The Effects of Targeted Learning Support

Evidence from a Regression Discontinuity Design

Gaute Eielsen

Supervisor:
Prof. Edwin Leuven

Master of Philosophy in Economics

Department of Economics

University of Oslo

May, 2014

The Effects of Targeted Learning Support

Evidence from a Regression Discontinuity Design

2014

The Effects of Targeted Learning Support:
Evidence from a Regression Discontinuity Design

Gaute Eielsen

http://www.duo.uio.no

Print: Reprosentralen, University of Oslo

Preface

This thesis is part of the ongoing evaluation of the "Ny GIV" initiative financed by the Norwegian Ministry of Education and Research. I extend the regression discontinuity analysis in Eielsen et al. (2013), the first of two evaluation reports for the Ministry, building on the analyses presented in that report. In the evaluation I held responsibility for the regression discontinuity analysis in close cooperation with Lars J. Kirkebøen. I am very grateful to Kirkebøen for his mentoring in applied research. I am also very grateful to my supervisor Edwin Leuven for his support and excellent supervision throughout the process. I would also like to thank my colleagues and fellow investigators on the evaluation project Marte Rønning and Oddbjørn Raaum, and the Ministry of Education and Science for comments on previous work. My family and friends also deserve a big thanks for the support they offered throughout a period of absent-mindedness on my part. And last but not least, thanks to Cormac Mangan, for great help and jokes at the end. I am responsible for any errors or omissions.

Gaute Eielsen, Oslo, May 2014

Abstract

This thesis evaluates the short-term effects of a Norwegian policy that aims to increase upper secondary education completion rates. The evaluated program provides learning support to low-performing students at the end of lower secondary school, seeking to improve their basic skills in reading, writing and numeracy. The explicit target group of the program is the bottom ten percent in the average grade distribution. However, the assignment rule has been interpreted differently, creating institution-specific thresholds that determine the participation offers to the students. I develop an approach to identify these thresholds that may also prove useful for other evaluations of targeted policies where lower level administrative units have implemented rules independently. For a relatively small sample the necessary assumptions for a regression discontinuity design are credible. I find no evidence of effects from the program.

Contents

List of Tables

List of Figures

1 Introduction

Low upper secondary school completion rates are a persistent cause of concern amongst policymakers in most high-income countries. Currently, one in four young people in OECD countries will *not* have passed one kind of upper secondary school by their 25th birthday (OECD, 2013).[1] Failure to complete secondary education comes at a great cost to both the individual and the society at large (Oreopoulos, 2007). For the individual, not only do lifetime earnings increase with additional schooling, there are also a number of nonpecuniary effects of education such as making better decisions about health, marriage and parenting style (Oreopoulos and Salvanes, 2011).

In Norway there has also been a growing concern over low and late completion. The share of a cohort completing secondary education within 5 years of finishing lower secondary school has been relatively stable at around 70 percent over the last decade.[2] The average rate however differ substantially by earlier performance as measured by final assessment grades at the end of lower secondary school. For the 10 percent lowest-performing students the average completion rate has been relatively stable at 16 percent over the last 6 years.[3] For the second and third deciles the corresponding figures are 35 and 50 percent, while for the top half of the distribution 90 percent have completed within 5 years. This association between earlier performance and the probability of completion is also found for the US, the UK and New Zealand (Falch et al., 2011).

In 2010 the Norwegian Ministry of Education and Science initiated several policies under the name "Ny GIV" to increase upper secondary completion rates.[4] A central part of the initiative, studied in this paper, is a remedial program targeting low-performing students at the end of their 10th academic year, the last compulsory year in school. Specifically the target group was the 10 percent lowest-performing students as judged by their first term GPA in 10th grade. The program aimed to increase basic skills in reading, writing and numeracy, and is generally implemented as adapted instruction in smaller groups. This is a substitute to ordinary classes, extra instruction time is not added.

This thesis analyzes the implementation and effects of the remedial program on short-term academic outcomes and progress through the first two years of upper secondary. Doing

[1] There are a number of difficulties comparing completion rates across countries and the share by some age is the most comparable. Problems still remain with different definitions of "completion" across countries and very different age profiles for completion. For a discussion see Lyche (2010).

[2] The theoretical duration for the academic and vocational study tracks is 3 and 4 years, respectively.

[3] Figure A.1 shows these completion rates using the complete cohorts finishing lower secondary school the years 2002-2007.

[4] The Ministry set a target of increasing the overall 5 year completion rate to 75 percent within year 2015 (Utdanningsdirektoratet, 2013).

so, I make two contributions to the literature. First, I develop an approach to find unknown cutoffs varying between units (here, schools or municipalities) for assignment to treatment. The program is explicitly targeted towards the lowest-performing 10 percent. However, this has been interpreted differently by different schools and municipalities, resulting in some schools having no clear cutoff. Other schools have cutoffs at unknown values of first term GPA, which can in turn be defined in different ways. The search procedure builds on the same idea used when looking for structural breaks in time-series econometrics, and is used by Card et al. (2008) to find "tipping points" in neighborhood population flows. However, to my knowledge, it has not previously been applied in the context of a policy evaluation. Although the search procedure should be considered work in progress, this application may prove useful in contexts where there exist rules that are open to different interpretations by different administrative units, resulting in effective variation in the assignment thresholds across units. Through employing a method to convincingly identify the rule applied, then it may still be possible to draw inferences in these contexts that were previously regarded as too "messy".

The second contribution is to use the identified threshold in a sub-sample of schools to estimate the causal effect of the remedial program on the outcomes of interest.

The evaluation compares students "just" below and a above a certain cutoff value in the first term grade point average (GPA) distribution. The idea being that while the students just above this cutoff have a much lower probability of receiving the intervention, they are similar in both observed and unobserved characteristics to those just below, and therefore qualify as a valid control group. Participation in the program is voluntary, therefore actually receiving the treatment is not a deterministic function of the first term GPA. This data generating process is what is known in the literature as a fuzzy regression discontinuity design (RDD). It depends on two crucial elements; the first is what generates the design: That actual implementation in the schools caused a discontinuity in the probability of receiving the treatment at some value of the first term GPA. The second is the key identifying assumption, first formalized in Hahn et al. (2001), that the potential outcomes are continuous in GPA at the discontinuity. In other words, there are no other factors that change discontinuously at the cutoff other than the difference in treatment probability. This assumption might seem strong, but the appeal of a regression-discontinuity design over other non-experimental evaluation strategies, such as difference-in-differences and (other types of) instrumental variable approaches, is that the implied local randomization can be verified much in the same way as a randomized controlled trial. Where in an experiment (globally) the observable characteristics should be balanced between the treated and the control group, this should be the case locally for students below and above the cutoff in a RDD (Lee, 2008).

If the identifying assumption holds, target group membership i.e. having a first term GPA equal to or lower than the cutoff, can be used as a valid instrument for participation. If being in the target group at least does not reduce the probability of participation (monotonicity), and the instrument has no independent effect on the outcomes (exclusion restriction), we can identify the local average treatment effect for the students who participate because they are in the target group (the compliers) in the proximity of the cutoff (Imbens and Lemieux, 2008).

The search procedure leads me to a sample of schools in the municipality of Stavanger, where there is a clear discontinuity in treatment probability and the continuity assumption seem to hold. For this sample I find no evidence of effects of the program on grades at the end of the final year of lower secondary school or in the first year of upper secondary school. Nor do I find any evidence of impact on progression through upper-secondary school. However, because of the limited precision, I cannot reject that there are effects of economical interest on these outcomes.

The thesis is structured as follows: In Section 2 I give a brief review of the relevant literature. Section 3 describes the institutional background, program studied, its participants, the data sources and finally applies the search algorithm to identify assignment rules. Section 4 develops the empirical strategy and the effect estimators. Section 5 presents and discusses the results from the estimations, while Section 6 explain in further detail why I cannot find any effects of the program, before Section 7 concludes.

2 Literature

In the economic literature of life cycle skill formation outcomes such as academic achievement and educational attainment are often modelled as a function of a set of skills, effort and various purchased inputs.[5] In this framework, social policies have an effect on outcomes by affecting skills such as cognitive ability and motivation or the effort of the student. In an influential study Carneiro and Heckman (2003) review the empirical evidence of policies that seek to improve various socioeconomic outcomes for disadvantaged children and adolescents and conclude that 1) early interventions are more effective than later interventions and 2) that personality skills are more malleable at earlier ages and that these can be as important determinants of later outcomes as cognitive skills.[6] There is a growing consensus that academic achievement and graduation rates are among the outcomes most effectively improved

[5] For a recent review see Heckman and Mosso (2014).

[6] Cognitive skills include such skills as memory and processing of new information while personality skills are among the noncognitive skills found to be important determinants of future socioeconomic outcomes.

by early interventions (Cook et al., 2014),[7] but there are a limited number of studies of remedial programs targeting adolescents that find positive impacts.

Lavy and Schlosser (2005) investigate the effect of providing individualized extra teaching to small groups of low-performing upper secondary students, finding that this increases graduation rates by 3.3 percentage points at the school level, implying an improvement of 6 percent.

De Haan (2012) studies a Dutch remedial program where schools get additional funding for each low-performing student. Non-parametrically bounding the effect she finds that graduation rates increase by at least 4 percentage points and reading and math performance also improve.

Perhaps most closely related to this study, Cortes et al. (2013) investigate an algebra policy implemented in Chicago in 2003 where students with achievement below the national median on an eighth grade exam in mathematics are assigned to algebra courses with double instructional time in ninth grade. Using a regression discontinuity design, they find sizable effects of the double-dosing in algebra on high school graduation rates, college entrance exam scores, and college enrollment rates. The intervention was most successful for students with relatively low reading skills.

Finally, a recent randomized experiment of an intervention that combines behavioral therapy with individualized academic remediation to 9th and 10th graders, also in Chicago public high schools, finds surprisingly large effects. Maths grades are reported to have improved by 0.67 of a control group standard deviation, and the expected graduation rate increased by 14 percentage points. Although it remains to be seen if these effects can be reproduced in the ongoing scaling up of the program, the cost-effectiveness of this program is much better than most other interventions targeting adolescents (Cook et al., 2014).

There is a large literature that more indirectly sheds light on the potential impacts of the program. The program implies a reduction in class size for both treated students and the remaining students in the cohort, which has been studied intensively empirically. Hanushek (1997) concludes in an influential review of this literature that there is not consistent evidence of positive impacts from a reduction of class-size, while Krueger (2003) reviews the same evidence concluding that there is a "subtle" positive impact. In a Norwegian context Leuven et al. (2008) find no effects on lower secondary school performance. Fredriksson et al. (2013) study the long-term effects of smaller class size over the last three years of primary school in Sweden and find that it not only improves non-cognitive and cognitive ability at age 16, but also improves secondary school completion rates and adult earnings. The intervention also

[7] Cook et al. (2014) argue, however, that this conclusion might be premature based on their findings.

changes the classroom composition, which can have a causal effect (Leuven and Rønning, 2011; Van Ewijk and Sleegers, 2010; Duflo et al., 2011). Additionally the ministry intended to change the pedagogy used. Similar interventions have been found to improve student outcomes in primary schools in England and India (Machin and McNally, 2008; Banerjee et al., 2007). Related to this, the curriculum also changed which, according to Cortes et al. (2013), can have a positive effect. Finally, in a Norwegian context, Falch et al. (2013) study the effect of randomly assigned exam subjects on performance and subsequent educational choices. They find a substantial effect of being assigned to mathematics, and argue that the effect of short-term (in this case only three to six days) intensive and focused training can be large.

3 Background

In Norway, compulsory schooling encompasses 10 grades. Student starts school at age 6, and leave compulsory school the year they turn 16. After compulsory school most students continue to upper secondary school. Upper secondary education has different tracks. Some of these tracks are academic, generally consisting of three years in school and intended to prepare students for further studies. A second path is vocational, generally consisting of two years in school followed by two years as an apprentice, leading to a certificate of apprenticeship. While not compulsory, students have a right to attend upper secondary school, and almost all students enroll in upper secondary school. However, the share completing upper secondary within five years of enrollment has for several years been stable at about 70 percent (Utdanningsdirektoratet, 2013). Completion in this context means obtaining a diploma from upper secondary school.

3.1 The program

The program's Norwegian name "Overgangsprosjektet", translated "the Transition Project", reveals the objective of easing the transition from lower to upper secondary school for the targeted students. The Ministry of Education and Science explicitly stated that the lowest-performing ten percent in terms of first term grades within each municipality are the target group. These students are considered at high risk of dropping out before the end of the remaining 3 or 4 years of their secondary education.[8]

The lack of basic skills in literacy, writing and numeracy for these students are thought to

[8]See Figure A.1 for completion rates within 5 years of using the complete cohorts finishing lower secondary school the years 2002-2007.

be the key reason for the low completion rates. Thus, to prepare for upper secondary, instead of following the regular curriculum in regular classes, these students are taught these basic skills in smaller groups. However, while the intervention changes the classroom composition and possibly the methods and content of the teaching, training in basic skills is intended to replace instruction time in the corresponding subject, and thus not supposed to change the relative time spent across subjects.

The intensive learning support was rolled out in three waves starting in the spring of 2011, each encompassing approximately one third of all students. The second and third waves were rolled out in the spring of 2012 and 2013 respectively, thus by spring 2013 all lower secondary schools in Norway were actively participating in the program.

In a letter from the Ministry describing the intervention, the schools were given substantial freedom in how to implement the program, but some features are still shared across schools. To describe the nature of the program I rely on survey responses from the principals after the first year, reported in Sletten et al. (2011). The response rate for the principals was 88 percent. Students and teachers (both those teaching intensive training lessons and others) were also surveyed, but the response was lower at approximately 30 and 40 percent of the populations. For this reason I use mainly responses from the principals in the following.

In the average school 12 students were offered the program and 10 of these accepted the offer. In most schools the program acted as a substitute to regular classes and typically accounted for about 6 to 7 hours of the 30-hour school week. In a minority of schools the targeted students also received classes in addition to the 30-hour school week. The average duration was 13 weeks, with a minimum of ten weeks and maximum of 18 weeks. There was some variation across schools in which skills the students received training; 80 percent of the participants received training in literacy and writing; 90 percent in numeracy; such that 70 percent received training in all three competencies. In 95 percent of the schools the students were taught outside of the regular class in smaller groups. In smaller schools all students in the program were mainly kept in one group, while in larger schools about half decided to split into groups depending on the competency being taught.

The group size was typically 10 students, but with much variation across schools. Among the responding teachers many had previous experience with teaching low-performing students. Furthermore, as a part of the program selected teachers received five days training focusing on teaching such students. The surveyed teachers state that they adapted their teaching to fit the challenges of the targeted students, and the extra training is reported to have strengthened the ability of the teachers to increase the students' motivation.

While the program targeted the lowest-performing students, it also affected the other students. The consequences for the remaining students was a temporary reduction in class

size, reduced within-class heterogeneity in terms of performance and possibly a reallocation of teaching resources. The majority of teachers who themselves did not teach in the program reported that it was easier to provide lessons to the remaining students. Only a minority of the teachers reported that the regular classes suffered in terms of teacher resources in the program period. Except for the five-day training there were no additional resources provided to the schools during the program from the Ministry. However, about half the principals responding said they received additional funds supplied by the municipalities to hire teachers in relation to the project. There is no information of whether these funds covered the extra cost of the teachers needed to carry out the program, or how the schools who did not receive these funds managed to supply the necessary teachers.

The larger initiative also involved other initiatives in upper secondary school. Notably, the responsibilities of school and other public agencies to follow up students at risk of dropping out were clarified. However, this does not impact the validity of identification as these policies are not exclusive to the participants of the intensive training. The later interventions should, nevertheless, be taken into account when interpreting the external validity of the effects, as these could be conditional on an environment where struggling students have extra resources available.

3.2 Data

I use administrative register data from Statistics Norway, covering the complete cohorts of lower secondary graduates of 2003 through 2011 for this analysis. The intensive learning support was rolled out in three waves starting in the spring of 2011, as explained above. This means that I can study the achievement and progression of the first wave of the program. The data will later be extended with more cohorts. Each cohort consists of roughly 60 000 students. For these students all first-term and final grades from lower secondary school are available, as well as information on their transition from lower secondary to upper secondary and their progress through upper secondary school. Individual-level data on participation in the program has been collected by NOVA, as part of their mappings of the program (Sletten et al. (2011)). The mean, standard deviations and the number of observations are presented in Table 1 for all observations. Further details on the variables are in the notes.

Table 1: Summary statistics wave 1

	Mean	SD	N
Characteristics			
Share female	0.479	0.500	18084
Mother's schooling	13.036	3.926	17189
Father's schooling	12.744	4.100	16454
Share immigrant	0.077	0.266	18084
Share immigrant parents	0.074	0.261	18084
Prior achievement			
Avg. on 8th grade tests	-0.016	0.897	17182
GPA 1st term	3.817	0.806	17689
Math grade 1st term	3.391	1.166	17314
Norwegian grade 1st term	3.679	0.962	17007
Achievement			
GPA teacher grades	4.000	0.842	17918
Written exam grade	3.471	1.155	17122
On-time enrollment 1st year	0.972	0.164	18038
On-time completion 1st year	0.790	0.407	18038
GPA upper sec.	36.476	11.425	17405
On-time enrollment 2nd year	0.838	0.368	18038

Notes. *GPA 1st term* is the average of all subject grades (for most students this is 12 grades) set by the students' teachers at the end of the first term of 10th grade. *Math* and *Norwegian* grades make up two of the grades in *GPA 1st term*. *Avg. on 8th grade tests* is the average of three standardized grades from a national exam in 8th grade in English, Norwegian and Maths. *Mother's* and *Father's schooling* are the number of years of schooling of the mother and father of the student, respectively. *Share female/immigrant/immigrant parents* are all dummy variables equal to one if the student is female, an immigrant or has immigrant parents, respectively. The enrollment and completion variables equals one if the student has enrolled or completed, respectively; zero otherwise. *GPA teacher grades* is the average of all grades (for most 13 grades) set by the students' teachers at the end of lower secondary school (10th grade). *Written exam grade* is the average of the three exams most students undertake in English, Maths and Norwegian. *GPA upper sec.* is the average of all grades the first year of upper secondary school multiplied by ten.

3.3 First wave participants

The target group of the program was the 10 percent lowest-scoring students in each municipality as per first term GPA (Sletten et al., 2011). Table 2, which compares participating students with other students in the participating schools, shows that these differ from the remaining students. The participating students have lower first term performance, in particular in Maths, are more likely to be boys and have a more adverse family background.

Table 2: Comparison of participant and other students in the first wave schools

	(1) Participants mean/sd	(2) Non-participants mean/sd	(3) Difference b/se
GPA 1st term	2.863	3.840	-0.977**
	(0.592)	(0.957)	(0.015)
Missing grades 1st term	0.013	0.023	-0.010**
	(0.112)	(0.150)	(0.003)
Math grade 1st term	2.178	3.543	-1.365**
	(0.658)	(1.126)	(0.018)
Norwegian grade 1st term	2.767	3.792	-1.025**
	(0.712)	(0.928)	(0.018)
Avg. on 8th grade tests	-0.806	0.081	-0.887**
	(0.677)	(0.872)	(0.017)
Share female	0.404	0.488	-0.084**
	(0.491)	(0.500)	(0.012)
Mother's schooling	11.287	13.244	-1.957**
	(4.169)	(3.844)	(0.102)
Father's schooling	11.209	12.921	-1.713**
	(3.962)	(4.079)	(0.102)
Share immigrant	0.129	0.070	0.059**
	(0.336)	(0.256)	(0.008)
Share immigrant parents	0.123	0.068	0.055**
	(0.328)	(0.251)	(0.008)
Observations	1972	16112	18084

Notes. Mean values of each characteristic are shown in column (1) and (2) for participants and non-participants, respectively. Standard deviations are in parentheses. Column (3) tests each difference with Welch's t-test, allowing for the difference in sample size and variance. Data are for the students in schools included in the program the first year (the first wave). Standard errors are in parentheses. Stars indicate the significance level (* $p<0.10$, ** $p<0.05$).

Although the program targeted the bottom ten percent, there is a lot of variation in program participation across the average grades distribution. Figure 1 shows how the share of participants varies over the municipality-specific distribution of first term GPA.[9] This shows that schools were using other criteria than the average first term grade alone to select students to the program.

[9]Figure A.2 in the appendix shows the first term GPA distribution for all students and the participating students.

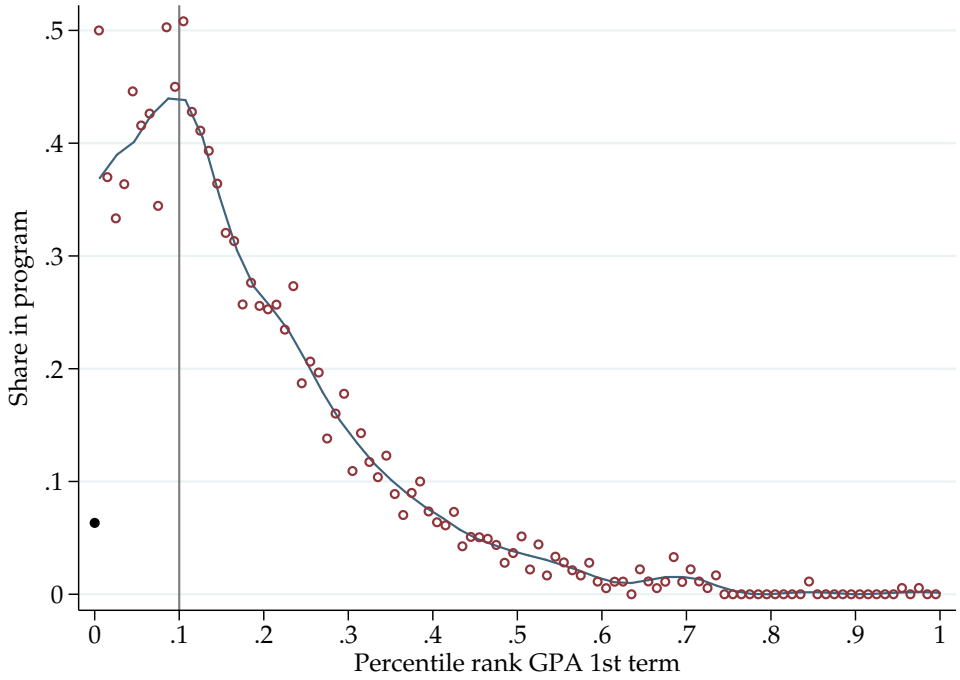

Figure 1: Program participation conditional on 1st term average grade

Notes. The x-axis shows the percentile rank, i.e. the percentage of average grades that are the same or lower, in the 1st term average grade distribution of each municipality. The solid circle indicates the percentage of participants missing 1st term grades. The hollow circles shows the mean percentage of studens participating conditional on the percentile rank point. On the vertical line and to the left are the 10 percent lowest-scoring pupils in each municipality. Also added is a fit estimated with a local linear regression weighted using the Epanechnikov kernel and a bandwidth of 2 percentile rank points. Data are for the students in schools included in the program the first year (the first wave).

Less than half of the target group, the 10 percent lowest-scoring students in each municipality, actually participates in the program. Within the first decile there is also variation, with the maximum participation rate of 50 percent around the 10th percentile and the minimum at 34 percent in the third. Estimating the conditional mean participation rate separately below and above the 10th percentile reveals no difference. There is no clear discontinuity either way.

There are several reasons why, in spite of the clear instruction from the Ministry, there is no clear discontinuity in participation around the 10th percentile. First, while the students should be selected based on first term grades, no clear advice was given on what weights should be attached to different subjects. All subjects could be given equal weight (as in Figure 1), or for example Maths and Norwegian grades could be given more weight, as some

coordinators of the programs report.

Second, some students were already receiving different kinds of special education. The Ministry explicitly stated that in such cases the program should only be offered if it was considered to be a better alternative. This seems unlikely given that these students already had an individually adapted curriculum and teaching. About 11 percent of 10th grade students have such individual programs. While these individuals cannot be identified in the data, they are likely overrepresented among the low-performers.[10] This may explain the relatively low training incidence below the 10th percentile.

Ten percent of the students in the first wave of the program participate in the training. This means, with some low-performing students not participating, that the schools include higher-performing students. With different shares of special needs students at different schools, this can therefore give rise to different participation thresholds.

Finally, schools or municipalities may determine participation on other criteria. There is, for example, anecdotal evidence that in some cases the selection of students for participation was based on the effect the teachers anticipated for a given student.

To conclude, some municipalities and schools probably chose students in a way that produced no discontinuities in the probability of participation. In these cases participating and non-participating students with similar first term GPA are not systematically different. In the next section I detail how I identify schools and municipalities that assigned students according to a local cutoff.

3.4 Searching for cutoffs and "strict" implementation

The directive of the Ministry of Education and Science suggested that all students below the 10th percentile would receive a treatment offer. For a given municipality, we can write this formally as

$$\tilde{d}_i = 1\{g_i \leq \tau_{10}\} \tag{1}$$

where $\tilde{d}_i$ is the binary offer variable, g_i the first term GPA of the student, τ_{10} the 10th percentile in the first term GPA distribution. Participation d_i then depends on the participation offer $\tilde{d}_i$ as follows

$$d_i = \gamma_0 + \gamma_1 \tilde{d}_i + u_i \tag{2}$$

[10] The number of subjects a student receives grades in may be a proxy for individual programs. Studying this, Eielsen et al. (2013) find that there are students with fewer grades over the entire GPA distribution, but that they are clearly overrepresented in the bottom. Furthermore, having few graded subjects reduces the probability of participation in the intensive training program for given GPA .

As explained above, municipalities could deviate from the 10th percentile rule, and use another threshold (if any). There were also different practices in terms of which grades made up the average grade, with five specific combinations reported by the local program administrators.[11] To investigate this possibility, I estimate for each municipality equations such as (2), while letting the threshold vary from the 1st to the 35th percentile in each of the five GPA distributions. The threshold that predicts observed treatment most accurately (the one with the highest R-squared), is then taken as the one the municipality applied.[12] This forms a course, municipality and cohort-specific assignment variable, which for every student is normalized to a cutoff of 0.

The same procedure is repeated at the school level, using the GPA distributions at the municipality level. This is to account for the possibility that there could be certain "strict" schools within a municipality that adhere to a (potential) percentile rule of the municipality.

Figure A.3 in the appendix shows that the percentiles that best explain program participation differ substantially, from the 5th to the 30th. For municipalities most fall in the range from 10 to 25, while for schools there is wider dispersion. How well the best models explain assignment also varies as shown in Figure A.5 (also in the appendix), but is overall rather low: Most schools have a share of explained variation (R^2) smaller than 0.6 and most municipalities smaller than 0.4. Figure 2 categorizes units by the share of variation explained, and shows program participation against the normalized assignment variable. There are clear differences in the discontinuities both at the school and municipality level, with a much larger drop for the strict municipalities.

[11] As part of the evaluation documented in Eielsen et al. (2013) we surveyed local administrators on their assignment practice.

[12] If assignment is strict, all students below the nth percentile would participate and the model would perfectly explain the variation in participation and thus yield an R^2 of 1.

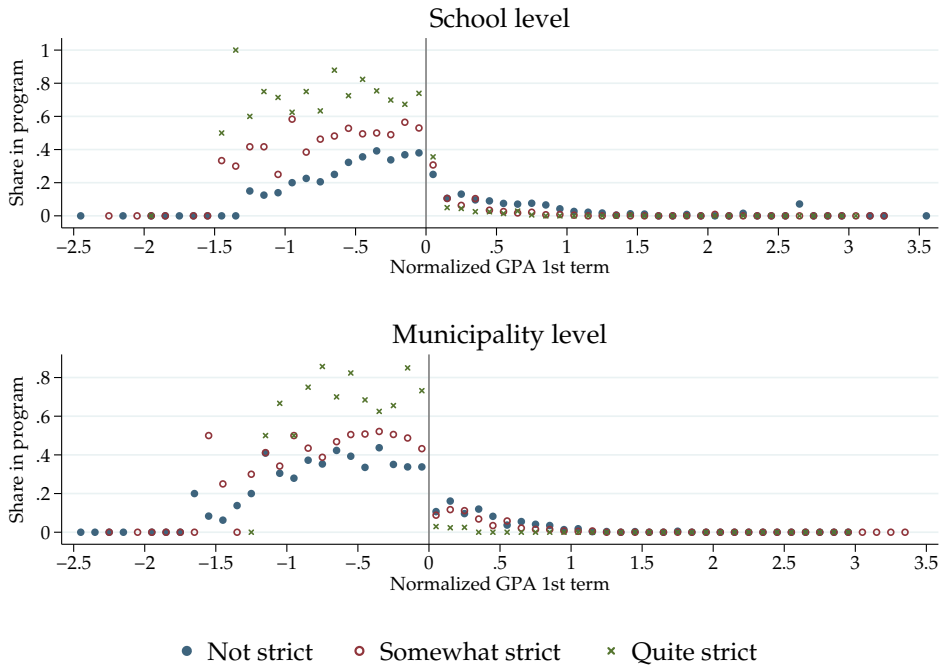

Figure 2: Probability of participation by "strictness" category

Notes. The y-axis shows the share in the program. The x-axis shows the normalized assignment variable for the best specification for all units. The mean participation rate for bins of 0.1 average grade-points is plotted at midpoints. In the upper panel the units are schools, while in the lower there are municipalities. The units are categorized by the share of variation explained. *Not strict* is defined as having a R-squared from the best specification in the interval [0,0.25]. Similarly, for *Somewhat strict* the R-squared is in the interval (0.25, 0.5], while for *Quite strict* in (0.5,1]

Among the municipalities Stavanger is the strictest, with an R-squared of 0.7. This matches well with reports from local administrators, as well as the plot of individual students in the schools in Figure 3. With the exception of one school (#9), the same municipality-specific cutoff at the 11th percentile predicts participation well. I therefore continue with the first cohort in Stavanger as my main estimation sample.[13] Table 1 in the appendix shows the summary statistics for this sample.

[13] As "School 9" looks to have a different practice I exclude that school from the sample. The R-squared from the estimation of the cutoff is the lowest in the municipality of Stavanger at 0.41. The results are not sensitive to this exclusion.

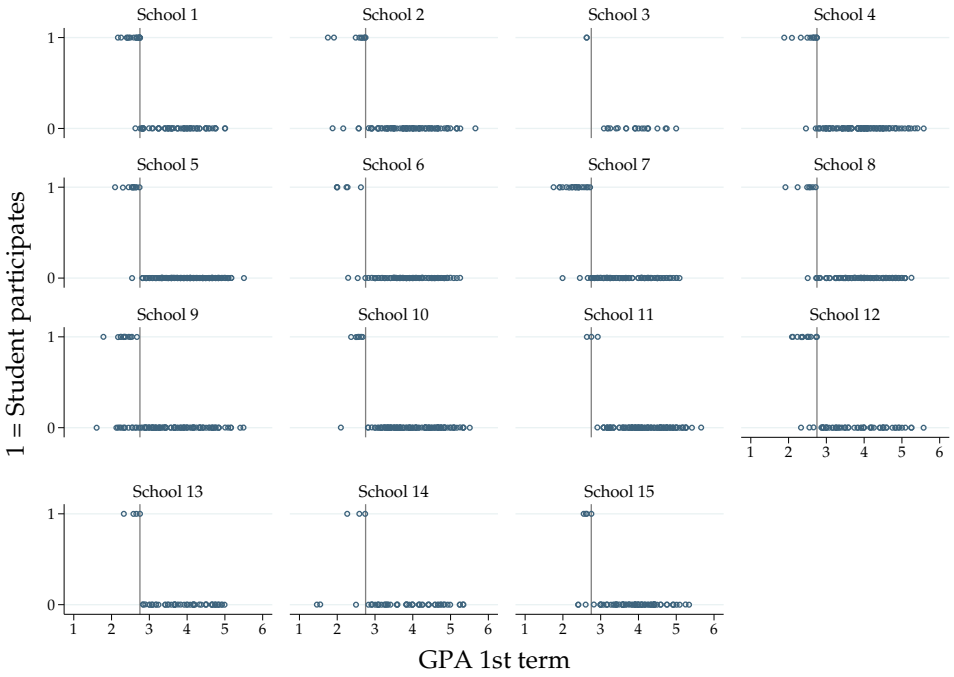

Figure 3: Assignment of students in Stavanger first cohort

Notes. Each school in the first wave of the program in Stavanger is plotted separately. The y-axes of all graphs shows whether a student participate, with y equal to one; zero otherwise. The x-axes shows the students' average first term grades plotted with some random noise (jitter) to show the relative weight of students along the axis. The vertical line indicates the estimated cutoff at the 11th percentile, or a GPA of 2.75.

Concerning the results from searching for school-specific cutoff, I do find some schools that seem to have a reasonably strict assignment, as shown in Figure A.5 in the appendix. I keep the schools with an R-squared larger than 0.5 as an alternative estimation sample for further inspection. This sample does at least consist of some actual discontinuities, as many of the strictest schools are found in Stavanger, but the algorithm can also have picked up spurious cutoffs when iterating over the large number of schools.

4 Empirical strategy

The challenge in estimating the causal effect of the intensive training program is addressing non-random selection into the program. Table 2 in the previous section showed that participants are different in many observable characteristics, including grades in Maths and Norwegian, which is to be expected when the targeted group is the first decile of the first term GPA distribution. Simply comparing students who attend with those who do not, will

likely result in effect estimates that are heavily downward biased.

To get credible causal effect estimates, the main identification strategy in this thesis relies on a directive from the Ministry of Education and Science stating that the bottom 10 percent of students should be offered the program. Sletten et al. (2011) report that most students accepted the program offer. If municipalities follow the rule-based assignment then there is a clear difference in the probability of participation across the cutoff that we can exploit.

This section starts with a presentation of this identification strategy, called the "fuzzy" regression discontinuity (FRD) design, and continues with a discussion of the estimation. I go on to assess whether the identifying assumptions are satisfied for these samples before I continue with a discussion of potential spillovers from the program and implications for the effect estimates. I conclude this section with an outline of an alternative identification strategy.

4.1 The effects of the intensive training program

The effect of the intensive training on an outcome y, say GPA at the end of the first year of upper secondary school, for student i can conceptually be defined by the difference in potential outcomes (Rubin, 1974). Let $y_i(1)$ be the GPA for the student if she participates, and $y_i(0)$ the GPA if she does not. The causal effect of the program for this student is then $y_i(1) - y_i(0)$. Depending on a student's treatment status we either observe $y_i(0)$ or $y_i(1)$, but never both. This is "the fundamental problem of causal analysis", coined by Holland (1986). The observed outcome, y_i, can be written in terms of potential outcomes as follows:

$$y_i = y_i(0) + d_i(y_i(1) - y_i(0)) \equiv \alpha + d_i\beta_i + \nu_i, \tag{3}$$

where $\beta_i \equiv y_i(1) - y_i(0)$, $\alpha = E[y_i(0)]$, $\nu_i = y_i(0) - E[y_i(0)]$ and $d_i = 1$ if student i participates, and is zero otherwise. Although we cannot estimate unit level treatment effects β_i, we can estimate *average* causal effects by comparing treated and untreated students who are on average identical.

The program was intended for the ten percent lowest-performing students as judged by their first term GPA in 10th grade, g_i. Students would thus receive a treatment offer if $g_i \leq c$, where c is the 10th percentile of the first term GPA distribution. Following (Hahn et al., 2001), I now discuss how to recover causal effects in the context of this treatment assignment mechanism. The probability of participation given g_i is defined as $Pr[d_i = 1 \mid g_i = g]$. The first requirement is that this probability is discontinuous at the 10th percentile cutoff c:

$$d^- \equiv lim_{\epsilon \uparrow 0} Pr[d_i \mid g_i = c + \epsilon] \neq lim_{\epsilon \downarrow 0} Pr[d_i \mid g_i = c + \epsilon] \equiv d^+ \tag{4}$$

The main identifying assumption is that the only thing that changes at the cutoff is treatment. This implies that average potential outcomes do not jump at the cutoff. More formally:

Assumption 1. *$E[y_i(0) \mid g_i = g]$ and $E[y_i(1) \mid g_i = g]$ are continuous at $g_0 = c$.*

This requires for example that students' average motivation does not change discontinuously at the cutoff. In practice the main threat to this assumption is that individuals sort around the cutoff. This may therefore seem like a strong assumption, but as long as there is an element of chance determining the assignment variable then there will be no self-selection *close* to the cutoff, even if students prefer one side of the cutoff over the other (Lee, 2008).

In the context of this study it seems plausible that there is a stochastic element to the first term average grade from the students' perspective, after all it depends on grading in several courses on multiple tests by different teachers. Schools may however "sort" students below or above the cutoff, perhaps based on perceived gains from the program. Assumption 1 implies however that students just below and above the cutoff should have the same predetermined characteristics. This provides a local balance test similar to the (global) one conducted between control and treated students in a randomized experiment. If sorting behavior by students and schools depends on expected benefits, and if we have access to predetermined characteristics that correlate with potential outcomes, then this should show up in the balance tests.

Now we can define a local intention to treat (ITT) parameter by looking at the difference in average outcomes on both sides of the cutoff c:

$$\beta^{ITT} = lim_{\epsilon \uparrow 0} E[y_i \mid g_i = c + \epsilon] - lim_{\epsilon \downarrow 0} E[y_i \mid g_i = c + \epsilon] \equiv y^- - y^+ \tag{5}$$

With perfect compliance, i.e. all students offered the program participated, this parameter equals the local average treatment effect.

With imperfect compliance, as is the case in this evaluation, Hahn et al. (2001) show that as long as crossing the threshold has a monotonous effect on treatment,[14] then we can identify the local average treatment effect (LATE) for the students induced to participate by the instrument, the so-called "compliers" (Angrist et al., 1996). It can be shown that the LATE is the ratio of the local ITT and the difference in treatment probability:

[14] This implies that there are no students who would not have participated with a a test score below the threshold, but who would have participated with a test score above the threshold.

$$\beta^{LATE} = \frac{y^- - y^+}{d^- - d^+} = E[\beta_i \mid \text{student } i \text{ is a complier, } g_i = c] \tag{6}$$

Note that this is the average effect of treatment for the sub-population that is 1) induced into the treatment if their score g_i falls below the threshold, and 2) has a GPA close to the 10th percentile in the distribution.

With heterogeneous effects of the program, and without further assumptions, this effect estimand is thus not valid for students that would get into the program regardless of their first term grades, nor those that would always decline an offer. This make intuitive sense as there are likely reasons for why some students accept an offer of participation and why others do not. With maximizing students one would expect the compliers to perceive their gains from treatment to be higher.

4.2 Estimation

The parameters derived above are the difference of the limits at each side of the cutoff. In practice there is however insufficient data for such local estimation, and I will need to use observations further away from the discontinuity in the estimations. In order to estimate the LATE I need estimates of the denominator and the numerator in Equation (6). I estimate the denominator, $d^- - d^+$, by regressing treatment d_i on target group membership $\tilde{d}_i$:

$$d_i = \mu_{j0} + \mu_{j1}\tilde{d}_i + f_j(g_i) + u_{ji} \tag{7}$$

where g_i is now normalized to 0 at the cutoff and $\tilde{d}_i = 1[g_i \leq 0]$. The estimate for the coefficient μ_{j1} is then the difference in probability of treatment in the sample, $\hat{d^-} - \hat{d^+}$. This probability is allowed to differ for the different j outcomes studied, as the population comprises of the students with non-missing values for each of the outcomes. To make sure that I capture the jump at the cutoff I need to control for a flexible function of the running variable $f_j(g_i)$.

Similarly I can estimate $y^- - y^+$ by estimating:

$$y_{ji} = \alpha_{j0} + \alpha_{j1}\tilde{d}_i + h_j(g_i) + v_i, \tag{8}$$

where the coefficient α_{j1} is the difference in sample averages of the observed outcomes at each side of the cutoff, $\hat{y^-} - \hat{y^+}$. This is interpreted separately as the estimator for the ITT parameter in Equation (5).

Taking the ratio of these two estimates gives the estimate for the LATE, which is equivalent to estimating the structural equation

$$y_{ji} = \beta_{j0} + \beta_{j1} d_{ji} + m_j(g_i) + \varepsilon_i, \tag{9}$$

using two-stage least squares and instrumenting d_i with $\tilde{d}_i$.[15]

The main challenge in practice is to specify the parametric models for the assignment variable $f_j(\cdot)$, $h_j(\cdot)$ and $m_j(\cdot)$, and because the identification is ultimately local, the restriction on the estimation sample around the cutoff. The nonparametric regression of program participation on the assignment variable for the main estimation sample, presented in the first graph in the upper left corner of Figure 4 below, suggests that a linear model on both sides is a good approximation to $f_j(\cdot)$. Similarly this also seems to be the case for $h_j(\cdot)$, judging the fits in Figure 7. I will thus estimate local linear regressions allowing the slope to differ at each side of the discontinuity in all equations presented above. So specifically for the structural equations, inserting for $m_j(g_i)$ for a bandwidth choice b yields:

$$y_{ji} = \beta_{j0} + \beta_{j1} d_{ji} + \beta_{j2} g_i + \beta_{j3} g_i \cdot \tilde{d} + \varepsilon_i \text{ for } -b \leq g_i \leq b \tag{10}$$

and similarly for Equation (7) and (8).

In my preferred specifications I will use a bandwidth of 1 average grade-point for all outcomes. This choice is based on the outcome-specific optimal bandwidths calculated for the different outcomes, all in the range of 0.7 to 1.15,[16] as well as inspection of Figure 7. To have one common bandwidth also eases comparisons of precision.[17]

In all models I use a triangle kernel function to weight the observations, in practice giving relatively more weight to observations closer to the cutoff. Finally, as the assignment variable is discrete there is the risk of introducing a random common component to the variance of all observations at the same values when we specify our model (Lee and Card, 2008). To correct for this I follow the recommendation of Lee and Card (2008) and cluster the sampling errors on these discrete values of the assignment variable.

[15] With the benefit of getting the standard errors for the estimates directly.

[16] Optimal bandwidths is calculated using the the Stata procedure rdob implementing the algorithm derived in Imbens and Kalyanaraman (2012)(Imbens, 2012).

[17] I assess the sensitivity of my estimates presenting ITT estimates for four other bandwidths, from a quarter of a grade point on each side of the cutoff to one and a half grade points in Table A.3 below. Further, Figure A.7 in the appendix shows the LATE estimates and their confidence intervals against a even wider range of bandwidths. Note also that the bandwidths are asymmetric when larger than 1 grade point, as there are only students within one grade-point below the cutoff.

4.3 Assessing the identifying assumption

The continuity assumption (Assumption 1) of the potential outcomes cannot be tested, but a consequence of the assumption is that baseline covariates should be balanced across the cutoff (Lee, 2008). If students are able to manipulate their first term GPA this should be revealed by balance tests. A second and more direct way to test for manipulation is to look at the density of the assignment variable (McCrary, 2008).

4.3.1 Local balance tests

Figure 4 shows how program participation and student characteristics change around the cutoff in the estimation sample. First, there is a clear discontinuity in program participation in the upper left plot, which drops from a stable level just below 80 percent to zero. The first requirement (Equation 4) for the design is satisfied for this sample.

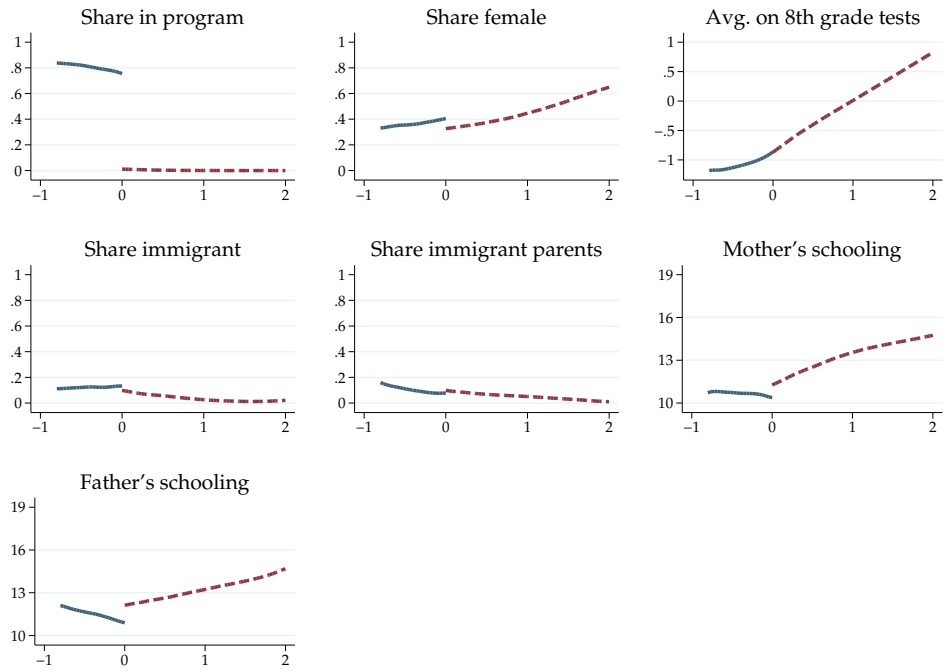

Figure 4: Balancing tests: Composition of student characteristics around cutoff

Notes. The fits are the smoothed values from local linear regressions of the first term GPA on participation, characteristics, and the prior achievement. All regressions estimated separately at each side of the the cutoff, weighted with a triangle kernel with a bandwidth of 1 average grade-point for all outcomes. The cutoff, normalized to zero, was identified by the search algorithm at the 11th percentile in the estimation sample.

Student performance, measured by performance on a national test in 8th grade, shows no

sign of discontinuities. On the other hand there is some indication of differences in the student composition with respect to gender and parental education. Table 3, presents estimates of the difference in characteristics across the cutoff while varying the bandwidth. For fathers' average education there is a significant difference for the larger bandwidths, but only at a ten percent level. With six characteristics this could be by chance, and the Wald test for a joint difference in the baseline characteristics is reassuring with a p-value of 0.44 for the preferred bandwidth.

Table 3: Composition of student characteristics around cutoff, main sample

	(1)	(2)	(3)	(4)	(5)
	.25	.50	.75	1.00	1.50
Share in program	0.742**	0.736**	0.737**	0.744**	0.758**
	(0.088)	(0.070)	(0.062)	(0.057)	(0.053)
Share female	0.187	0.107	0.092	0.078	0.091
	(0.187)	(0.119)	(0.098)	(0.086)	(0.075)
Avg. on 8th grade tests	-0.287	0.131	0.062	0.006	-0.072
	(0.398)	(0.203)	(0.154)	(0.130)	(0.110)
Share immigrant	-0.124	-0.040	0.025	0.034	0.032
	(0.153)	(0.089)	(0.070)	(0.058)	(0.049)
Share immigrant parents	0.149	0.025	-0.015	-0.020	-0.013
	(0.106)	(0.068)	(0.058)	(0.051)	(0.044)
Mother's schooling	-1.857	-0.298	-0.976	-0.886	-0.905
	(1.919)	(1.214)	(0.980)	(0.837)	(0.734)
Father's schooling	-2.119	-0.885	-1.270*	-1.232*	-1.119*
	(1.392)	(0.888)	(0.756)	(0.664)	(0.589)
Observations	171	311	456	608	919
Wald test of joint significance, all but 'Share in program'	7.395	2.607	5.062	5.897	7.574
p-value Wald test	0.286	0.856	0.536	0.435	0.271

Notes. Heteroskedasticity robust standard errors clustered at the discrete values of the assignment variable in parentheses. Stars indicate the significance level (* p<0.10, ** p<0.05). Data are for the students in the main estimation sample. Column (1) - (5) presents balance tests for bandwidths of .25 - 1.5 average grade-points. The cutoff, normalized to zero, was identified by the search algorithm at the 11th percentile. For means, standard deviations and explanations of variables see the summary statistics in Table A.1 with notes.

For the sample of strict schools Table 4 shows that there is a significant difference in probability of participation across the cutoff. The difference, however, depends more on the chosen bandwidth and is smaller (50 percentage points compared to 74 in the main sample for the preferred bandwidth of one average grade point). Graphic balance tests for this sample are shown in the appendix in Figure A.6. For the observed characteristics there is a

significant difference in average education of the students' mothers across the cutoff. This difference is significant at the five percent level for all presented bandwidths. The joint test is significant at a ten percent level and close to significant at a five percent level, suggesting that the algorithm might have picked up schools where there was in fact no rule-based assignment to the program. This suggests a violation of the continuity assumption, such that I cannot draw credible causal inference from this sample.

Table 4: Composition of student characteristics around cutoffs, alt. sample

	(1)	(2)	(3)	(4)	(5)
	.25	.50	.75	1.00	1.50
Share in program	0.170**	0.363**	0.445**	0.502**	0.568**
	(0.085)	(0.059)	(0.047)	(0.041)	(0.035)
Share female	0.074	0.095	0.051	0.027	0.024
	(0.086)	(0.062)	(0.051)	(0.045)	(0.039)
Avg. on 8th grade tests	0.044	-0.039	-0.055	-0.096	-0.120**
	(0.134)	(0.098)	(0.080)	(0.070)	(0.061)
Share immigrant	0.036	0.031	0.038	0.030	0.026
	(0.055)	(0.042)	(0.034)	(0.030)	(0.026)
Share immigrant parents	-0.040	-0.028	-0.030	-0.027	-0.023
	(0.056)	(0.035)	(0.028)	(0.025)	(0.021)
Mother's schooling	-1.788**	-1.098**	-1.117**	-0.970**	-0.724**
	(0.672)	(0.528)	(0.443)	(0.394)	(0.348)
Father's schooling	-1.192*	-1.031**	-0.664	-0.480	-0.322
	(0.656)	(0.508)	(0.425)	(0.380)	(0.338)
Observations	568	1097	1611	2109	3081
Wald test of joint significance, all but 'Share in program'	12.171	12.176	12.667	12.148	11.747
p-value Wald test	0.058	0.058	0.049	0.059	0.068

Notes. Heteroskedasticity robust standard errors clustered at the discrete values of the assignment variable in parentheses. Stars indicate the significance level (* p<0.10, ** p<0.05). Data are for the students in the alternative sample of schools identified as "Quite strict" (R-squared>0.5). Column (1) - (5) presents balance tests for bandwidths of .25 - 1.5 average grade-points. The cutoff, normalized to zero, was identified by the search algorithm at the 11th percentile. For means, standard deviations and explanations of variables see the summary statistics in Table A.1 with notes.

4.3.2 The first term GPA distribution

Studying the distribution of the assignment variable in Figure 5 in high resolution (bin width of 0.05 average grade-points) there does seem to be more mass to the left of cutoff, indicated by the vertical line. These peaks appear at regular intervals, thus also at values where there are no incentives for individuals to act strategically. This is explained by the data-generating

process of the variable: The number of subjects that enter first term GPA varies between individuals, with 12 being by far the most common number. As subject grades are integers, this will produce "heaps" at multiples of 1/12. The cutoff identified in Stavanger, 2.75, is such a multiple.

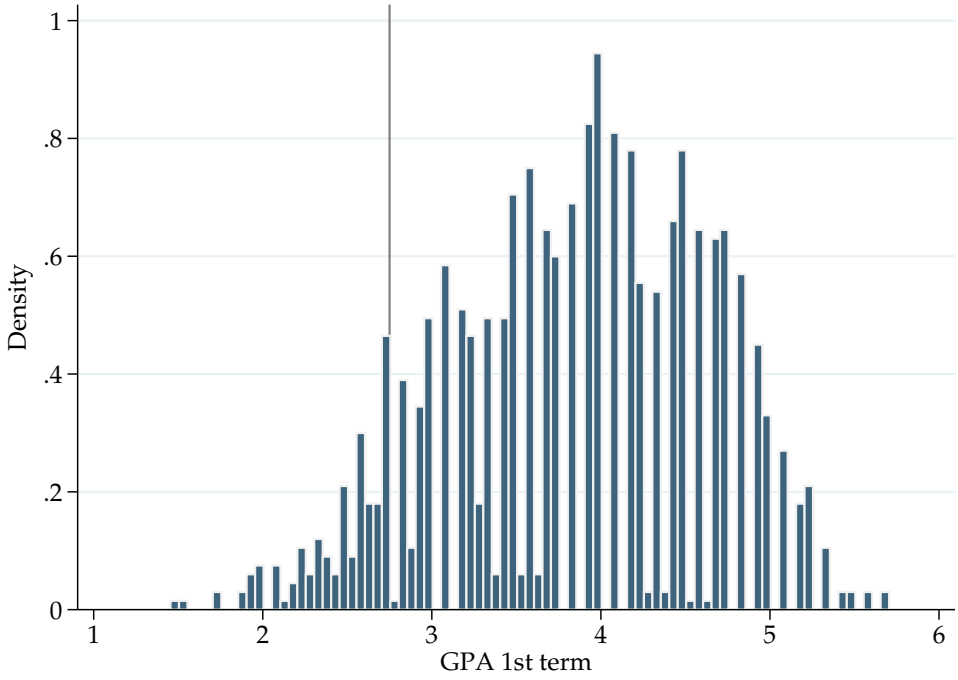

Figure 5: Distribution of first term GPA in the estimation sample

Notes. Distribution of the assignment variable for the first cohorts in Stavanger with a bin width of 0.05 first-term average grade-points. The vertical lines indicate the located cutoff at the 11th percentile.

Even in the absence of strategic behavior the bunching in the distribution could cause problems. Barreca et al. (2011) find that (non-random) heaping causes bias in the estimates of marginal returns to medical care for newborns in Almond et al. (2010).[18] Students with 12 grades could be systematically different. For one, they are less likely to be defined as special needs. I therefore follow Barreca et al. (2012) and plot the three potentially problematic covariates against first term GPA in Figure 6. There is no indication of any systematic differences between the heaps and the neighboring values.[19]

[18]Poorer hospitals are more likely to round off the birth weight of the newborn babies and thus the composition of babies at every multiple of a 100 grams are different from the neighboring values. The babies at the cutoff at 1500 grams are thus not comparable to those "just" above.

[19]The heaps are closer to the overall average, but this is natural with more observations making up the average characteristic at these values.

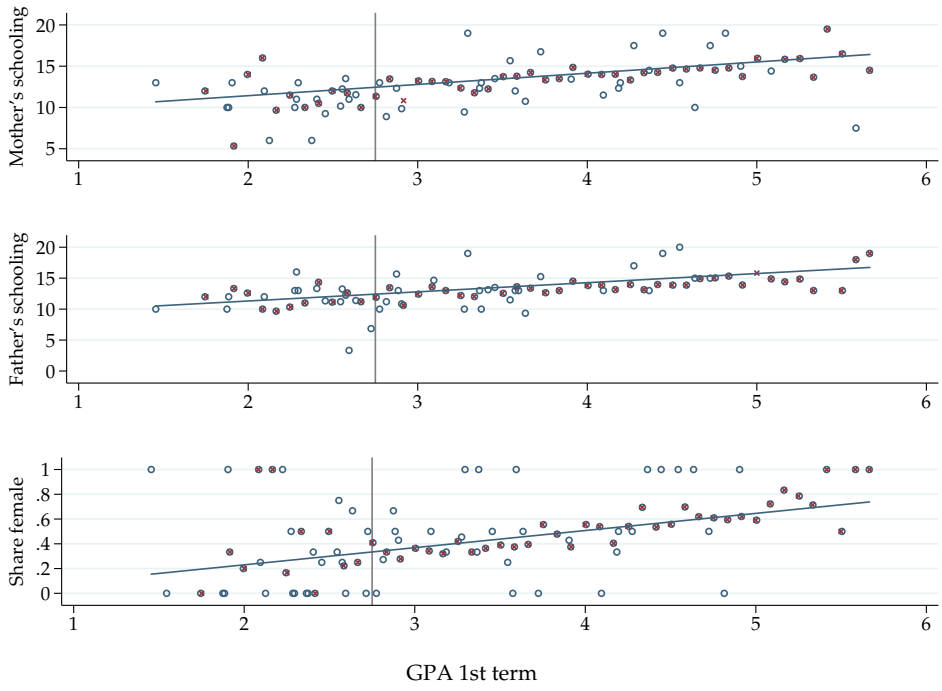

Figure 6: Covariates vs. assignment

Notes. The y-axes in the top two panels show the number of years of schooling for the students' mother and father, respectively. The y-axis in the bottom panel shows the share that is female. The hollow circles show the the characteristic means for bins of 0.01 average grade-points plotted at midpoints, while the x's show the characteristic means at multiples of 1/12. Linear fits of the characteristics on the first-term GPA is shown in each panel. The vertical lines indicate the cutoff at the 11th percentile.

4.4 Difference-in-Differences estimation

An alternative evaluation strategy is a difference-in-differences (DiD) estimation at the school level, exploiting the fact that the program was implemented over three years. Ideally one would like the introduction of the program to be random. As this was not the case, in order to draw causal conclusions we have to assume that the trends in average school outcomes would have been the same in the absence of program implementation in schools included and schools to be included. This is found to be a fair assumption, at least for a sub-sample of schools (Eielsen et al., 2013). By comparing how the students' outcomes evolve in the schools where the program was offered earlier to other schools we can estimate an intention to treat effect at the school level. Eielsen et al. (2013) present fairly precise estimates showing no evidence that the intervention has an effect at the school level. Even in the presence of effects for program participants we could fail to detect an effect at the school level. The potential

of a sub-group analysis is however limited for the first cohorts due to the lack of rule-based assignment. The problems this creates for a sub-group analysis is further discussed in the next sub-section.

4.5 Potential spillover effects

One concern is that treatment may affect the students who do not receive intensive training (the students scoring above the cutoff). Such spillovers may arise if schools reallocate teachers, essentially shifting resources from the remaining students to the participants. We saw in section 3.1, however, that participating schools to some degree were compensated for the increase in teacher demand from the program. Thus the consequence for at least a substantial part of the students may have been a reduction in class size, and a reduction of skill heterogeneity in the class, found by Duflo et al. (2011) to be important.[20] Even if the teacher hours stayed the same, the average quality of the teachers teaching the remaining students may have suffered, if for instance more motivated or able teachers were used in the program.

To explore the relative size of direct program effects and the spillover effects we might, with the availability of data on more cohorts in the program, use a sub-group difference-in-difference evaluation. As mentioned above, this is not possible for the first cohorts as there is only a limited number of schools with strict implementation. Still, let us assume for a moment that this was *not* the case and only students in the first decile participated in the first cohort. We could then compare that quantile in the first wave schools with the same quantile in the remaining schools to get an ITT estimate of the direct effects. Similarly, comparing the upper 90 percent of the distribution in the first wave schools with the same part of the distribution in other schools would give an estimate of the spillover effects.

In reality participation is not limited to the first decile. Figure 1 in Section 3 shows that there are participating students in the all of the lowest four deciles and thus if we found effects on the upper 90 percent of the distribution this could be both direct effects of the program and spillovers effects. So I cannot separate these effects with the available data, but it is nevertheless important for the overall evaluation of the program; for instance I could fail to find a positive local average treatment effect on the compliers if there is also a positive spillover effect on the non-participants. In the presence of these potential spillover effects I can still estimate local treatment effects: the effect of the program on marginal individuals' outcomes, relative to not being assigned to the program, but still being in a program school.

[20] Admittedly in a very different context: large primary school classes in rural Kenya were randomly divided in half by previous achievement.

5 Results

Before discussing the result from the TSLS estimation outlined in Section 4 above, I start with non-parametrically estimating the intention-to-treat effects. This is done by estimating local linear regressions at both sides of the cutoffs, using the preferred bandwidth of one average grade-point.

Figure 7 shows the results, and gives a visual preview of the effects of being offered the program on the achievement and progress outcomes of interest. There is no indication that the program affected GPA and exam scores, since they vary more or less continuously around the cutoff. For the outcomes measuring progression there is some indication of negative effects of the program.

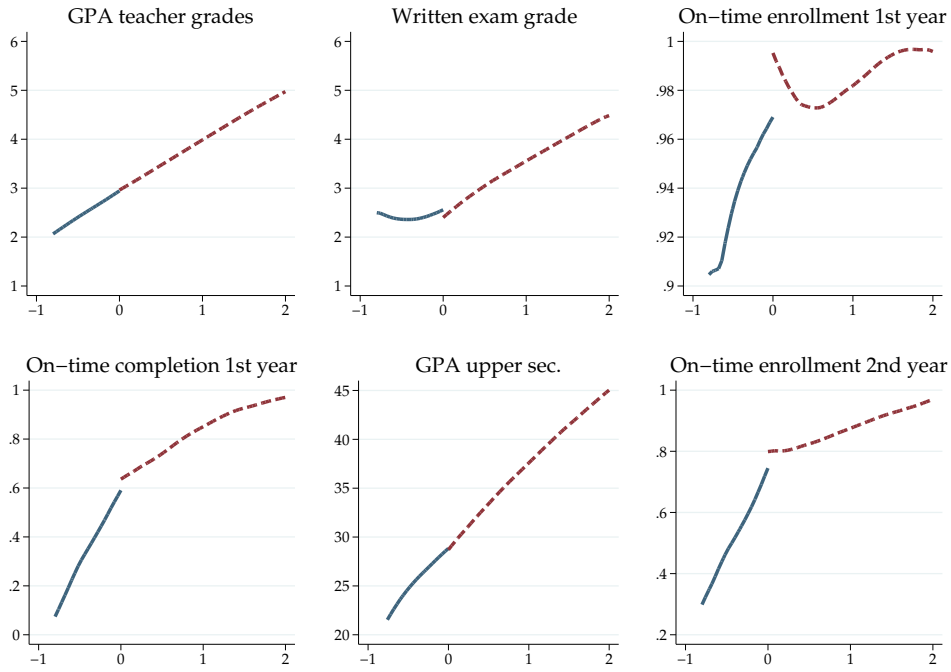

Figure 7: Average outcomes around estimated cutoff

Notes. The fits are the smoothed values from local linear regressions of the first term GPA on the outcomes, estimated separately at each side of the the cutoff, weighted with a triangle kernel with a bandwidth of 1 average grade-point for all outcomes. The cutoff, normalized to zero, was identified by the search algorithm at the 11th percentile in the estimation sample.

Column (2) in Table 5 shows the results from reduced form regressions, where treatment and outcomes are regressed on treatment assignment conditional on the first term GPA, as in Equation (8) in section 4.2. The first row shows that there is a strong relationship between

treatment assignment and actual treatment. At the threshold the probability of being treated is 74.4 percentage points higher than just above. This effect is highly significant, and shows that the necessary requirement for the design in Equation (4) above is satisfied.

The following rows show the reduced form results for the different outcomes. Students at the cutoff are estimated to be on average 2.6 percentage points less likely to enroll on-time the first year of upper secondary than those right above, but this difference is not precisely estimated and insignificant at conventional levels. Students at the cutoff is also estimated to have a 1.7 percentage points *lower* teacher grade, a 0.159 *higher* grade-point average on written exam and be 4.6 percentage points *less* likely to complete the first year of upper secondary school. None of these estimates are however close to be significantly different from zero.

Table 5: The local effects of participating in the program on the outcomes of interest (LATE)

	(1) Obs. in bwidth count	(2) ITT b/se	(3) LATE b/se
Share of compliers (First stage)	608	0.744***	1.000***
		(0.057)	(0.000)
GPA teacher grades	607	-0.017	-0.023
		(0.049)	(0.066)
Written exam grade	581	0.159	0.210
		(0.148)	(0.194)
On-time enrollment 1st year	608	-0.026	-0.035
		(0.028)	(0.037)
On-time completion 1st year	608	-0.046	-0.061
		(0.085)	(0.114)
GPA upper sec.	545	0.135	0.181
		(1.313)	(1.755)
On-time enrollment 2nd year	608	-0.054	-0.073
		(0.073)	(0.098)

Notes. Heteroskedasticity robust standard errors errors clustered at the discrete values of the assignment variable in parentheses. Stars indicate the significance level (* $p<0.10$, ** $p<0.05$). Data are for the students in the main estimation sample. Column (1) shows the number of observations for each of the outcome variables in the estimations with the preferred bandwidth of one grade-point. Column (2) shows the ITT estimates for the preferred bandwidth. Column (3) shows the LATE estimates, which for the outcomes with no missing values could be calculated by dividing the ITT with the difference in probability of treatment (.74). This probability is slightly different for example for the *Written exam grade*, as the first stage estimation also only include the 581 students for which we observe written exam grades. The cutoff, normalized to zero, was identified by the search algorithm at the 11th percentile for the estimation sample. For means, standard deviations and explanations of variables see the summary statistics in Table A.1 with notes.

Column (3) shows the LATE estimates from the TSLS estimation of the structural equation in (10) above, instrumenting for participation with target group membership (being below the cutoff). These are obtained by dividing the ITT estimates in Column (2) by the difference in participation in the first row in Column (2)). This yields effect estimates for the so-called compliers; students who participate in the the program if their first grade test score is below the cutoff but who would not have participated otherwise.

We see that the compliers are 3.5 percentage points less likely to enroll on-time the first year. This estimate is also far from statistically significant. Moving down to the final row shows that he compliers at the cutoff are estimated to be 7.3 percentage points less likely to enroll on time the second year because of the program, but again, the estimate is insignificant.

Table A.3 in the appendix shows the sensitivity of the ITT estimates for five different bandwidths. Even though participating students are consistently found to have slower progress than comparable non-participating students, at the current level of precision I cannot reject the null of no effects. Only for enrollment the first year, for a bandwidth of half a grade-point, is there a significant effect at the ten percent level, but with the number of tests this could very likely be spurious. This interpretation is supported by Figure A.7 in the appendix that shows that this estimate is highly sensitive to the bandwidth, with the estimate for a half a grade-point bandwidth being particularly negative.

While I do not find evidence of any effect on any of the outcomes studied, I cannot rule out substantial effects. For example, in Table 5, the standard error on the LATE estimate for completion of the first year of upper secondary is over 11 percentage points which is about one quarter of a standard deviation in the sample. Similarly the standard error on the written exam score is .19 grade points, or about one sixth of a standard deviation for this variable. Thus, any effect would need to be very large in order for me to be able to reject the null with a sample of this size.

6 Discussion

There are two categories of explanations for why I cannot find any effects of the program. First, the implementation in practice is not suitable for evaluation. Second, the program may be ineffective.

6.1 Program implementation and evaluation

The main difficulty in evaluating the policy with the currently available data is that there are a limited number of schools and municipalities that follow a strict assignment rule. This

reduces the size of the sample that can be used, and therefore the precision of the estimates. This is amplied by the fact that only about ten percent of the students in a school are directly affected by the program. A school level analysis will therefore also have limited statistical power, as described in Eielsen et al. (2013).[21]

While currently the analysis relies only on Stavanger, the standard way to increase precision is to extend the sample to other municipalities or schools that follow a strict assignment. This was investigated for the current cohort, and did result in more precise estimates but at the cost of comparability of the treated and untreated students. This could be because of measurement error in the estimated cutoffs as iterating over a large number of units and specifications will result in spurious cutoffs. Spurious cutoffs will lead to noise and may cause bias in the final effect estimates.

However, with data on more cohorts the applicability of the cutoff estimation procedure may increase.[22] With outcome data on additional cohorts there would be two main improvements. First, sample sizes would be larger. Second, cutoffs could be more reliably estimated. More specifically, if cutoffs are persistent over time and spurious cutoffs are random events then we may be able to identify (true) strict schools by selecting only those identified as strict over two or three years. The resulting sample of schools should contain fewer "false" strict schools. The likelihood of this circumstance and resulting potential bias can be explored with simulation studies.

Identifying a sample of strict schools could potentially also make possible proper subgroup analysis with a difference-in-differences framework, and it might be possible to isolate direct effects from potential spillover effects.

Another reason for the lack of precise results is the likely differences in the way the program was implemented between schools. Sletten et al. (2011) do not give information on the municipality of Stavanger separately, but report substantial variance in the group size in which the trainings took place for the whole sample. With treatment heterogeneity there could be both effective and ineffective versions of the program canceling each other out.

6.2 Ineffective program?

In this evaluation I have not studied the outcome explicitly targeted by the program, graduation rates, but rather related outcomes associated with completing upper secondary school. Participating students have not yet completed upper secondary, and grades are not a perfect

[21] The program only directly affects ten percent of the students in the average school and spillover effects on the other students are likely limited.

[22] For this thesis I only have available data on the assignment for two cohorts, and for the outcomes studied, only for the first.

measure of basic skills. Cortes et al. (2013) study an intensive training program and find that there is an effect on graduation despite a lack of immediate effects on performance. I cannot rule out this possibility here.

In light of the existing empirical literature it still seems likely that the program has at best small effects. There is both theory and evidence that suggests that early interventions are more effective than later remediation (Cunha et al., 2006; Carneiro and Heckman, 2003). Intervening at the end of compulsory school may be too late to make a large impact. Cook et al. (2014) find sizable effects from a program at the high school level, however, and warn that this conclusion might be premature. They argue that the focus of previous remediation programs have been wrong and failed to recognize the actual needs of the students that have fallen behind.

Even if it is not too late to target students at age 16, the focus of the intensive program studied here might have been too narrowly targeted at basic skills. In the review by Carneiro and Heckman (2003), non-cognitive skills such as motivation are found to be more easily malleable at later stages in the life cycle. Moreover, the apparently successful program studied in Cook et al. (2014) combined non-academic support and individualized academic remediation and improved expected graduation rates by 14 percentage points for the sample of disadvantaged high school students in Chicago. This seems like an interesting model, but it is important to stress that it still remains to been seen if the short-term effects hold up and can be replicated for different samples. Finally regarding the focus of the program, the remedial education program in Lavy and Schlosser (2005) also targeted improving the self-image of the students as one of its aims and it achieved 6 percent increase in the graduation rates.

A final reason the program may be ineffective is the limited size of the intervention compared to effective comparable programs. For example, the Ministry does not provide additional resources, as opposed to De Haan (2012), nor does the intervention (for the majority) increase the amount of instruction time as in Cortes et al. (2013) where it was doubled. Rather the intervention involves only changes to the group size and composition, and the pedagogy.

Still, the limited size of the program makes it a relatively cheap intervention in terms of costs per treated student. With large returns (to the individual and society) from completing upper secondary, even small effects can be economically relevant.

7 Conclusion

I have shown how a search over possible definitions and values of the first term GPA has successfully recovered the assignment threshold in the first term GPA distribution for a sample of schools. For this sample there is a large difference in probability of participating in the program "just" below and above the cutoff, while the students are otherwise similar. Comparing the two groups close to the cutoff I find no effect estimates significantly different from zero. The results are very imprecise, and thus I cannot reject economically interesting effects. The literature on comparable interventions and the larger literature on skill formation over the life cycle, however, suggest that the program very well might be ineffective. Future studies that investigate additional cohorts and more years of schooling of existing cohorts will be able to extend the outcomes investigated and should be able to better identify any possible effects.

8 References

Almond, D., Doyle, J. J., Kowalski, A. E., and Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, 125(2):591–634.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). Remedying education: Evidence from two randomized experiments in india. *The Quarterly Journal of Economics*, 122(3):1235–1264.

Barreca, A. I., Guldi, M., Lindo, J. M., and Waddell, G. R. (2011). Saving babies? revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics*, 126(4):2117–2123.

Barreca, A. I., Lindo, J. M., and Waddell, G. R. (2012). Heaping-induced bias in regression-discontinuity designs.

Card, D., Mas, A., and Rothstein, J. (2008). Tipping and the dynamics of segregation. *The Quarterly Journal of Economics*, 123(1):177–218.

Carneiro, P. and Heckman, J. (2003). Human capital policy.

Cook, P. J., Dodge, K., Farkas, G., Fryer Jr, R. G., Guryan, J., Ludwig, J., Mayer, S., Pollack, H., and Steinberg, L. (2014). The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in chicago. Technical report, National Bureau of Economic Research, Inc.

Cortes, K., Goodman, J., and Nomi, T. (2013). Intensive math instruction and educational attainment: Long-run impacts of double-dose algebra.

Cunha, F., Heckman, J. J., Lochner, L., and Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, 1:697–812.

De Haan, M. (2012). The Effect of Additional Funds for Low-Ability Pupils - A Nonparametric Bounds Analysis. Technical report.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, 101(5):1739–74.

Eielsen, G., Kirkebøen, L. J., Leuven, E., Rønning, M., and Raaum, O. (2013). Effektevaluering av intensivoppæringen i overgangsprosjektet, Ny GIV.

Falch, T., Nyhus, O. H., and Strøm, B. (2011). Grunnskolekarakterer og fullføring av videregående opplæring.

Falch, T., Nyhus, O. H., and Strom, B. (2013). Causal effects of mathematics. Working Paper Series 15013, Department of Economics, Norwegian University of Science and Technology.

Fredriksson, P., Öckert, B., and Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly Journal of Economics*, 128(1):249–285.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational evaluation and policy analysis*, 19(2):141–164.

Heckman, J. J. and Mosso, S. (2014). The Economics of Human Development and Social Mobility. NBER Working Papers 19925, National Bureau of Economic Research, Inc.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Imbens, G. (2012). RD: Stata module for optimal bandwidth choice.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal*, 113(485):F34–F63.

Lavy, V. and Schlosser, A. (2005). Targeted remedial education for underperforming teenagers: Costs and benefits. *Journal of Labor Economics*, 23(4):839–874.

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. S. and Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674.

Leuven, E., Oosterbeek, H., and Rønning, M. (2008). Quasi-experimental estimates of the effect of class size on achievement in norway*. *The Scandinavian Journal of Economics*, 110(4):663–693.

Leuven, E. and Rønning, M. (2011). Classroom grade composition and pupil achievement.

Lyche, C. (2010). Taking on the completion challenge: A literature review on policies to prevent dropout and early school leaving, oecd education working papers.

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

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

OECD (2013). Indicator a2 how many students are expected to complete upper secondary education?, in education at a glance 2013: OECD indicators.

Oreopoulos, P. (2007). Do dropouts drop out too soon? wealth, health and happiness from compulsory schooling. *Journal of public Economics*, 91(11):2213–2229.

Oreopoulos, P. and Salvanes, K. G. (2011). Priceless: The nonpecuniary benefits of schooling. *The Journal of Economic Perspectives*, 25(1):159–184.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sletten, M. A., Bakken, A., and Haakestad, H. (2011). Ny start med Ny GIV? kartlegging av intensivopplæringen i regi av Ny GIV-prosjektet skoleåret 2010/11.

Utdanningsdirektoratet (2013). Gjennomføringsbarometeret 2013:2.

Van Ewijk, R. and Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review*, 5(2):134–150.

9 Appendix

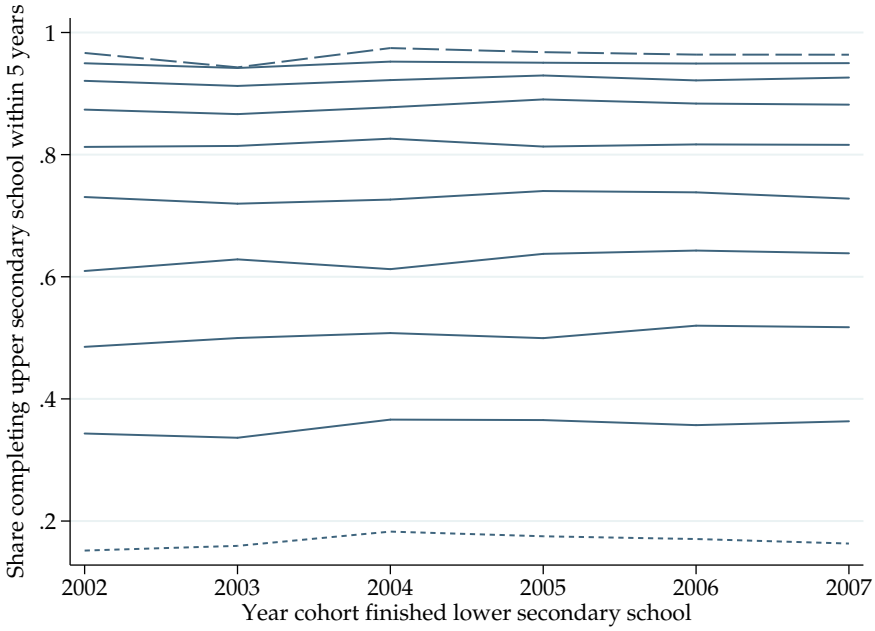

Figure A.1: Completion upper secondary school within 5 years in Norway, by achievement deciles of GPA

Notes. The completion rates of upper secondary school within 5 years of finishing lower secondary school are plotted by deciles in the achievement distribution at the end of lower secondary school (as measured by grade point average from the final assessment grades). The shortdashed line at the bottom is the average completion rate for the ten percent lowest-performing (the first decile), which for the 2007-cohort was 16 percent. The longdashed line is the rate for the top ten percent and was at 96 percent for the same cohort. The remaining lines, from the bottom up, shows the completion rate for the 2nd through the 9th decile. The sample consists of all Norwegian students completing lower upper secondary school in the period 2002-2007.

Table A.1: Summary statistics estimation sample

	Mean	SD	N
Characteristics			
Share female	0.483	0.500	1347
Mother's schooling	13.440	3.771	1301
Father's schooling	13.316	4.131	1276
Share immigrant	0.041	0.198	1347
Share immigrant parents	0.047	0.211	1347
Prior achievement			
GPA 1st term	3.829	0.783	1334
Math grade 1st term	3.407	1.147	1309
Norwegian grade 1st term	3.673	0.918	1309
Avg. on 8th grade tests	0.085	0.883	1310
Achievement			
GPA teacher grades	4.048	0.832	1343
Written exam grade	3.666	1.145	1305
On-time enrollment 1st year	0.982	0.132	1345
On-time completion 1st year	0.819	0.386	1345
GPA upper sec.	38.554	9.157	1265
On-time enrollment 2nd year	0.850	0.357	1345

Notes. *GPA 1st term* is the average of all grades (for most students this is 12 grades) set by the students' teachers at the end of the first term of 10th grade. *Math* and *Norwegian* grades make up two of the grades in *GPA 1st term*. *Avg. on 8th grade tests* is the average of three standardized grades from a national exam in 8th grade in English, Norwegian and Maths. *Mother's* and *Father's schooling* is the number of years of schooling of the mother and father of the student, respectively. *Share female/immigrant/immigrant parents* are all dummy variables equal to one if the student is female, a immigrant or have immigrant parents, respectively. The enrollment and completion variables equals one if the student has enrolled our completed, respectively; zero otherwise. *GPA teacher grades* is the average of all grades (for most 13 grades) set by the students' teachers at the end of lower secondary school (10th grade). *Written exam grade* is the average of the three exams most students undertake in English, Maths and Norwegian. *GPA upper sec.* is the average of all grades the first year of upper secondary school multiplied by ten.

Table A.2: Comparison of participants and other students in estimation sample

	(1) Participants mean/sd	(2) Non-participants mean/sd	(3) Difference b/se
GPA 1st term	2.451	3.921	-1.470**
	(0.343)	(0.786)	(0.039)
Missing grades 1st term	0.008	0.010	-0.001
	(0.092)	(0.098)	(0.009)
Math grade 1st term	1.912	3.548	-1.637**
	(0.576)	(1.085)	(0.063)
Norwegian grade 1st term	2.411	3.791	-1.380**
	(0.578)	(0.853)	(0.060)
Avg. on 8th grade tests	-1.009	0.186	-1.195**
	(0.616)	(0.834)	(0.063)
Share female	0.398	0.491	-0.092*
	(0.492)	(0.500)	(0.047)
Mother's schooling	10.514	13.702	-3.188**
	(4.477)	(3.589)	(0.445)
Father's schooling	11.277	13.491	-2.214**
	(3.225)	(4.154)	(0.343)
Share immigrant	0.161	0.029	0.132**
	(0.369)	(0.169)	(0.034)
Share immigrant parents	0.093	0.042	0.051*
	(0.292)	(0.201)	(0.027)
Observations	118	1229	1347

Notes. Data are for the main estimation sample. Mean values of each characteristic is shown in column (1) and (2) for participants and non-participants, respectively; standard deviations are in parentheses. Column (3) tests each difference with a Welch's t-test, allowing for the difference in sample size and variance; standard errors are in parentheses; stars indicate the significance level (* $p<0.10$, ** $p<0.05$).

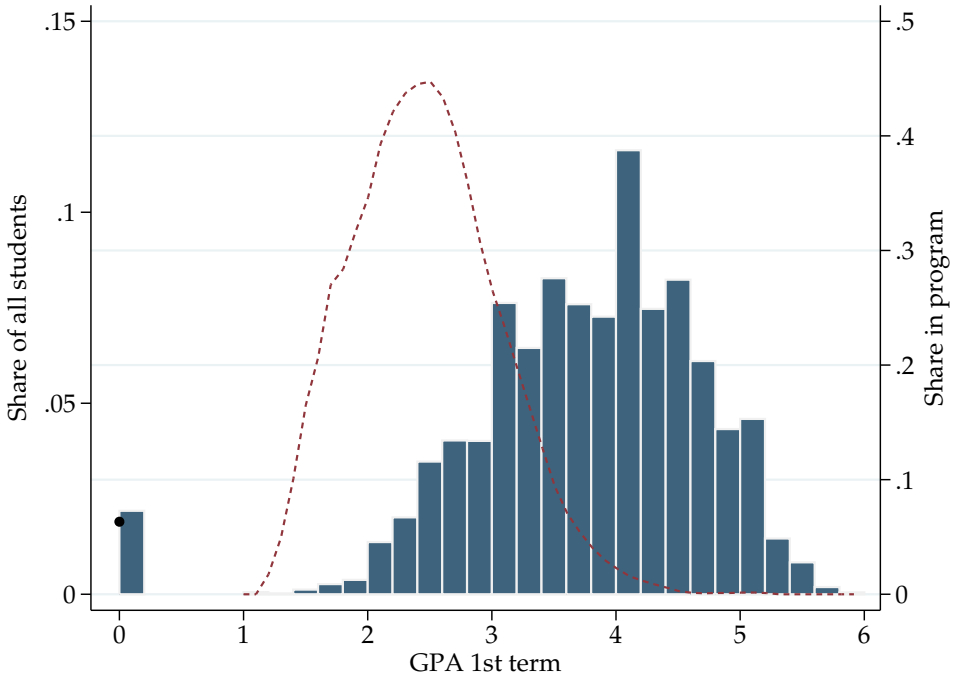

Figure A.2: Pupils in wave 1 schools

Notes. The histogram shows the first term grade distribution of all students, while the graphed Epanechnikov kernel density estimation shows the same distribution for only students participating.

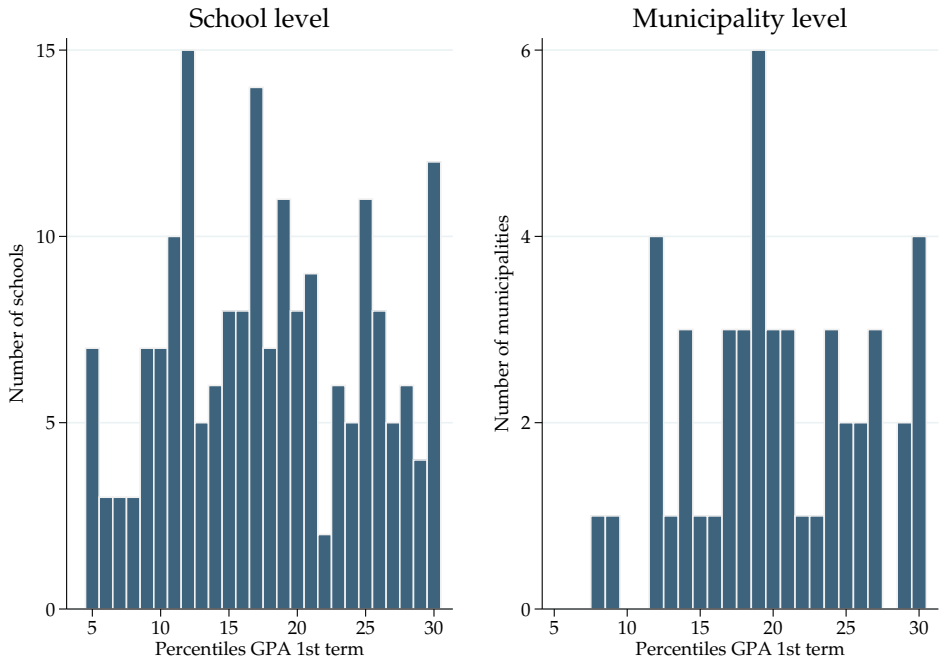

Figure A.3: Percentiles identified as cutoffs

Notes. Histograms shows the number of times the percentiles from the 5th to the 30th were identified as cutoffs for schools and municipalities. The percentiles are from the specification that best explain program participation.

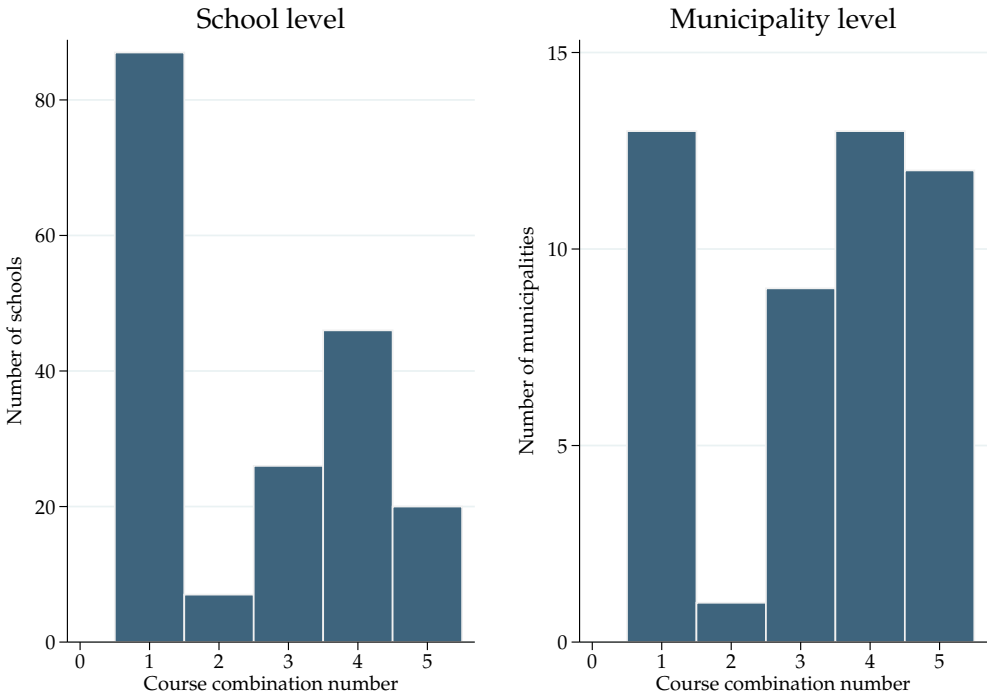

Figure A.4: Course combinations

Notes. The histograms shows the number of times different grade combinations were found to be the best assignment variable, as identified by the search procedure.

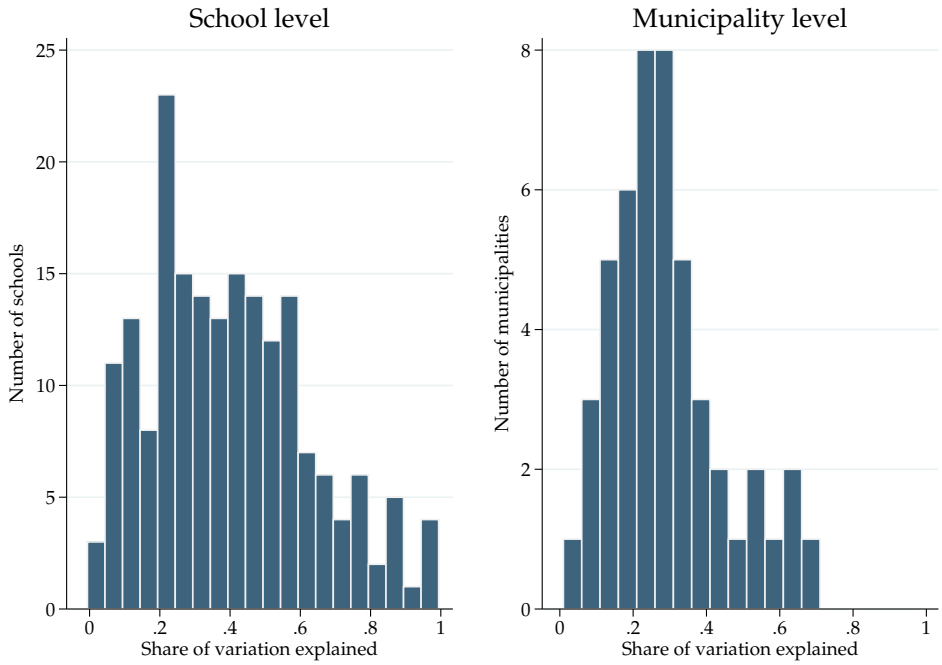

Figure A.5: Degree of strict assignment

Notes. Histograms shows the frequencies of the R-squared from the best specifications found with search procedure.

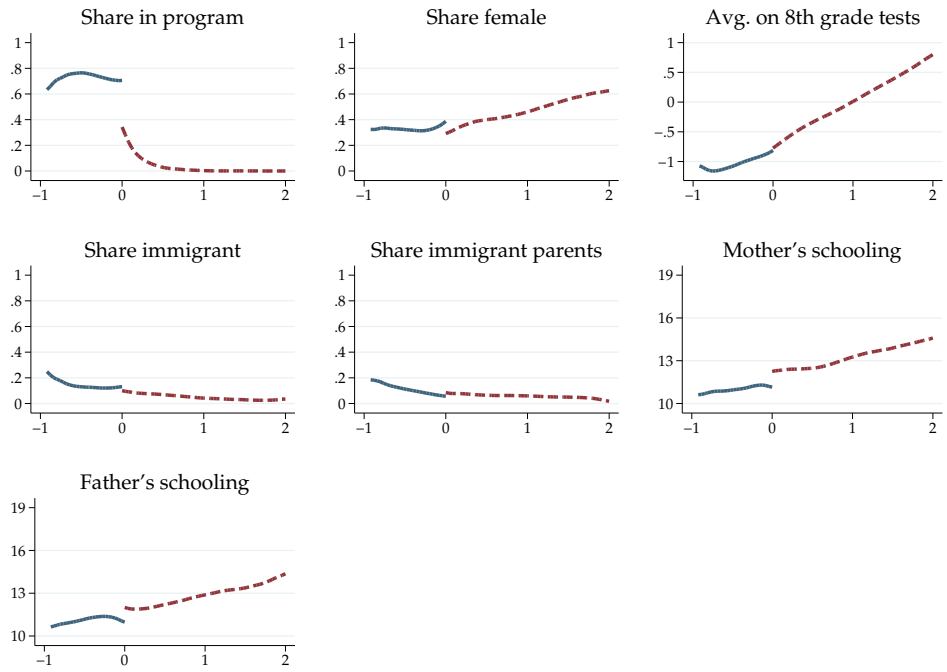

Figure A.6: Composition of student characteristics around cutoff for alt. sample

Notes. The fits are the smoothed values from local linear regressions of the first term GPA on participation, characteristics, and the prior achievement. All regressions estimated separately at each side of the the cutoff, weighted with a triangle kernel with a bandwidth of 0.5 average grade-point for all outcomes. The assignment variable first term GPA is normalized to zero.

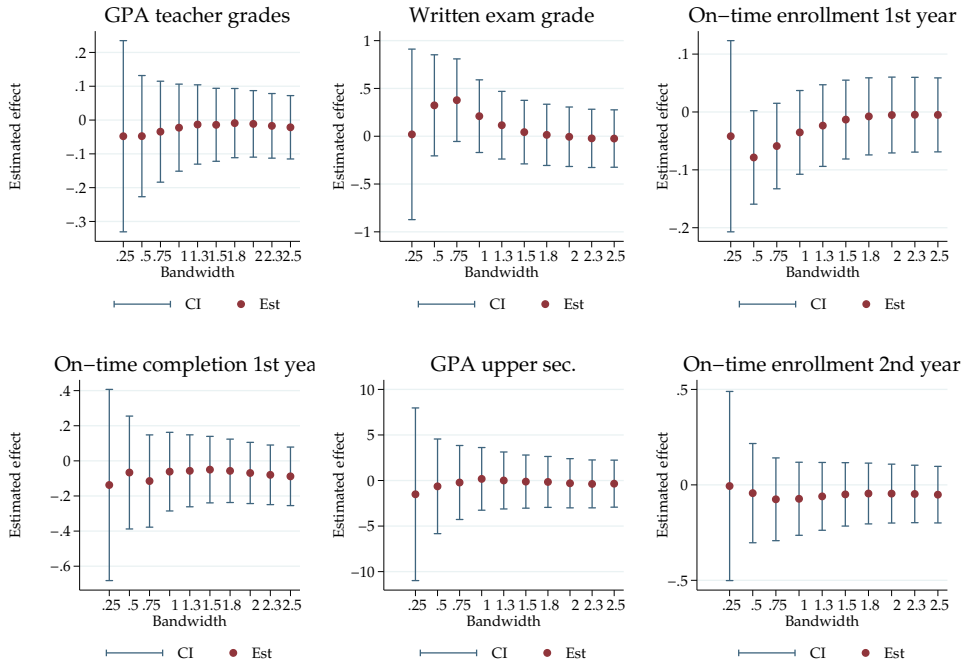

Figure A.7: Dependence on choice of bandwidth

Notes. Graphs of estimates (with 95% confidence intervals) versus bandwidths for all six outcome variables. Data are for the main estimation sample.

Table A.3: The local effects of being offered the program on the outcomes of interest (ITT)

	(1)	(2)	(3)	(4)	(5)
	.25	.50	.75	1.00	1.50
GPA teacher grades	-0.036	-0.035	-0.026	-0.017	-0.011
	(0.110)	(0.068)	(0.057)	(0.049)	(0.042)
Written exam grade	0.015	0.235	0.279*	0.159	0.034
	(0.351)	(0.200)	(0.166)	(0.149)	(0.133)
On-time enrollment 1st year	-0.034	-0.058*	-0.043	-0.026	-0.010
	(0.035)	(0.031)	(0.028)	(0.028)	(0.026)
On-time completion 1st year	-0.104	-0.049	-0.085	-0.046	-0.038
	(0.210)	(0.121)	(0.098)	(0.085)	(0.073)
GPA upper sec.	-1.150	-0.462	-0.160	0.135	-0.089
	(3.719)	(1.927)	(1.533)	(1.317)	(1.130)
On-time enrollment 2nd year	-0.005	-0.032	-0.056	-0.054	-0.038
	(0.192)	(0.098)	(0.082)	(0.073)	(0.064)
Observations	171	311	456	608	919
Wald test of joint significance	1.398	5.556	6.630	3.011	0.890
p-value Wald test	0.966	0.475	0.356	0.807	0.989

Notes. Heteroskedasticity robust standard errors clustered at the discrete values of the assignment variable in parentheses. Stars indicate the significance level (* $p<0.10$, ** $p<0.05$). Data are for the students in the main estimation sample. Column (1) - (5) presents ITT estimates for bandwidths of .25 - 1.5 average grade-points. The cutoff, normalized to zero, was identified by the search algorithm at the 11th percentile for the estimation sample. For means, standard deviations and explanations of variables see the summary statistics in Table A.1 with notes.