	0.50	*0.60*	0.25	0.28	*0.51*
12. Average class size (of started students)	16.0	16.0		19.2	20.3	*0.02*
Outcome variables (after one year)						
13.a.School dropout[e]	0.00	0.06	*0.33*	0.17	0.07	*0.00*
13.b. Switch to another study	0.38	0.25	*0.46*	0.21	0.19	*0.54*
13.c. Still in same study	0.63	0.69	*0.72*	0.62	0.74	*0.19*
Number of classes	1	1	2	15	8	23
Observations	16	16	32	288	162	450

Notes:

(a) Controlling for cohort.

(b) All students under 16 are obliged to go to school in any case. Students of 16 and 17 are obliged to be enrolled in formal education if they have not completed a degree that counts as a 'start qualification' (i.e. ISCED level 3 or higher).

(c) Security (second cohort) and Sales (first cohort) have only been sampled in one cohort.

(d) Controlling for cohort and vocational course.

(e) School dropout is defined as having left education without having obtained a start qualification (i.e. ISCED level 3 or higher).

5.6 The effect after one year of coaching

The estimated effects of one year of coaching are shown in Table 5.3. Panel A shows the estimated effects on school dropout based on linear probability models in which school dropout is regressed on coaching using different sets of control variables. Column (1) controls for cohort and vocational course, column (2) also controls for socioeconomic and personal characteristics, and column (3) includes controls for previous schooling and cognitive skills. Columns (4) and (5) show the results for the first or second cohort respectively. Standard errors have been corrected for clustering at the class level. Panel B distinguishes between switching to another vocational course and school drop-out, and shows the marginal effects of one year of coaching. The estimates are based on multinomial logit models using the same specifications as in panel A.

Table 5.3 Estimates of the effect after one year on school dropout

	(1)	(2)	(3)	(4)	(5)
Panel A: Linear probability models					
Coaching on school dropout	-0.096***	-0.073***	-0.071***	-0.065**	-0.082***
	(0.023)	(0.021)	(0.018)	(0.028)	(0.022)
Panel B: Multinomial logit model					
(Reference category = 'same study')					
Coaching on school dropout	-0.097***	-0.075***	-0.072***	-0.076***	-0.087***
	(0.019)	(0.019)	(0.018)	(0.024)	(0.027)
Coaching on switching	-0.012	-0.014	-0.017	-0.025	-0.008
	(0.042)	(0.041)	(0.040)	(0.049)	(0.053)
Sample	pooled	pooled	pooled	cohort 1	cohort 2
Controls					
Cohort and vocational course	yes	yes	yes	yes	yes
Socioeconomic and personal factors	no	yes	yes	yes	yes
Previous education and cognitive skills	no	no	yes	yes	yes
Observations	450	450	450	216	234

Notes: Column one controls for cohort and type of study. Column (2) also includes controls for gender, born in the Netherlands, compulsory education status, living with both parents, timing of study choice, having problems. Column (3) also includes highest level of education attained, verbal and numeric skills. Panel B shows marginal effects from multinomial logit models. Standard errors corrected for clustering at the class level in parentheses. * / ** / *** significant a 1, 5 or 10 %-level.

The estimates in columns (1) to (3) of panel A show that one year of coaching reduces the probability of school dropout between 7.1 and 9.6% points. Including controls lowers the estimated effect towards 7.1% points. From a base level of 17% this corresponds to a reduction of more than 40%. The estimation sample consists of students that started in their courses. A concern with the pooling of the data in columns (1) to (3) may be that the first cohort was offered a program of two years and the second cohort a program of one year. In addition, effects may differ between cohorts because of a difference in experience with running the program. Therefore, we have also estimated the effects for the two cohorts separately. Results are shown in columns (4) and (5). The estimated effects on school dropout turn out to be robust to cohort, but slightly larger for the second cohort.

As mentioned in the previous sections, this sample deviates from the original assignment sample due to students that did not start; 12% of the students in the control group and 8% in the treatment group did not start (Table 1). The estimate of the treatment effect might be biased due to the higher proportion of students in the control group that did not start. Because non-starting differs from non-compliance, we cannot apply an instrumental variable approach (see section 5.2.3). To assess this possible bias we calculated bounds of the treatment effect based on Lee (2002 and 2009).[75] For the first model of Table 3 we find a lower bound of the treatment effect of 0.093 (0.024), and an upper bound of 0.138 (0.057). For the model of column (3) with all covariates the lower bound estimate is 0.067 (0.020) and the upper bound is 0.112 (0.056). It should be noted that bias due to non-starting is unlikely because the coaching program was announced after the students had made their decision to start or not in the specific vocational course.

The coaching program includes both preventive action, aimed at reducing school dropout, and curative actions, aimed at guiding students to other courses when it was likely that the students would drop out of their current course. To analyze the importance of these two types of action we further investigated whether coaching not only has an effect on school dropout but also on switching to other courses. We have estimated multinomial logit models in which the dependent variable has three categories: same vocational course, switching to another vocational course and school dropout. Panel B of Table 5.3 shows the marginal effects of one year of coaching; the category ‘same vocational course’ is the reference category. The multinomial logit estimates show that one year of coaching reduces school dropout with 7% points but has no effect on switching to other vocational courses. Hence, the estimates in panel A and B consistently show that coaching significantly reduces school dropout. The results also indicate that the coaching program is successful in preventing school dropout but do not seem successful with respect to the curative actions aimed at guiding potential dropouts towards other courses. An explanation for the latter finding might be that the ‘care as usual’ received by the control group mainly consists of curative actions. These might have the same effects as the curative actions undertaken in the coaching program.

[75] The trimming proportion is 4 percent, which equals the difference in non-starting between the treatment and control group.

Heterogeneous effects of one year of coaching on school dropout

The effect of coaching may differ among subgroups of students. We investigated this issue by interacting the treatment variable with specific covariates. Table 5.4 shows the estimated effect for specific subgroups using the full sample of students. For instance, row (1a) shows the treatment effect for males and row (1b) shows the difference in the treatment effect between males and females. We observe that the treatment effect for male students is larger than for females. The difference in the treatment effect is 6 percentage points, which is nearly statistically significant (p-value of 0.13). In addition, we find that the treatment effect is 7 percentage points larger for students that are no longer obliged to be in formal education (see row 2b). Moreover, the treatment effect appears to be 8 percentage points larger for students that are not living with their parents (see row 3b). This difference is again nearly statistically significant (p-value of 0.13). These results suggest that the treatment effect is larger for subgroups with a higher ex-ante probability of dropping out of school as can be observed in the last column of Table 5.4. Hence, coaching seems more effective for groups with a relatively large probability of dropping out. We further investigated this by estimating a probit regression that predicts the probability of dropping out of school as a function of individual covariates, vocational course, and cohort. This regression is estimated on the control sample only and is used to generate ex-ante school dropout probabilities for both treated and non-treated students. The fitted probabilities are used to split the sample into two subgroups of roughly equal size, one with a relatively low probability of dropping out of school, and one with a relatively high probability (top half). Students in the top half have an average ex-ante school dropout probability of 28%, whereas students in the bottom half have an ex-ante school dropout probability of 3%. Next, we constructed a dummy variable that distinguishes students in the bottom half of the school dropout probability distribution from students in the top half and estimated the effect of the interaction of this variable with the treatment. Row (4a) shows that the treatment effect for students with a high ex-ante dropout probability is approximately 13 percentage points. The estimate of the interaction effect (see row 4b) shows that the treatment is much less effective for students with a low ex-ante dropout probability, the difference in the estimated effect is statistically significant. This finding is consistent with Rodríguez-Planas (2012b), who finds that the Quantum-Opportunity Program is ‘extremely successful’ in improving educational and behavioral outcomes for those most at risk.[76] Our results suggest that the efficiency of the coaching program may be improved by targeting the coaching interventions on groups with a higher ex-ante probability of dropping out.

Table 5.4 Heterogeneous effects of one year of coaching on school dropout by different student characteristics, pooled sample of both cohorts

	Coefficient	Dropout in control group for relevant subgroup
1.By gender		
1.a. Male	-0.111***	0.238
	(0.037)	
1.b. Interaction term treatment * female	0.063	0.137
	(0.040)	
2. By compulsory schooling status		
2.a. No longer obliged to stay in formal education	-0.107***	0.227
	(0.030)	
2.b. Interaction term treatment * obliged to stay in formal education	0.071*	0.089
	(0.040)	
3. By living situation		
3.a. Not living with both parents	-0.107**	0.236
	(0.041)	
3.b. Interaction term treatment * living with both parents	0.082	0.067
	(0.052)	
4. By ex-ante probability of school dropout		
4.a. High predicted school dropout probability (top half)	-0.129***	0.278
	(0.035)	
4.b. Interaction term treatment * low predicted dropout probability (bottom half)	0.114**	0.032
	(0.044)	

Notes: The estimates are derived from regressions including a treatment dummy and an interaction term for the denoted subgroup with the treatment dummy. All models include the complete set of covariates as in column (3) of Table 5.4. Standard errors are corrected for clustering at the class level. ***/**/* denotes effects are significant at a 10/5/1 percent significance level.

Recent research suggests that the development of boys is more sensitive to unstable home environments. For instance, Bertrand et al. (2013) find that disruptive behavior of boys is associated with one parent families. We find a similar pattern in our data. We observe that for

[76] Rodríguez-Planas (2012b) identifies students most at risk as the students in the top-half of the predicted drug use distribution. High school graduation increased by 20 percent for this group and college enrollment increased by 28 percent due to QOP.

students not living with both parents school dropout occurs much more often among male students than among female students (i.e. 43% versus 17%). In this group male students report twice more often having problems with police and drugs than female students, and one third more often financial problems and problems with housing. Reporting (one of) these problems is associated with a higher probability of school dropout. Moreover, for this group we find that coaching is much more effective for male students than for female students. The estimated effect is -0.29 (with a standard error of 0.12) for male students versus -0.04 (with a standard error of 0.05) for female students. Coaches reported that they had put a lot of effort into helping students to resolve personal problems, such as financial problems and problems with housing (or in guiding them to the appropriate organizations that could help them). This may have had a positive impact on the decision to stay in school, particularly among the group of boys not living with both parents, of which a relatively large proportion reported these problems.

5.7 Effects after two years on school dropout and start qualifications

The first cohort received two years of coaching. This allows us to compare the effect after two years of coaching to the effect after one year of coaching for the same sample of students. In addition, we can investigate the effect of coaching on obtaining a start qualification. For obtaining this qualification a student has to complete all elements of the study with a nominal duration of successfully. The table in Appendix B shows that treatment and control groups of the first cohort are quite similar. Regarding the outcome variables we observe that two years after the start the treatment group is less likely to dropout and more likely to obtain a start qualification.

Table 5.5 shows the estimated effects after one year and after two years of coaching for the first cohort that participated in the experiment. We find that after one year of coaching the effect on school dropout was statistically significant and points at a reduction of school dropout by 6.5 percentage points. The effect after two years of coaching is 7.3 percentage points, which is statistically significant as well. This estimate suggests a reduction in dropout from 22% to 14%. Hence, the estimated effect increases by approximately one percentage point in the second year.

Unfortunately, the second year of coaching was not randomly assigned in our experimental design; we cannot compare the effects after two years of treatment group that was assigned to one year of coaching with a treatment group that was assigned to two years of coaching. This implies that the evidence on the relative effectiveness of the first and second year of coaching should be considered as indicative and not as conclusive. If we assume that the gains from the first year of coaching are not lost in the second year then our findings suggest that of the gain in reducing school dropout comes from the first year of coaching. Two further observations seem to support this indication. First, national dropout figures show that most school dropout takes place in the first study year.[77] Second, many coaches have reported that the coaching capacity for the second year was (too) high, particularly since the original treatment groups had been reduced by on average a quarter after one year due to either switching to other courses or school dropout.

Table 5.5 OLS estimates of the effect of two years of coaching on school dropout and having obtained a start qualification, first cohort (starting in study year 2009-10)

	School dropout	Start qualification
Outcome variable	(1)	(2)
1a) One year of coaching	-0.065**	
	(0.027)	
1b) Two years of coaching	-0.073**	0.063
	(0.022)	(0.040)
Observations	216	216

Notes: All models include the full set of covariates as in column (3) of table 5.4. Standard errors are corrected for clustering at the class level. ** Significant at a 5 %-significance level.

The estimated effect of coaching on having obtained a start qualification two years after the start is quite similar to the estimated effects on school dropout but statistically insignificant (p-value 0.14). The point estimate would imply an increase in start qualification attainment share from 49% to 56% due to two years of being offered intensive coaching. It is likely that the insignificance of the effect on obtaining a start qualification can be explained by the timing of the second data collection. These data were collected after two years, which is exactly the nominal duration of the program and only the best students graduate within these two years. Hence, t it is somewhat early to evaluate the effects of coaching on educational

[77] See the national website on school dropout prevention of the Dutch Ministry of education: http://www.aanvalopschooluitval.nl/beleid/beleidsthemas/van-vmbo-naar-mbo.

attainment since only half of the students in our population manages to obtain a start qualification within two years.

Heterogeneous treatment effects

We also investigated heterogeneous treatment effects for the models that regress obtaining a start qualification (i.e. graduating from the two-year vocational program) on coaching. The estimates show that coaching has a strong and significant effect on the probability of having obtained a start qualification for students no longer obliged to be in formal education of 0.163 (0.063). This finding suggests that coaching has increased the probability of obtaining a start qualification among this group from 42% to 58%. The estimated effect for the subgroup of students obliged to stay in formal education is zero. A regression with an interaction term for treatment group and not obliged to be in formal education produces an effect estimate of 0.21 (0.05). This larger effect for the group not obliged to stay in education is in line with the larger effect on dropout reduction we found in Table 5.4. Descriptive statistics show that the share of students obtaining a start qualification within two years is around 60% among both coached and non-coached students that are obliged to be in formal education.

5.8. Cost-benefit analysis

To assess the possible impact of the coaching program on societal welfare we have performed a tentative cost-benefit analysis. The details of this analysis are shown in Appendix C. The cost-benefit calculations suggest that the internal rate of return (IRR) of one year of coaching is 6.9% and the IRR of two years of coaching is 3.7%. To put these estimates into perspective, Angrist & Lavy (2009) estimated an internal rate of return of 8.6% for a program offering financial incentives for high school students in Israel upon passing high school matriculation exams. They find a higher rate of return mainly because of the much lower average cost of their program per student. Cost-benefit calculations of the Education Maintenance Allowance program in the UK also point at a net social gain (see Dearden et al., 2009).[78] In sum, our cost-benefit analyses for the Dutch coaching program suggest that one year of coaching is likely to generate a net social gain. However, two years of coaching may not generate a net social gain. It should be noted that the program may well have other gains

[78] The required return to break even is estimated by Dearden et al. (2009) at 7.7%, whereas research for the UK shows that the returns from staying on in post-compulsory education are 11% for males and 18% for females.

to society which have been linked to dropout reduction but are not taken into account in our cost-benefit analyses, such as reducing crime. On the other hand, it is not yet clear whether the assumed wage increases will occur for students in our study who are affected by the coaching program in terms of not dropping out and instead obtained a start qualification.

5.9 Conclusions and discussion

In this chapter we investigated the effect of an intensive coaching program in secondary education using data from a randomized experiment. The coaching program can be characterized as a high quality/ high intensity program because of the educational experience and training of the coaches, the student/coach ratio and the broad range of interventions that were applied for supporting the students. One year of intensive coaching reduces the probability of dropping out of school from 17% to 10%, that is, a reduction of more than 40%.[79] One additional year of coaching reduces the dropout rate by a further 1 percentage point. These findings suggest that the second year of coaching adds little to the total reduction in school dropout; most of the gain of the coaching program seems to be generated in the first year of coaching. A cost-benefit analysis shows that the internal rate of return of one year of coaching is 6.9 percent and of two years of coaching is 3.7 percent. Hence, only the first year of coaching seems cost-effective. The coaching program did not have an effect on the decision of students to switch to other vocational courses. These findings suggest that the 'preventive elements' of the program worked well, but that the interventions aimed at keeping students in the education system once they had dropped out from a particular study have been no more successful than the interventions in the care-as-usual situation.

An investigation of heterogeneous treatment effects shows large differences between subgroups. The estimated effects are larger for male students, students no longer obliged to be in formal education, and students not living with both parents. Moreover, the effects are much larger for boys not living with their parents than for girls not living with their parents. This is consistent with recent research that shows that the development of boys is more sensitive to unstable home environments. The groups for which we find large effects of the program have a relatively large ex-ante probability of dropping out of school. The information to identify groups with a high ex-ante probability can be collected relatively easy

[79] This average effect size compares favorably to for instance the dropout reduction of 13% found by Dearden et al. for the Education Maintenance Program in the UK (see Dearden et al., 2009).

and can be used to target the program on students most likely to be helped by the coaching program. A more targeted approach of the coaching program on those students being most vulnerable to school dropout will likely improve the cost effectiveness of the coaching approach.

The coaching program that we evaluated consists of multiple interventions. In our experimental setup we cannot distinguish which intervention is most important. However, surveys among students, coaches and staff of the coaching program may provide some indications about this. Surveys among students of the first cohort (response of 133 students) show that personal guidance was valued highest (6.2 at a scale of 1-10), whereas home visits by the coach were valued lowest (5.2). In between are group activities with the coach (6.1) and visit(s) at the internship of the student (5.6). More than half of students states that the most important value added of the coach lies in involvement with study progress, about a quarter claims it lies in involvement with the students personal living situation, and about one out of five claims it lies in interference with the student's internship. Surveys among coaches point at 1-on-1 conversations with the students being valued highest in terms of the contribution to dropout prevention (8.5 at a scale of 1 to 10), whereas group activities were valued lowest (6.8). In between are monitoring absence and dropout (7.6), visit at internship (7.4), home visits (7.3), and after-care in case of dropout (7.2). The program management in a self-evaluation together with the coaches state that the three most valuable interventions have been (in descending order): working on study skills (planning and organizing; focus on self-reliance); counselling on personal problems of students (social-emotional; referring to internal and external assistance); contact with parents.

For assessing the potential impact of the program it seems important to note that the coaching program was implemented in a period with a strong policy focus on reducing school dropout rates. In this period two nationwide policies aimed at reducing school dropout rates were also introduced in the Netherlands. First, students below the age of 18 and without a start qualification (i.e. ISCED level 3) were obliged to stay in formal education.[80] Second, financial incentives for schools to reduce school dropout rates were introduced. Schools could receive additional funds if they succeeded in reducing school dropout rates compared to a base level of school dropout. Hence, the new policies stimulated both students and schools to reduce dropout. If these policies have been successful they might have reduced the

[80] The school leaving age used to be 16 and it did not matter whether or not a start qualification had been obtained.

effect of the coaching program. A final factor that may have affected the effectiveness of the coaching program is the relative high unemployment at the time of the experiment. Between 2008 and 2010 youth unemployment increased from 8.4 to 11.7%. This may have created additional incentives for all students in the treatment as well as in the control groups to stay in school, in line with empirical findings about the positive relationship between youth unemployment and school enrolment (Rivkin, 1995; Card & Lemieux, 2000; Clark, 2011, Rice, 1999, Messchi et al. 2011). These contextual factors might have limited the effectiveness of the program.

Not only the particular policy and economic context may have affected the effectiveness of the program, but also the particular target group of students aged 16-25 within intermediate post-secondary vocational education. Another particular feature is that the program was targeted at a group that had just made the transition from secondary to post-secondary education. It has been well documented that this is a phase in which the risk of dropping out is eminent, since it may be hard for students to make the right choice for a new study program and a new school out of many options, students often have to change study and travel routines, and students have to get used to new classmates and teachers at their new schools, etc.[81] The coaching program attempted to address several of difficulties that students encountered after this transition phase.

A further issue is to which extent the results of this experiment will also hold for other schools and other geographical areas. We note that the experiment took place in a school with a student population that is rather representative for the Dutch context (see also Appendix Table A1). The school had a rather average school dropout rate ranking 17th highest out of 42 institutions offering intermediate vocational education. Moreover, the vocational programs that we study are offered all over the country. Finally, early school leaving in the Netherlands is concentrated in intermediate vocational education (75 % of all early school leaving), which is the level of education in which the experiment took place. This probably suggests that our findings are generalizable to other schools and other geographical areas as well. It is difficult to assess whether our findings would also hold for younger students or for students in academic tracks. School dropout is much less of a problem among students younger than 16 years and in academic tracks. It seems also likely that students in academic tracks will be quite different from students in vocational tracks.

[81] Behavioural barriers within and after the transition phase from secondary to post-secondary education have been well documented in Lavecchia et al. (2014) and Ross et al. (2013).

The main finding of this study is that the coaching program has substantially reduced school dropout rates. Therefore, we conclude that intensive coaching can be a successful instrument for reducing school dropout, especially among students with a high ex-ante probability of dropping out.

Appendix A. School dropout in the Netherlands and policy context of the coaching experiment

The Netherlands had almost 40 thousand new school dropouts (or early school leavers) in the school year 2010-11 (source: www.aanvalopschooluitval.nl).) The official definition of an early school leaver is a student aged 12-22 that is (i) in education on the first of October (start of the school year), (ii) not in education one year later, and (iii) has not obtained a so-called 'start qualification' in the meanwhile. A start qualification is equal to a degree of upper secondary education or of intermediate post-secondary vocational education of at least level two (i.e. ISCED level 3 or higher). School dropout is largely concentrated in intermediate post-secondary vocational education (MBO), which has 75 percent of all school dropouts and which has a little less than 500 thousand students. Within MBO, 40 percent of all school dropouts are enrolled at level two. This is the level at which the coaching program took place. Thirty percent of all new school dropouts in the Netherlands drop out from level two. Official study duration at level two is two years and completing this level yields a start qualification. The national average school dropout rate at level two was 13 percent in school year 2010-11, that is, one out of every seven students at this level leaves education without a start qualification every year. See https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Netherlands:Overview for a schematic overview of the Dutch education system. The coaching experiment has been carried out in "MBO Basisberoepsopleiding", which is the same as level two.

The target of the current national action program against dropout " Aanval op de Uitval" is to reduce the yearly number of school dropouts to 25,000 by 2014-2015.[82] Total (yearly) public expenditures on Dutch dropout policy have been estimated to be around 400 million Euro in 2011 (Ecorys, 2009). An important part of this budget has been invested through regional covenants with a contact municipality and schools for secondary general education as well as vocational education within each region. The covenants describe targets for the subsequent years for each of the 39 regions which add up to the national dropout reduction target. Part of the provided funds are provided unconditionally, another part of the budget is paid to the

[82] The target in European perspective is based on another measure of school dropout. This is the share of students aged 18-24 with only lower secondary education at best and not in education or training. The EU 2020 target for the Netherlands to which the Dutch government has committed itself is 8%. The 2010 rate was 9.1% down from 15.1% in 2000. The EU-27 average rate was 13.5% in 2010, down from 17.6% in 2000. (source: http://europa.eu/rapid/pressReleasesAction.do?reference=IP/12/577&format=HTML&aged=0&language=EN&guiLanguage=en)

schools conditional on reaching preset targets for dropout reduction. This implies that there is in part a financial incentive to the schools to reduce school dropout. Schools and regions have full autonomy over their choice of anti-dropout measures.

The coaching experiment took place at ROC Rijnijssel, a large school for intermediate post-secondary vocational education. The school is located in Arnhem, a medium-sized city belonging to the 30 largest cities in the Netherlands. The school had a little less than ten thousand participants in school year 2010-11 of which 9.4 percent dropped out of education without a start qualification. ROC Rijnijssel had the 17th highest dropout rate out of 42 institutions offering intermediate post-secondary vocational education in the Netherlands. Appendix table A1 shows that the school at which the experiment took place is rather average as well in terms of student characteristics, domains of vocational courses offered and size of institution. The share of students living in a poverty accumulation area is somewhat higher than average though at ROC Rijnijssel (19 versus 12 percent).

Appendix Table A1 ROC Rijnijssel (i.e. institution where the experiment took place) versus all institutions offering intermediate vocational education in the Netherlands (source: www.aanvalopschooluitval.nl, figures for school year 2010-2011)

	ROC Rijnijssel	All institutions offering intermediate post-secondary vocational education
Dropout		
% of dropouts	9.4	8.4
% of dropouts at level 2	15.9	14.7
Student characteristics		
Male (%)	53	53
Dutch (%	77	76
Living in poverty accumulation area (%)	19	12
Domains of vocational courses offered		
Economic	37	35
Technical	22	27
Care and health	38	31
Agriculture	0	6
Combination	3	2
Economic	37	35
Institution size		
Average number of students	9512	10039

In 2009 the Dutch Ministry of Education was actively looking for opportunities to gain more (convincing) evidence on promising dropout interventions and invited institutions that offered intermediate post-secondary vocational education to see whether they would be willing to participate in a randomized dropout prevention experiment. ROC Rijnijssel in Arnhem was interested in an experiment. They were thinking about expanding an intensive coaching setting from MBO level 1 to level 2. Experiences with this intensive coaching setting at level 1 had been satisfying and it was felt that this approach contributed to dropout prevention, though convincing evidence was lacking. The school agreed to participate in a randomized experiment at level 2.

Appendix B. Descriptive statistics of first cohort, starting in school year 2009-10, having received two years of coaching.

Characteristic	All courses pooled		
	Control	Treated	p-value[a]
Student characteristics			
1. Male	0.23	0.38	*0.51*
2. Age (in years)	18.2	17.9	*0.76*
3. Obliged to stay in formal education[b]	0.45	0.61	*0.29*
4. Born in the Netherlands	0.88	0.93	*0.19*
5. Living with both parents	0.57	0.62	*0.76*
6. Having problems in at least one of the following areas: finance, police and justice, family and friends, or living/housing situation	0.48	0.35	*0.11*
7. Score on verbal skills at intake test (1-5)	3.2	3.3	*0.37*
8. Score on numeric skills at intake test (1-5)	3.1	3.2	*0.19*
9. Highest previous attained degree (1-6)	2.3	2.2	*0.96*
10. Already obtained a start qualification at the start of the experiment	0.05	0.13	*0.14*
11. Late choice (July or later)	0.27	0.28	*0.82*
12. Average class size (of started students)	20.9	20.7	*0.30*
Outcome variables (after one year)			
13.a School dropout[c]	0.12	0.04	*0.06*
13.b Switch to another study	0.26	0.23	*0.43*
13.c Still in same study	0.63	0.73	*0.19*
Outcome variables (after two years)			
14.a. School dropout[c]	0.22	0.12	*0.06*
14.b. Obtained 'start qualification'[d]	0.49	0.60	*0.11*
Number of classes	8	4	12
Observations	142	74	216

Notes:

A missing value on the background characteristics is limited to maximum six percent of the pooled sample.

(a) Controlling for cohort and vocational course.

(b) All students under 16 are obliged to go to school. Students of 16 and 17 are obliged to stay in education if they have not completed a degree that counts as a start qualification (i.e. ISCED level 3 or higher).

(c) School dropout is defined as having left education without having obtained a start qualification.

(d) A start qualification is equal to ISCED level 3 (or higher).

Appendix C. Cost-benefit analysis

We started with the calculation of the rate of return of a program of one year of coaching. The costs of one year of coaching amount to 3,000 euro per treated student (i.e. 60 k euro for a FTE equivalent of coaching per group divided by 20 students per group). The returns per year are calculated making the following assumptions: we use average annual earnings of workers without a start qualification as a base (25,265 euro)[83]; we use the effect estimate of (minus) 7.1 percentage points of the effect on school dropout after one year as a proxy for the definitive effect on school dropout; we assume this seven percent of the treated population not becoming an school dropout due to coaching receives two extra years of schooling[84], each year yielding a rate of return of 10%[85]; public as well as private costs of these two extra years of schooling for seven percent of the treated population are taken into consideration. These costs consist of around 5k euro public contribution and 1k euro private contribution per study year per student in intermediate post-secondary vocational education.

The yearly return can then be calculated as follows: 25265*0.10*2*0.071 = 353 euro per year. These returns are assumed to start occurring in the fifth year after the coaching started (to take into account extra study duration and the time to labor market entry) and are

[83] This is a weighted average of wage income of three different subgroups varying by their distance to a start qualification (and thus by their completed years of education), the weights corresponding to relative occurrence of these subgroups in our sample. Wage figures are taken from Arbeidsmarktpanel 2009 (Statistics Netherlands). We used average yearly wage income of workers aged 20-64.

[84] The reasoning for using two years is as follows. The distance to a start qualification of the group without a start qualification in terms of years of completed education is one year for the group with MBO level 1, two years for the subgroups with completed secondary vocational education, and six years for the subgroup with just primary education. The shares of these subgroups in the sample without a start qualification at the start are 10, 77 and 13% respectively. This would imply the average distance to a start qualification in terms of completed years of education is 2.4 years in our sample. Furthermore, a start qualification gives access to higher levels of intermediate vocational education (whereas this access is not granted without a start qualification), such that the definitive difference in years of completed education among students managing to obtain a start qualification and those that do not is probably even larger than two years. Nevertheless we use a conservative assumption of two years of additional education linked to those students not becoming a school dropout due to coaching.

[85] OECD (2012) shows that people (aged 25-64) having attained less than upper secondary education earn 19 percent less than people having attained upper secondary education in the Netherlands. The average earnings difference in OECD is 24% between these two groups. This earnings difference increased somewhat in the last decade (up from 20 percent in 2000), despite a rather strong decline in the share of people having attained less than upper secondary education (from 36% to 26%). This suggests that relative demand for people with below upper secondary education (relative to those with upper secondary education) has fallen. These earnings differences need not represent the causal effects of obtaining an upper secondary level, but come close to other courses which have used credible designs to detect the returns to education (see e.g. Card, 1999 and Heckman et al., 2006 for reviews of this literature).

assumed to be maintained for 42 years.[86] Bringing these costs and benefits all together yields an internal rate of return of one year of coaching of 6.9%. At the advised discount rate of 5.5%[15] one year of coaching would then yield a positive net present value of 18 k euro per coached group (at an initial investment of 60 k euro per group). To put it differently, we would need a sustained effect of at least 5.5% point less school dropout in order for one year of intensive coaching to break even at a discount rate of 5.5%.

The estimated internal rate of return of two subsequent years of coaching is 3.7%, which implies a net social loss at the advised discount rate of 5.5%. This estimate is based on (i) the effect estimate of two years of coaching of -7.3% point on school dropout (based on the first cohort sample), (ii) € 6,000 of initial investments in coaching per treated student (i.e. two years of €3,000 euro), and (iii) for the rest on the same assumptions as above. To put it differently, we would need an effect of 10% points less school dropout in order to break even with the two year coaching program at its current costs (and at the advised discount rate of 5.5%).

[86] Average age at start of the experiment is 18. Official pension age was 65 at the time of the experiment but is agreed to go up to 67 by 2025. We assume benefits of higher educational attainment will continue up to the age of 65.

Summary

This thesis aims to add to the literature and policy debate by providing additional insights to measuring teacher quality and to the effectiveness of education policies to raise the quality and quantity of teachers. These are policy relevant questions since it has been found that teachers matter a great deal for pupil performance as well as for later success in life. At the same time relatively little is known about the determinants of teacher quality and about effective policies to raise the quality and quantity of teachers.

The first policy consists of the introduction of schooling vouchers for teachers that can be used to finance a bachelor or master degree program to gain a higher or additional certification or additional skills. The second policy is a regional teacher remuneration improvement policy that triggered a higher teacher pay in an urbanized region in the Netherlands with less qualified teachers and more socio-economically disadvantaged pupil populations. This policy had the goal to attract and maintain more teachers in this hard-to-staff region.

The final chapter is not related to teachers but investigates the effectiveness of an intensive coaching program for students in post-secondary vocational education on student dropout. This chapter adds to the small, but growing literature on the effectiveness of coaching or mentoring interventions on school success (see e.g. a section in Lavecchia et al, 2014 for an overview of these studies).

In this thesis I use well-known empirical approaches for identification of the causal relationships I am interested in. These include instrumental variables, differences-in-differences, fuzzy regression discontinuity and a setup and evaluation of a randomized control trial. I refer to Angrist & Pischke (2009) for more detailed descriptions of these approaches.

The chapters of this thesis are summarized below.

Chapter 2 examines the relationship between teacher evaluations and pupil performance gains in primary education. Teacher evaluations have been conducted by trained external evaluators who scored teachers on a detailed rubric containing 75 classroom practices. These practices reflect pedagogical, didactical and classroom organization competences considered crucial for effective teaching. Conditional on previous year test scores and several pupil and classroom characteristics the score on this rubric significantly predicts pupil performance

gains on standardized tests in math, reading and spelling. Estimated test score gains are in the order of 0.4 standard deviations in math and spelling and 0.25 standard deviations in reading if a pupil is assigned a teacher from the top quartile instead of the bottom quartile of the distribution of the evaluation rubric. The observation rubric particularly seems to have potential to identify the weaker teachers.

Chapter 3 investigates the effects of schooling vouchers for teachers on enrollment and completion of higher education programs, as well as on retention of teachers. This is done by employing a so-called fuzzy regression discontinuity design. The discontinuity in the probability of being assigned a voucher arises due to budget constraints in the first application period. The estimates suggest that effects of voucher assignment on both higher education enrollment and completion rates are in the order of 10 to 20 percentage points as measured five and a half years after application. Relative to a baseline enrollment rate of 77 percent and a baseline completion rate of 54 percent (i.e. of applicants that were not assigned a voucher), these effect estimates correspond to a 12 to 29 percent higher enrollment and to a 17 to 42 percent higher completion. Effects on enrollment and completion are relatively small for shorter studies (up to one year) and for teachers that had already started at the time of application. The teacher voucher crowds out funding by schools out of their regular professional development budgets as well as own financial contributions by teachers. Our results suggest small positive effects of voucher assignment on retention in education as measured four years after application.

Chapter 4 investigates the effects of higher teacher pay for secondary school teachers on their teacher retention decision and enrollment in additional schooling. I exploit regional variation in teacher pay that is induced by the introduction of a new teacher remuneration policy. This policy provided schools in an urbanized region with extra funds to place a larger share of teachers in a higher salary scale. We exploit this policy in an instrumental variable setup to estimate the effects of higher teacher pay on our outcomes. The main finding is that we find no effects of higher teacher pay on the probability to stay in the teaching profession. The policy however succeeded in keeping a slightly larger share of teachers in the targeted region. In addition, our findings suggest that the policy increased teachers' enrollment in bachelor or master degree programs from 2.3% to 3.2%. This finding is consistent with the setup of the policy in which one of the criteria for placement in a higher salary scale is that teachers would obtain extra qualifications or gain extra expertise.

Chapter 5 investigates the effect of an intensive coaching program aimed at reducing school dropout rates among students aged 16 to 20 in post-secondary vocational education. Within the coaching program students were offered fulltime support and guidance with their study activities, personal problems and internships in firms. The coaching program lasted one or two years. Students were randomly assigned to classes and the coaching program was randomly assigned to classes as well. We find that one year of coaching reduced school dropout rates by more than 40 percent from 17 to 10 percentage points. The second year of coaching further reduced school dropout by one percentage point. The program is most effective for students with a high ex-ante probability of dropping out, such as students no longer obliged to be in formal education, male students, and students not living with both parents. Cost-benefit analysis suggests that one year of coaching is likely to yield a net social gain.

Nederlandse samenvatting

Motivatie en onderzoeksvragen

Een groeiende literatuur laat consistent zien dat de kwaliteit van leraren een bepalende factor is voor de prestaties van leerlingen. Leerlingen die worden toegewezen aan een leraar met een standaarddeviatie hogere kwaliteit winnen 0.1 tot 0.3 standaarddeviatie meer in termen van prestaties op cognitieve toetsen (zie bv. Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Kane & Staiger, 2008; Hanushek & Rivkin, 2010). In een recent paper van Chetty et al. (2014) wordt bovendien gevonden dat de invloed van leraren zich ook uitstrekt tot uitkomsten later in het leven. Een jaar in een klas met een leraar met een 1 standaarddeviatie hoger dan gemiddelde kwaliteit houdt verband met een toename in de kans op deelname aan college van 0.8 procent, met 0.4 procent in de kans om werk te hebben en met 350 dollar per jaar hoger inkomen, beiden op 28-jarige leeftijd.

Het is alleen minder duidelijk wat nu precies de kwaliteit van leraren bepaalt en hoe de kwaliteit kan worden bevorderd. Observeerbare kenmerken als opleidingsniveau en ervaring – opvallend genoeg vaak de belangrijkste determinanten van de beloning van leraren - blijken slechts een beperkt gedeelte van de variatie in lerarenkwaliteit te verklaren, zie overzichtsstudies van Hanushek en Rivkin (2006) en Harris en Sass (2011). Anders gezegd, het grootste gedeelte van de verschillen in de kwaliteit van leraren bevindt zich binnen leraren met hetzelfde opleidings- en ervaringsniveau. Deze lage verklarende kracht van observeerbare factoren heeft er in de afgelopen jaren toe geleid dat onderwijsonderzoekers hun focus zijn begonnen te verleggen naar de vraag welke docentpraktijken en –vaardigheden ertoe doen en in hoeverre verschillende soorten meetinstrumenten van docentkwaliteit voorspellend zijn voor leerwinst van leerlingen. Een recent groot onderzoeksproject dat zich op deze onderzoeksvragen werpt is het Measures of Effective Teaching project in de VS. Hoofdstuk 2 van dit proefschrift sluit aan op deze smalle maar groeiende literatuur en onderzoekt in hoeverre een gedetailleerd observatie-instrument gericht op het meten van pedagogische, didactische en organisatorische vaardigheden van leraren voorspellend is voor leerwinst van leerlingen.

Gegeven het belang van leraren voor onderwijskwaliteit is het niet vreemd dat beleidsmakers in de hele wereld allerhande beleidsmaatregelen inzetten om de kwaliteit en kwantiteit van leraren te verhogen. Investeringen in in scholing van leraren en in een hogere beloning van leraren behoren tot de beleidsmaatregelen die frequent worden ingezet. De evidentie onder deze beleidsmaatregelen levert een gemengd beeld op en overtuigende effectstudies zijn over het algemeen nog vrij schaars. Dit komt vooral door een gebrek aan experimentele of quasi-experimentele variatie in deze beleidsinterventies die benut kan worden om effecten betrouwbaar te schatten. Hoofdstuk 3 en 4 van dit proefschrift dragen bij aan het vergaren van meer kennis over de effectiviteit van beleidsmatige investeringen in het stimuleren van scholingsdeelname onder leraren respectievelijk een hogere beloning van leraren.

Naast de stevige evidentie dat leraren ertoe doen voor het succes van leerlingen is er een beperktere maar groeiende literatuur die erop wijst dat coaching of mentoring van jongeren gedurende het onderwijs positief kan uitpakken voor de schoolloopbanen van jongeren. Zie bv. Lavecchia (2014) voor een overzicht van enkele van deze studies. Hoofdstuk 5 van dit proefschrift voegt toe aan deze literatuur door een experimentele evaluatie van een intensieve coaching aanpak voor studenten aan de onderkant van het middelbaar beroepsonderwijs.

De vier hoofdonderzoeksvragen die in vier achtereenvolgende hoofdstukken van dit proefschrift aan de orde komen in dit proefschrift zijn de volgende:

1. In welke mate is de score op een gedetailleerd observatie-instrument van leraren voorspellend voor de leerwinst die hun leerlingen boeken? (hoofdstuk 2)
2. Wat is het effect van scholingsvouchers voor leraren op hun deelname aan en afronding van hoger onderwijsopleidingen? (hoofdstuk 3)
3. Wat is het effect van een hogere beloning van leraren op hun kans op behoud voor het lerarenberoep? (hoofdstuk 4)
4. Wat is het effect van een intensieve coaching aanpak voor studenten aan de onderkant van het middelbaar beroepsonderwijs op de kans op voortijdig schoolverlaten? (hoofdstuk 5)

Methoden

Een belangrijke uitdaging in dit proefschrift is het identificeren van de oorzakelijke effecten van de (beleids-)interventies op de relevante uitkomstvariabelen. Dat is vaak lastig omdat er geregeld sprake is van selectie-effecten. Dat wil zeggen dat er vaak sprake is van niet-geobserveerde verschillen tussen degenen die wel en degenen die niet aan een bepaalde interventie deelnemen. Dat kunnen factoren zijn als motivatie en aanleg die tegelijkertijd van invloed zijn op de uitkomstvariabelen. Als er sprake is van dergelijke niet-geobserveerde verschillen zullen simpele vergelijkingen in uitkomsten tussen degenen die wel en niet de interventie hebben ondergaan een vertekend beeld geven van de daadwerkelijke effecten. De afgelopen decennia zijn in de empirische economische literatuur steeds vaker nieuwe empirische methoden toegepast om het endogeniteitsprobleem bij effectstudies te adresseren Angrist en Pischke (2010) spreken in dit verband van de 'credibility revolution in empirical economics'. Het gaat hier om methoden die gebruik maken van experimentele of quasi-experimentele variatie in de toekenning van de interventies aan individuen. Door deze variatie ontstaan geloofwaardige controlegroepen van individuen die de interventie niet ondergaan hebben voor de individuen die de interventie wel ondergaan hebben en kunnen effecten betrouwbaarder worden geschat.

In dit proefschrift wordt een variëteit van deze empirische onderzoeksmethoden toegepast om de relevante oorzakelijke verbanden vast te stellen. Het betreft onder meer het benutten van een zogenoemde 'fuzzy regressie-discontinuïteit' (hoofdstuk 3), instrumentele variabele en verschil-in-verschillen schattingen (hoofdstuk 4), en benutting van experimentele variatie in een gerandomiseerd experiment (hoofdstuk 5). Ik verwijs graag naar het boek 'Mostly harmless econometrics' van Angrist en Pischke (2009) voor een uitgebreide beschrijving van deze methoden en de complicaties die zich hierbij kunnen voor doen en hoe daarmee om te gaan. In hoofdstuk 2 gaat het niet om een effectstudie maar om het bepalen van de voorspellende kracht van een meetinstrument voor prestaties van leerlingen op cognitieve toetsen. Om deze voorspellende kracht zo goed mogelijk vast te stellen wordt gebruik gemaakt van een regressie waarin het verband wordt geschat tussen de score op het meetinstrument en prestaties van leerlingen op taal- en rekentoetsen. Hierbij wordt gecontroleerd voor een uitgebreide set van relevante achtergrondkenmerken van leerlingen (waaronder de voorgaande scores van leerlingen en de sociaal-economische achtergrond) en klassen.

Resultaten

Hoofdstuk 2 onderzoekt de relatie tussen evaluaties van leraren, uitgevoerd door getrainde beoordelaars tijdens lesbezoeken, en leerlingprestaties in het basisonderwijs in een grote stad in Nederland. De evaluaties zijn gebaseerd op een gedetailleerd scoresysteem waarin leraren door getrainde observanten tijdens lessen gescoord worden op het wel of niet laten zien van 75 gedragsaspecten. Deze gedragsaspecten geven uiting aan organisatorische, didactische en pedagogische competenties die gerelateerd zijn aan opbrengstgericht lesgeven. De analyses laten zien dat de evaluatiescores van leraren significant voorspellend zijn voor de vooruitgang in leerlingprestaties van de leerlingen in hun klas. De gemiddelde voorspelde vooruitgang in toets-scores bij rekenen en spelling is 0.4 standaarddeviatie als een leerling is toegewezen aan een leraar uit het bovenste kwartiel van de gemeten vaardigheidsverdeling in plaats van een leraar uit het onderste kwartiel. Dit betekent dat een basisschoolleerling die twee jaar op rij een zwakke leraar krijgt toegewezen in plaats van een goede leraar, hierdoor een heel niveau lager terecht kan komen in het vervolgonderwijs, dus van bijvoorbeeld in potentie vwo-niveau naar havo-niveau. De resultaten suggereren dat met het evaluatie-instrument gedragsaspecten van leraren gemeten worden die ertoe doen voor leerlingprestaties. Vooral zwakkere leerkrachten kunnen met het observatie-instrument worden geïdentificeerd. De resultaten suggereren dat het evaluatie-instrument potentie heeft om gebruikt te worden voor ontwikkelings- en personeelsbeleid.

Hoofdstuk 3 onderzoekt het effect van toewijzing van een lerarenbeurs op deelname en afronding van hoger-onderwijsopleidingen (bachelors of masters) door leraren. De effecten worden geschat door het benutten van een zogenoemde regressiediscontinuïteit in de kans om een beurs toegewezen te krijgen. De discontinuïteit is ontstaan doordat het aantal aanvragen in de eerste ronde veel hoger was dan het budget toeliet en de lerarenbeurzen vervolgens zijn toegewezen op volgorde van binnenkomst van de aanvragen. Het geschatte positieve effect van toewijzing van een lerarenbeurs op zowel de kans op deelname als op afronding ligt gemiddeld in de orde van 10 tot 20 procentpunt. Dit komt overeen met een relatief effect van 12-29 procent op deelname en 17-42 procent op afronding. De effecten op deelname en afronding zijn hoger dan gemiddeld voor aanvragers die nog niet gestart waren met de opleiding op het moment van aanvraag, maar lager dan gemiddeld voor aanvragers voor kortere opleidingen tot en met een jaar. Deze effecten zijn gemeten vijfeneenhalf jaar na aanvraag van de lerarenbeurs. De lerarenbeurs vormt grotendeels een substituut voor zowel financiering uit reguliere scholingsbudgetten van scholen als voor financiering uit eigen bijdragen van leraren. Er zijn aanwijzingen voor beperkte positieve effecten van de lerarenbeurs op de kans om in het onderwijs te blijven. Deze effecten lijken geconcentreerd bij leraren in het voortgezet onderwijs en bij leraren die nog niet gestart waren met de opleiding op het moment van de aanvraag.

Hoofdstuk 4 onderzoekt de vraag in hoeverre een hogere beloning voor leraren in het voortgezet onderwijs hun beslissing om leraar te blijven beïnvloedt. Om deze vraag te kunnen beantwoorden, wordt gebruik gemaakt van een beleidsmaatregel die heeft geleid tot regionale verschillen in de beloning van leraren in het voortgezet onderwijs in Nederland. Vanaf 2009 kregen scholen in de Randstad, in vergelijking met scholen buiten de Randstad, extra geld om meer leraren in een hogere salarisschaal te plaatsen. Het betrof een totale additionele bekostiging van 290 miljoen euro over de periode 2009-2014. Deze zogenoemde versterking van de Functiemix in de Randstad heeft ertoe geleid dat in 2014 bijna 20 procentpunt meer leraren in een hogere salarisschaal zijn geplaatst dan op scholen buiten de Randstad. Eenmaal in de hogere salarisschaal, kregen de leraren uitzicht op een 17 procent hogere beloning, ofwel 7200 euro bruto op jaarbasis. Doel van deze beleidsmaatregel was om de beloningsachterstand ten opzichte van banen buiten het onderwijs te verkleinen en (toekomstige) lerarentekorten in de Randstad te bestrijden. De belangrijkste bevinding van dit onderzoek is dat we geen effecten vinden van deze hogere beloning op de kans om leraar te blijven. We hebben de uittreedkans van leraren in de Randstad vergeleken met die van leraren buiten de Randstad. Elk jaar treedt ongeveer 7% van de leraren uit het lerarenberoep, zowel binnen als buiten de Randstad. Dit percentage is na de invoering van de hogere beloning voor de leraren in de Randstad niet veranderd ten opzichte van leraren buiten de Randstad. De hogere beloning heeft er wel voor gezorgd dat een iets groter deel van de leraren in de Randstad blijft werken en niet kiest om elders een baan als leraar te aanvaarden. De jaarlijkse kans voor een leraar om te switchen van een school in de Randstad naar een school buiten de Randstad is gedaald met 0,4 procentpunt. Dit komt overeen met ongeveer 125 leraren per jaar die niet van regio veranderd zijn op een totaal van circa 30 duizend werkzame leraren in de Randstad. Onze bevindingen suggereren daarnaast dat het beleid heeft geleid tot meer deelname aan formele scholing. Het jaarlijkse aandeel leraren dat een aanvraag doet voor een lerarenbeurs, om zo deel te nemen aan een bachelor of masteropleiding, is door de regionale versterking van de Functiemix gestegen van 2.3 naar 3.2 procent. Deze bevinding is consistent met de opzet van het beleid, waarbij een van de overeengekomen criteria voor plaatsing in een hogere salarisschaal is dat leraren meer kwalificaties of extra expertise verwerven.

Hoofdstuk 5 onderzoekt het effect van intensieve coaching van studenten aan de onderkant van het middelbaar beroepsonderwijs op voortijdig schoolverlaten op basis gegevens van een gerandomiseerd experiment. Het coaching-programma bood studenten intensieve ondersteuning en advies bij hun studieactiviteiten, persoonlijke problemen en bij hun stages. Het coaching programma duurde 1 of 2 jaar. Studenten werden aselect toegewezen aan klassen en klassen werden aselect toegewezen aan de coachingaanpak. De belangrijkste bevinding is dat één jaar coaching voortijdige schoolverlaten met meer dan 40 procent reduceert: van 17 naar 10 procentpunt. Het tweede jaar coaching zorgde voor een verdere daling van het voortijdige schoolverlaten met 1 procentpunt. Het programma is het meest effectief voor studenten met een hoge ex-ante kans op uitval. Het betreft studenten die niet meer onder de kwalificatieplicht vallen, mannelijke studenten en studenten die niet bij beide ouders wonen. Een kosten-batenanalyse suggereert dat bij één jaar coaching de maatschappelijke baten groter zijn dan de kosten.

Curriculum Vitae

Marc van der Steeg was born in 1979 in Leiderdorp. He studied Economics at Erasmus University Rotterdam from 1997 to 2003. He obtained his MA degree in 2003. His master thesis was on India’s vulnerability to macroeconomic shocks for which he went to India to the Institute of Economic Growth in Delhi for half a year. He joined CPB Netherlands Bureau for Economic Policy Analysis in 2004 and has been working there till the end of 2014. His work at CPB was mainly on education and education policy evaluation, but he has also worked on innovation and science policy. Since the end of 2014 he is working at the Dutch Ministry of Education, Culture and Science as a senior advisor at the Knowledge Directorate. In 2012 he started writing his PhD thesis.

Bibliography

Introduction and summary (as far as not listed under the chapters)

Angrist, J. and J. Pischke, 2009, Mostly harmless econometrics: An empiricist's companion, Princeton University Press, Princeton.

Angrist, J. and J. Pischke, 2010, The credibility revolution in empirical economics: how better research designs is taking the con out of econometrics, *Journal of Economic Perspectives*, vol. 24, no. 2, pp. 3-30.

Hahn, J., P. Todd, and W. van der Klaauw, 2001, Identification and estimation of treatment effects with a regression discontinuity design, Econometrica, vol. 69, no. 1, pp. 201-209.

Hanushek, E., 2011, The economic value of higher teacher quality, *Economics of Education Review*, vol. 30, no. 3, pp. 466-479.

Heckman, J., J. Stixrud, and S. Urzua, 2006, The effects of cognitive and non-cognitive abilities on labor market outcomes and social behavior, *Journal of Labor Economics*, vol. 24, no. 3, pp. 411-482.

Lavecchia, A., H. Liu, and P. Oreopoulos, 2014, Behavioural economics of education: Possibilities and progress, *NBER Working Paper*, no. 20609.

Lee, D., and T. Lemieux, 2010, Regression discontinuity designs in economics, *Journal of Economic Literature*, vol. 48, pp. 281–355.

Chapter 2

Aaronson, D., L. Barrow, and W. Sander, 2007, Teachers and Student Achievement in the Chicago Public High Schools, *Journal of Labor Economics*, vol. 25, no. 1, pp. 95–135.

Angrist, J., and A. Krueger, 1991, Does compulsory school attendance affect schooling and earnings?, *Quarterly Journal of Economics*, vol. 106, no. 4, pp. 979–1014.

Araujo, M., P. Carneiro, Y. Cruz-Aguayo, and N. Schady, 2016, Teacher quality and learning outcomes in kindergarten, *IZA Discussion Paper*, no. 9796.

Chetty, R., N. Friedman, and J. Rockoff, 2013a, The long-term impact of teachers: teacher value-added and student outcomes in adulthood, *NBER Working Paper*, no. 17699.

Chetty, R., N. Friedman, and J. Rockoff, 2013b, Measuring the impact of teachers I: Evaluating bias in teacher value-added estimates, *NBER Working Paper*, no. 19423.

Clotfelter, C., H. Ladd, and J. Vigdor, 2006, Teacher–Student Matching and the Assessment of Teacher Effectiveness, *NBER Working Paper*, no. 11936.

Hansen, K., J. Heckman, and K. Mullen, 2004, The effect of schooling and ability on achievement test scores, *Journal of Econometrics*, vol. 121, no. 1–2, pp. 39–98.

Hanushek, E., and S. Rivkin, 2010, Using Value-Added Measures of Teacher Quality, *American Economic Review*, vol. 100, no. 2, pp. 267–71.

Harris, D., and T. Sass, 2011, Teacher training, teacher quality and student achievement, *Journal of Public Economics*, vol. 95, pp. 798–812.

Harris, D, and T. Sass, 2014, Skills, Productivity and the evaluation of teacher performance, *Economic of Education Review (forthcoming)*, http://dx.doi.org/10.1016/j.econedurev.2014.03.002

Holtzapple, E., 2003, Criterion-related validity evidence for a standards-based teacher evaluation system, *Journal of Personnel Evaluation in Education*, vol. 17, no. 3, pp. 207-219.

Jacob, B., 2007, The Challenges of Staffing Urban Schools with Effective Teachers, *The future of Children*, vol. 17, no. 1, pp. 129–54.

Jacob, B., and L. Lefgren, 2008, Principals as agents: Subjective performance measurement in education, *Journal of Labor Economics*, vol. 26, no. 1, pp. 101-136.

Kane, T., and D. Staiger. 2008, Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, *National Bureau of Economic Research Working Paper*, no. 14601.

Kane, T., E. Taylor, J. Tyler, and A. Wooten, 2011, Identifying effective classroom practices using student achievement data, *Journal of Human Resources*, vol 46, no. 3, pp. 587-613.

Kane, T., and D. Staiger, 2012, Gathering feedback for teaching: combining high-quality observations with students surveys and achievement gains, *Measures of Effective Teaching Research Paper*.

Kane, T., D. McCaffrey, T. Miller, and D. Staiger, 2013, Have we identified effective teachers? Validating measures of effective teaching using random assignment, *Measures of Effective Teaching Research Paper*.

Ministry of Education, 2013, Kerncijfers 2008-2012: Onderwijs, cultuur en wetenschap, The Hague.

Nye, B, S. Konstantopoulos and L. Hedges, 2004, *Educational Evaluation and Policy Analysis*, vol. 26, no. 3, pp. 237-257

Rivkin, S., E. Hanushek, and J. Kain. 2005, Teachers, Schools and Academic Achievement, *Econometrica*, vol. 73, no. 2, pp. 417–58.

Rockoff, J., 2004, The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data, *American Economic Review*, vol. 94, no. 2, pp. 247-252.

Rockoff, J., and C. Speroni, 2011, Subjective and objective evaluations of teacher effectiveness: Evidence from New York City, *Labour Economics*, vol. 18, pp. 687-696.

Rockoff, J., D. Staiger, T. Kane, and E. Taylor, 2012, Information and employee evaluation: evidence from a randomized intervention in public schools, *American Economic Review*, vol. 94, no. 2, pp 3184-3213.

Rothstein, J., 2010, Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement, *Quarterly Journal of Economics*, vol. 25, no. 1, pp. 175-214.

Staiger, D., and J. Rockoff, 2010, Searching for effective teachers with imperfect information, *Journal of Economic Perspectives*, vol. 23, no. 3, pp 97-118.

Taylor, E., and J. Tyler, 2012, The effect of evaluation on teacher performance, *American Economic Review*, vol. 102, no. 7, pp. 3628-3651.

Tyler, J., E. Taylor, T. Kane and A. Wooten, 2009, Using student performance data to identify effective classroom practices, Working Paper.

Tyler, J., E. Taylor, T. Kane, and A. Wooten, 2010, Using student performance data to identify effective classroom practices, *American Economic Review Papers and Proceedings*, vol. 100, pp. 256-260.

Webbink, D., and S. Gerritsen, 2013, How much do children learn in school? Evidence from school entry rules, *CPB Netherlands Bureau for Economic Policy Analysis Discussion Paper*, no. 255.

Weisberg, D., S. Sexton, J. Mulhern, and D. Keeling, 2009, The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Effectiveness, New York.

Wiswall, M., 2013, The dynamics of teacher quality, *Journal of Public Economics*, vol. 100, issue C, pp. 61-78.

Chapter 3

Aaronson, D., L. Barrow, and W. Sander, 2007, Teachers and Student Achievement in the Chicago Public High Schools, *Journal of Labor Economics*, vol. 25, no. 1, pp. 95–135.

Abramovsky, L., E. Battistin, E. Fitzsimons, A. Goodman, H. Simpson, 2011, Providing employers with incentives to train low-skilled workers: evidence from the UK Employer Training Pilots, *Journal of Labor Economics*, vol. 29, no. 1, pp. 153-193.

Angrist, J., and V. Lavy, 2001, Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools, *Journal of Labor Economics*, vol. 19, no. 2, pp. 343-369.

Angrist, J., and J. Pischke, 2009, Mostly harmless econometrics: An empiricist's companion, Princeton University Press.

Berndsen, F., and H. van Leenen, 2013, IPTO Bevoegdheden 2011, Amsterdam: Regioplan.

Berndsen, F., J. Brekelmans, B. Dekker, and C. Bergen, 2014, Onderwijs werkt!; Rapportage van een enquête onder docenten en management uit het po, vo, mbo en hbo Meting 2013, Amsterdam: Regioplan.

Boom, E. van der, and M. Stuivenberg, 2014, Teaching and Learning International Survey (Talis) 2013 Nationaal rapport Nederland.

Campbell, D., 1969, Reforms as Experiments, *American Psychologist*, vol. 24, pp. 409–429.

Caridad Araujo, M., P. Carneiro, Y. Cruz-Aguayo, and N. Schady, 2016, Teacher Quality and Learning Outcomes in Kindergarten, *IZA Discussion Paper*, no. 9796.

Chetty, R., J. Friedman & and J. Rockoff, 2014, Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood, *American Economic Review*, vol. 104, no. 9, pp 2633-79.

Garet, M.S., S. Cronen, M, Eaton, A. Kurki, M. Ludwig, W.,Jones, K. Uekawa, A. Falk, H. Bloom, F. Doolittle, P.,Zhu, and L. Sztejnberg, 2008, The impact of two professional development interventions on early reading instruction and achievement. U.S. Department of Education, NCEE, Washington, DC.

Garet, M., A. Wayne, F. Stancavage, J. Taylor, K. Walters, M. Song, S. Brown, S. Hurlburt, P. Zhu, S. Sepanik en F. Doolittle, 2010, Middle school mathematics professional development impact study: findings after the first year of implementation, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, NCEE, Washington DC.

Gelman, A., and G. Imbens, 2014, Why high-order polynomials should not be used in regression discontinuity designs, *NBER Working Paper*, no. 20405.

Gerritsen, S., E. Plug, and D. Webbink, 2014, Teacher quality and student achievement: evidence from a Dutch sample of twins, chapter in PHD dissertation of S. Gerritsen at Erasmus School of Economics.

Hahn, J., P. Todd, and W. Van Der Klaauw, 2001, Identification and Estimation of Treatment Effects with a Regression Discontinuity Design, *Econometrica*, vol. 69, pp. 201-209.

Hanushek, E., and S. Rivkin, 2010, Using Value-Added Measures of Teacher Quality, *American Economic Review*, vol. 100, no. 2, pp. 267–71.

Harris, D. and T. Sass, 2011, Teacher training, teacher quality and student achievement, *Journal of Public Economics*, vol. 95, no. 7, pp. 798-812.

Hidalgo, D., H. Oosterbeek, and D. Webbink, 2014, The impact of training vouchers on low-skilled workers, *Labour Economics*, vol. 31, pp. 117-128

Inspectorate of Education, 2012, De Staat van het Onderwijs: Onderwijsverslag 2010/2011, Utrecht.

Inspectorate of Education, 2013, Professionalisering als gerichte opgave; verkennend Onderzoek naar het leren van leraren, Utrecht.

Inspectorate of Education, 2014, De Staat van het Onderwijs: Onderwijsverslag 2012/2013, Utrecht.

Jacob, B., and L. Lefgren, 2004, The Impact of Teacher Training on Student

Achievement Quasi-Experimental Evidence from School Reform Efforts in Chicago, *Journal of Human Resources*, vol. 39, no. 1, pp. 50-79.

Jacob, B., and L. Lefgren, 2011, The impact of NIH postdoctoral training grants on scientific productivity, *Research Policy*, vol. 40, pp. 864-874.

Kane, T., and D. Staiger. 2008, Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, *National Bureau of Economic Research Working Paper*, no. 14607.

Kordes, J., Bolsinova, M., G. Limpens, and R. Stolwijk, 2013, Resultaten PISA 2012: Praktische kennis en vaardigheden van 15-jarigen; Nederlandse uitkomsten van het Programme for International Student Assessment (PISA) op het gebied van wiskunde, natuurkunde en leesvaardigheid in het jaar 2012, Arnhem: CITO.

Lee, D., and T. Lemieux, 2010, Regression discontinuity designs in economics, *Journal of Economic Literature*, vol. 48, pp. 281–355.

Leuven, Edwin, and Hessel Oosterbeek. 2004. Evaluating the effect of tax deductions on training. *Journal of Labor Economics*, vol. 22, no.1, pp. 461–488.

McCrary, J., 1998, Manipulation of the running variable in the regression discontinuity design: A density test, *Journal of Econometrics*, vol. 142, no. 2, pp. 698-714.

Ministry of Education, 2007, Actieplan Leerkracht van Nederland: beleidsreactie op het advies van de commissie leraren.

OECD, 2014, TALIS 2013 Results: An International Perspective on Teaching and Learning.

Rivkin, S., E. Hanushek, and J. Kain. 2005, Teachers, Schools and Academic Achievement, *Econometrica*, vol. 73, no. 2, pp. 417–58.

Rockoff, J., 2004, The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data, *American Economic Review*, vol. 94, no. 2, pp. 247-252.

Schochet, P., T. Cook, J. Deke, G. Imbens, J. Lockwood, J. Porter, and J. Smith, 2010, Standards for Regression Discontinuity Designs, Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Schwerdt, G., D. Messer, L. Woessmann, and S. Wolter, 2012, The impact of an adult education voucher program: evidence from a randomized field experiment, *Journal of Public Economics*, vol. 96, pp. 569-583.

Staiger, D., and J. Stock, 1997, Instrumental Variables Regression with Weak Instruments, *Econometrica*, vol. 65, no. 3, pp. 557-586.

Trochim, W., 1984, Research Design for Program Evaluation: The Regression-discontinuity Design. Beverly Hills, CA: Sage Publications.

Vink, R., H. Marien, and A. Vloet, 2012, Tijd voor (na)schooling: Tweede rapportage evaluatie (na)schooling en de lerarenbeurs voor scholing, IVA Beleidsonderzoek en Advies.

Wiswall, M., 2013, The dynamics of teacher quality, *Journal of Public Economics*, vol. 100, pp. 61-78.

Chapter 4

Bonhomme, S., G. Jolivet, and E. Leuven, 2015, School characteristics and teacher turnover: assessing the role of preferences and opportunities, *Economic Journal* (forthcoming), retrieved from http://onlinelibrary.wiley.com/doi/10.1111/ecoj.12279/epdf

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff, 2002, Initial matches, transfers and quits: career decisions and the disparities in average teacher qualifications across schools, retrieved from http://www.teacherpolicyresearch.org/ResearchPapers/tabid/103/Default.aspx.

Chevalier, A., P. Dolton, and S. McIntosh, 2007, Recruiting and retaining teachers in the UK: An analysis of graduate occupation choice from the 1960s to the 1990s, *Economica*, vol. 74, pp. 69-96.

Clotfelter, C., H. Ladd, and J. Vigdor, 2008, Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina, *Journal of Public Economics*, vol. 92, no. 5-6, pp. 1352-1370.

Clotfelter, C., H. Ladd, and J. Vigdor, 2010, Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects, *Journal of Human Resources*, vol. 45, no. 3, pp. 655-681.

Clotfelter, C., H. Ladd, J. Vigdor, 2011, Teacher mobility school segregation and pay-based policies to level the playing field, *Education Finance and Policy*, vol. 6, no. 3, pp 399-438.

Dolton, P., 1990, The economics of UK teacher supply: the graduate's decision, *Economic Journal*, vol. 100, pp. 91-104.

Dolton, P., 2006, Teacher supply, in: Hanushek en Welch (eds), *Handbook of Economics of Education*, vol. 2, chapter 18.

Dolton, P., and W. van der Klaauw, 1995, Leaving teaching in the UK: a duration analysis, *Economic Journal*, vol. 105, pp. 431-444.

Dolton, P. and O. Marcenaro-Gutierrez, 2011, If You Pay Peanuts Do You Get Monkeys? A Cross-Country Analysis of Teacher Pay and Pupil Performance, *Economic Policy*, Vol. 26, no. 65, pp. 5-55.

Gilpin, G.A., 2011, Re-evaluating the effect of non-teaching wages on teacher attrition, *Economics of Education Review*, vol. 30, no. 4, pp. 598-616.

Goldhaber, D., and D. Brewer, 2000, Does teacher certification matter? High school teacher certification status and student achievement, *Educational Evaluation and Policy Analysis*, vol. 22, no. 2, pp. 129-145.

Greaves, E., and L. Sibieta, 2014, Estimating the effects of teacher pay on pupil attainment using boundary discontinuities, *IFS Working Paper*, W14/03.

Hanushek, E., J. Kain & S. Rivkin, 1999, Do Higher Salaries Buy Better Teachers?, *NBER Working Paper*, no. 7082.

Hanushek, E., J. Kain, and S. Rivkin, 2004, Why public schools lose teachers, *Journal of Human Resources*, vol. 39, no. 2, 326-354.

Hendricks, M., 2014, Does it pay to pay teachers more? Evidence from Texas, *Journal of Public Economics*, vol. 109, pp. 50‑63.

Hendricks, M., 2015, Towards an optimal teacher salary schedule: designing base salary to attract and retain effective teachers, *Economics of Education Review*, vol. 47, pp. 143-167.

Heyma, A., D. de Graaf, and C. van Klaveren, 2006, Exploratie van beloningsverschillen in het onderwijs 2001-2004, Stichting Economic Onderzoek (SEO).

Imazeki, J., 2005, Teacher salaries and teacher attrition, *Economics of Education Review*, vol. 24, no. 4, pp 431-449.

Manski, C., 1987, Academic ability, earnings, and the decision to become a teacher: Evidence from the National Longitudinal Study of the High School Class of 1972: University of Chicago Press.

Ministry of Education, 2015, Functiemix, Letter to parliament on May 29 2015.

Murnane, R., J. Singer, and J. Willett, J., 1989, The influences of salaries and "opportunity costs" on teachers' career choices: Evidence from North Carolina, *Harvard Educational Review*, vol. 59, no. 3, pp. 325-34.

Reed, D., K. Rueben, and E. Barbour, 2006, Retention of new teachers in California, retrieved from http://www.ppic.org/content/pubs/report/R_206DRR.pdf.

Researchned, 2015, Invulling en inrichting van een tegemoetkoming studiekosten lerarenopleidingen voortgezet onderwijs.

Rickman, D., H. Wang, and J. Winters, 2015, Adjusted state teacher salaries and the decision to teach, *IZA Discussion Paper*, no. 8984

Staiger, D., and J. Stock, 1997, Instrumental variables regression with weak instruments, *Econometrica*, vol. 65, no. 3, pp. 557-586.

Steeg, M. van der, and R. van Elk, 2015, The effect of schooling vouchers on higher education enrollment and completion of teachers: A regression discontinuity analysis, *CPB Discussion Paper*, no. 305.

Wolter, S., and S. Denzler, 2003, Wage elasticity of the teacher supply in Switzerland, *IZA Working Paper*, no. 733, Bonn.

Chapter 5

Attenasio, O., E. Fitzsimons, A. Gomez, M. Gutiérrez, C. Meghir, and A. Mesnard, 2010, Children’s Schooling and Work in the Presence of a Conditional Cash Transfer Program in Rural Colombia, *Economic Development and Cultural Change*, vol. 58, no. 2, pp. 181-210.

Angrist, J. & V. Lavy, 2009, The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial, *American Economic Review*, vol. 99, no. 4, pp. 1384–1414.

Bertrand, M., and J. Pan, 2013, The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior, *American Economic Journal: Applied Economics*, vol. 5, no. 1, pp. 32-64.

Bettinger, E. and R. Baker, 2013, The Effects of Student Coaching: An Evaluation of a Randomized Experiment in Student Advising, *Educational Evaluation and Policy Analysis*, vol. 35, no. 3.

Card, D., 1999, The Causal Effect of Education on Earnings, *Handbook of Labour Economics*, vol. 3, pp. 1801-1863.

Card, D. and T. Lemieux, 2000, Adapting to Circumstances: The Evolution of Work, School,and Living Arrangements among North American Youth, http://www.nber.org/chapters/c6805.

Carneiro, P., and J. Heckman, 2003, Human Capital Policy, *NBER Working Paper*, no. 9495.

Clark, D., 2011, Do Recessions Keep Students in School? The impact of youth unemployment on enrolment in post-compulsory education in England, *Economica*, vol. 78, Issue 211, pp. 523-545.

Dearden, L., C. Emmerson, C. Frayne, and C. Meghir, 2009, Conditional cash transfers and school dropout rates, *Journal of Human Resources*, vol. 44, no. 4, pp. 828-857.

Dynarski, M., P. Gleason, A. Rangarajan, and R. Wood, 1998, Impacts of dropout prevention programs, Mathematica Policy Research, Princeton (New York).

Ecorys, 2009, MKBA Voortijdig schoolverlaten, Rotterdam.

Grossman, J., and J. Tierney, 1998, Does Mentoring Work? An Impact Study of the Big Brothers Big Sisters Program, *Evaluation Review*, vol. 22, no. 3, pp 402-426.

Harmon, C., H. Oosterbeek, and I. Walker, 2003, The Returns to Education: Microeconomics. *Journal of Economics Surveys*, vol. 17, pp. 115-155.

Heckman, J.J., J. Stixrud, and S. Uruza, 2006, The effects of cognitive and non-cognitive abilities on labor market outcomes and social behavior, *NBER Working Paper*, no. 12006.

Herrera, C., J. Grossman, T. Kauth, A. Feldman, J. McMaken, and L. Jucovy, 2007, The big brothers big sisters school-based mentoring impact study, Public/Private Ventures, Philadelphia.

Johnson, A., 1999, Sponsor-A-Scholar: Long-term impacts of a youth mentoring program on student performance, Princeton: Mathematica Policy Research, Inc.

Lavecchia, A., H. Liu, and P. Oreopoulos, 2014, Behavioural economics of education: possibilities and progress, *NBER Working Paper*, no. 20609.

Lee, 2002, Trimming for bounds on treatment effects with missing outcomes, *Center for Labor Economics Working Paper*, no. 38, University of California, Berkeley.

Lee, 2009, Training, wages, and sample selection: estimating sharp bounds on treatment effects, *Review of Economic Studies*, vol. 76, pp. 1071-1102.

Lleras-Muney, A., 2005, The relationship between education and adult mortality in the United States, *Review of Economic Studies*, vol. 72, no. 1, pp. 189-221.

Lochner, L., and E. Moretti, 2004, The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports, *American Economic Review*, vol. 94, no. 1, pp. 155-189.

Machin, S., O. Marie, and S. Vujic, 2012, Youth crime and education expansion, *IZA Discussion Paper*, no. 6582.

Meschi, E., J. Swaffield, and A. Vignoles, 2011, The relative importance of local labour market conditions and pupil attainment on post-compulsory schooling decisions, *IZA Discussion Paper*, no. 6143.

Millenky, M., D. Bloom, and C. Dillon, 2010, Making the Transition
Interim Results of the National Guard Youth ChalleNGe Evaluation, MDRC.

Ministry of Finance, 2009, Lange termijn discontovoet: aanvullend voorschrift (long-term discount rate: supplementary prescriptions, Brief aan tweede kamer (letter to house of parliament).

OECD, 2012, Education at a Glance 2012, OECD Indicators, Paris.

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

Oreopoulos, P., R. Brown, A. Lavecchia, 2014, Pathways to Education: an integrated approach to helping at-risk high school students, *NBER Working Paper*, no. 20430.

Rice, P., 1999, The impact of local labour markets on investment in further education: evidence from the England and Wales youth cohort courses, *Journal of Population Economics*, vol. 12, no. 2, pp. 287-312.
Rivkin, S.G., 1995, Black/White differences in Schooling and Employment, *Journal of Human Resources*, vol 30, no. 4, pp. 826-852.

Rodríguez-Planas, N, 2012a, Longer-Term Impacts of Mentoring, Educational Services, and Learning Incentives: Evidence from a Randomized Trial in the United States, *American Economic Journal: Applied Economics*, vol. 4, no. 4, pp. 121–139.
Rodríguez-Planas, N, 2012b, School and Drugs: Closing the Gap, Evidence from a Randomized Trial in the US, *IZA Discussion Paper*, no. 6770.

Ross, R., S. White, J. Wright, and L. Knapp, 2013, Using behavioral economics for post-secondary success, Ideas42.

Schirm, A., E. Stuart, & A. McKie, 2006, The Quantum Opportunity Program demonstration: Final impacts, Washington DC: Mathematica Policy Research.

Schultz, T., 2004, School subsidies for the poor: Evaluating the Mexican Progresa poverty program, *Journal of Development Economics*, vol. 74, no.1, pp. 199–250.

Sinclair, M., S. Christenson, D. Evelo, and C. Hurley, 1998, Dropout prevention for youth with disabilities: Efficacy of a sustained school engagement procedure, *Exceptional Children,vol. 65, no.* 1, pp. 7–21.

Sinclair, M., S. Christenson, M. Thurlow, 2005, Promoting school completion of urban secondary youth with emotional or behavioral disabilities. *Exceptional Children, vol. 74, no. 4*, pp. 465–482.

Essays on Teacher Quality and Coaching

Teacher quality is key to the performance of pupils in education. Improvements in teacher quality can therefore generate large returns. It is less clear however what drives teacher quality and how the quality of teachers can be improved. This dissertation aims to provide more insight into the determinants of teacher quality and the effectiveness of policies that aim to improve teacher quality. The first paper examines the relationship between teacher evaluations and pupil performance gains in primary education. It is shown that the score on a detailed observation rubric measuring pedagogical, didactical and classroom organization competences of teachers significantly predicts pupil performance gains on standardized tests in math, reading and spelling. The observation rubric particularly seems to have potential to identify the weaker teachers. The second paper investigates the effects of schooling vouchers for teachers by employing a fuzzy regression discontinuity design. Effects of voucher assignment on both higher education enrollment and completion rates are in the order of 10 to 20 percentage points, suggesting substantial crowding out. The third paper investigates the effects of higher teacher pay for secondary school teachers on their teacher retention decision and enrollment in additional schooling. This is done by exploiting regional variation in teacher pay that is induced by the introduction of a new teacher remuneration policy that provided schools in an urbanized region with extra funds to place a larger share of their teachers in a higher salary scale. No effects are found on the probability of remaining in the teaching profession. The policy however succeeded in keeping a slightly larger share of teachers in the targeted region. In addition, the findings suggest that the policy slightly increased teachers' participation in continuous schooling. The fourth paper investigates the effect of an intensive coaching program aimed at reducing school dropout rates among students in post-secondary vocational education. The coaching program was set up as a randomized experiment. I find that one year of coaching reduced school dropout rates by more than 40 percent. Cost-benefit analysis suggests that one year of coaching is likely to yield a net social gain.

Marc van der Steeg was born in 1979 in Leiderdorp. He studied Economics at Erasmus University Rotterdam from 1997 to 2003. He obtained his MA degree in 2003. His master thesis was on India's vulnerability to macroeconomic shocks for which he went to India to the Institute of Economic Growth in Delhi for half a year. He joined CPB Netherlands Bureau for Economic Policy Analysis in 2004 and has been working there till the end of 2014. His work at CPB was mainly on education and education policy evaluation, but he has also worked on innovation and science policy. Since the end of 2014 he is working at the Dutch Ministry of Education, Culture and Science as a senior advisor at the Knowledge Directorate. In 2012 he started writing his PhD thesis.