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

Gaute Eielsen† Lars J. Kirkebøen‡

April 2014

Preliminary and incomplete - do not cite

Abstract

This paper evaluates the short-term effects of an education program providing intensive learning support to low-performing students at the end of compulsory school. The explicit target group is the bottom ten per cent students. However, this has been interpreted differently, creating institution-specific performance thresholds for assignment. We develop an approach to identify these thresholds, and estimate the effects of the program using a regression discontinuity design. We don't find any significant effects with this approach and this conclusion is supported by a difference-in-differences analysis comparing schools starting the programs at different times. However, the approach for identifying local thresholds is likely to be relevant also in other situations.

1 Introduction

Low upper secondary completion rates continue to cause concern among policymakers in most OECD countries (OECD, 2013). Failure to complete secondary education comes at a great

*This paper is part of the ongoing evaluation of the "Ny GIV" intiative financed by the Norwegian Ministry of Education and Research. The paper extends the analyses documented in Eielsen et al. (2013), building heavily on these, as well as further analyses documented in Eielsens's thesis for the M.Phil of Economics degree at the University of Oslo. In particular we want to thank our colleagues and fellow investigators Edwin Leuven, Marte Rønning and Oddbjørn Raaum. Edwin Leuven also has supervised Gaute Eielsen's thesis. Furthermore, we want to thank the Ministry for comments on previous work, and Anders Bakken and Mira A. Sletten at NOVA and Solveig Holen and Berit Lødding at NIFU for providing data from their mappings of the program. Any errors remain our own.

†Research Department, Statistics Norway. E-mail: gae@ssb.no

‡Research Department, Statistics Norway. E-mail: kir@ssb.no

cost to both the individual and the society at large (Oreopoulos, 2007). For the individual, not only does 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). Academic skills, imperfectly measured by grades, are closely associated with completing another year of schooling, and ultimately a secondary education, in all countries reviewed in Falch et al. (2011), including Norway.

In Norway the share completing a secondary education within 5 years has been relatively stable over the last decade at around 69 percent of the cohorts[1]. In 2010 the Ministry of Education and Science set a target to increase this indicator to 75 percent within 2015. At the same time, several policies were initated under the name "Ny GIV" to achieve this goal (Utdanningsdirektoratet, 2013). A main 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. The program is aimed at increasing basic skills in reading, writing and numeracy, and is generally implemented as specially adapted instruction and training in smaller groups. This is a substitute to ordinary classes, extra instruction time is not added.

This paper analyzes the implementation and the effects of the remedial program on relatively short-term outcomes: Academic performance leaving compulsory schooling and progress through the early parts of upper secondary. Doing so, we make two contributions to the literature. First, we develop an approach to find unkown 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 differently interpreted in different schools and municipalities, resulting in some schools having no clear cutoff, while other have cutoffs at unknown values of first term GPA, which in turn can be defined in different ways. Our search procedure builds on the same idea as when looking for structural breaks in time-series econometrics, and is used by (Card et al., 2008). However, to our knowledge, it has not previously been used in the context of a policy evaluation. We believe it might prove useful in contexts where there exist rules, but due to room for different interpretations from different administrative units, the actual rule applied varies across units. If we can find a method to convincingly find the rule applied, then there may still be possible to say something in these contexts, perhaps previously regarded as too "messy".

Our second contributions is to estimate, using RD estimation and the thresholds for participation identified, the causal effect of the remedial program on the outcomes of interest. Doing so for schools where we can identify a clear cutoff, we argue that our estimate does

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

indeed give a credible estimate of the program for marginal participants. We supplement this analysis with a difference-in-differences analysis, using the gradual roll-out of the program and comparing different schools.

The RD analysis compares students just below a certain cutoff value in the first term grade point average (first term GPA) distribution, who were likely to receive the intervention, with those just above with a much lower possibility of receiving the treatment. The idea behind the evaluation is that the students just above this cutoff value are similar in both observed and unobserved characteristics to those just below, and is therefore a valid control group. As participation in the program is voluntary, actually receiving the treatment is not deterministically a function of first term GPA, and thus the data generating process is what is known in the literature as a "fuzzy" RD design. We depend on two crucial elements; the first is what gives rise to the design: that actual implementation in the schools caused discontinuities 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 is continuous in GPA at the discontinuity, or in other words, that there are no other factors that change discontinuously at the cutoff other than the difference in treatment probability. The key 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 to a greater extent be verified. Much in the same as with a randomized controlled trial where (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 (Lee and Lemieux, 2010). If the identifying assumption holds and the assumption of monotonicity and the exclusion restriction also hold (discussed below in section 4), we can use the first term GPA as a valid instrument for participation and identify the local average treatment effect for the students accepting participation, the compliers, in the proximity of the cutoff (Hahn et al., 2001).

The major challenge in this evaluation is to find schools where the first element discussed above, a discontinuity in treatment probability, is satisfied. In most schools the cutoff was not implemented strictly and there was different selection practices across schools. To find the specific selection rule employed by the municipalities and schools, respectively, we use an algorithm to search for discontinuities in the probability of participation. This leads us to a sample of schools in the municipality of Stavanger, for which the identifying assumption seem to hold.

This study focuses on the short-term effects on outcomes most closely linked to cognitive skills. We do not find any effects of the program on the outcomes we study, neither using

RD estimation nor DiD. However, because of the limited precision, we cannot reject that there are effects of economical interest on these outcomes. Furthermore, we are still unable to study the longer-term outcome that the program is meant to influence, i.e., completion of upper secondary school.

The paper proceeds a follows: In section 2 we briefly review some previous studies. Section 3 describes the institutional background, program studied, its participants and the data sources used in more detail. Section 4 develop the empirical strategy, including our search procedure and our effect estimators. Section 5 presents and discusses the results from the estimations, while section 6 concludes.

2 Previous literature

Motivating the study from an overall policy perspective, there is a large literature that finds both large social and private returns to another year of education, and several studies from the US shows that high school completion is a high-return investment.

There is a limited literature evaluating comparable remedial programs. Lavy and Schlosser (2005) study the effect of providing extra teaching to low-performing upper secondary students, finding that this increases graduation rates.

De Haan (2012) study 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, reading and math performance also improves.

Maybe most closely related to our paper, Cortes et al. (2013) study an algebra policy implemented in Chicago in 2003. Students with achievement below the national median result in 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 seems to have been 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 this in the Chicago public high schools, find surprisingly large effects. Math grades are reported to have improved by 0.67 of a control group standard deviation, and expected graduation rate with 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 addressing the different components of the program. The program implies a reduction in class size for both treated students and the remaining students in the cohort, features which have been studied intensively empirically independently (see e.g. Hanushek (1997) and Krueger (2003) for a summary of the international literature on the short-term effects), in a Norwegian context by Leuven et al. (2008), and also theoretically by Lazear (2001). Fredriksson et al. (2013) study the long-term effects of smaller class size the last three year of primary school 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 internvention also changes the classroom composition, which can have a causal effect (Leuven and Rønning, 2011; Van Ewijk and Sleegers, 2010). Additionally the ministry intended to change the pedagogy used, which may have an effect (Machin and McNally, 2008; Banerjee et al., 2007). Related to this the curriculum changed, which according to Cortes and Goodman (ming)also 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 years 1-10, with students leaving compulsory school the year they turn 16. Private schools are rare, about 98 percent of lower secondary school students, and almost as many upper secondary school students attend public schools. Public lower secondary schools are owned and financed by the municipalities, and all follow the same national curriculum.

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 group are vocational, generally consisting of two years in school followed by to years as an apprentice, giving 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. Completion in this context means obtaining a diploma from upper secondary school. Thus, some non-completers may have attended school or completed their apprenticeship, but without earning a diploma because of a previously failed compulsory subject.

3.1 The program

The program's Norwegian name "Overgangsprosjektet" in Norwegian, 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 within each municipality was 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.[2]

According to the Ministry, the lack of basic skills, in literacy, writing and numeracy, is a key reason of the low completion rates. Thus, to prepare the students for upper secondary, instead of following the regular curriculum in regular classes, they are taught such 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, not changing the allocation of the students' time across subjects.

The intensive learning support was rolled out in three waves starting in the spring of 2011, each with schools encompassing approximately one third of the students. The second and third wave 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 the 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. As this study will only use the sample of schools in the first wave of the program we rely on survey responses from the principals after the first year, reported in Sletten et al. (2011). The response rate for the principals was at 88 percent. Students and teachers (both teachers teaching intensive training lessons and other teachers) were also surveyed, but the response was only at approximately 30 and 40 percent of the populations, respectively, and thus we rely mainly on responses from the principals in the following.

In most schools the program was a substitute to regular classes and typically took about 6 to 7 hours of the 30 hour school week for the students. In a minority of schools the targeted students also received classes outside of the 30 hour school week. The average duration was approximately 13 weeks, with a minimum of ten weeks and maximum of 18 weeks. There was some variation across schools in whether the students received training in all three of the competencies; 80 percent of the participants received training in literacy and writing; 90

[2] Of the 2002-2007 compulsory school graduation cohort, only 15-17 percent of the lowest-performing decile had completed upper secondary school within five years of completing compulsory school. For the second and third deciles the corresponding figures are approximately 35 and 50 percent, while about 90 percent of the top half complete within five years.

percent in numeracy; such that 70 percent received training in all three competencies. In all but 5 percents 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 in groups depending on the competency taught.

The group size was typically 10 students, but with much variation across schools. From the probably not very representative survey of teachers we learn that, of these, 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 teachers surveyed 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 a the lowest-performing students, it was also a school-level intervention. The consequences for the remaining students was a reduction in class size, reduced within-class heterogeneity in terms of performance and possibly a change/reallcation 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. 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 to hire teachers in relation to the project. This must then have been supplied by the municipalities. We have no information of whether these funds covered the extra 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 programs in upper secondary school, that will affect the students that participated. Notably, there the responsibilities of school and other public agencies to follow up students in risk of dropping out were clarified. However, as these policies are not exclusive to the participants of the intensive training, the potential effect of the program will, given that the identification is valid at this stage, be causally linked to the training. The later interventions will, however, be important when discussing the external validity of the potential results, as these could be conditional on an environment where struggling students have extra resources available.

3.2 Data

We use administrative register data from Statistics Norway, covering the complete cohorts of lower secondary graduates of 2003 through 2011 for this analysis. This means that we will

be able to study the first wave of the program. The data will later be extended with more cohorts. Each cohort consists of roughly 60 000 students. On these students we have all final and mid-term grades from lower and upper secondary school in addition to a wide array of variables for the students's parents such as income and education. From the Norwegian National Educational data base we also have data on the students's 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)).

At the school level we have information on the schools' financial resources, teacher density and experience, and also information from a yearly student survey on a range of topics from perception of learning to motivation and general happiness.

3.3 The participants in the first wave

The target group of the program was the 10 percent lowest-scoring students in each municipality (Sletten et al., 2011). From Table 1, which compares participating students with other students in the participating schools, we clearly see that these differ. The participating students have lower first term performance, in particular in math, are more often boys and has a more adverse family background.

Table 1: Comparison of participants and non-participants

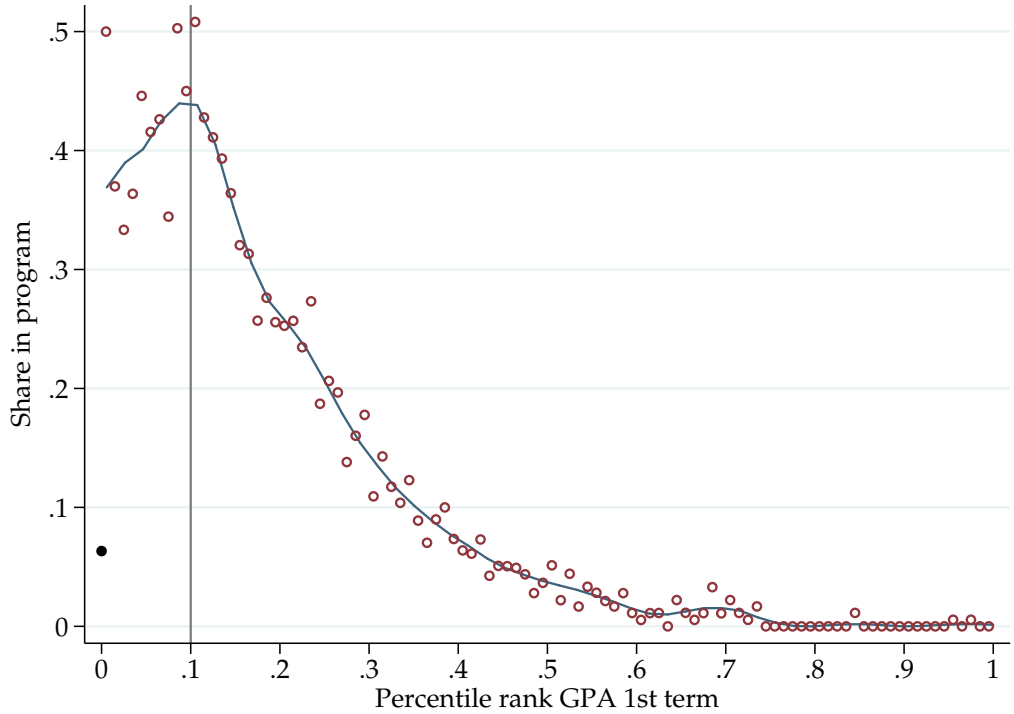
	(1) Participants mean/sd	(2) Non-participants mean/sd	(3) Difference b/se
GPA 1st term	2.863	3.840	-0.977**
	(0.592)	(0.957)	(0.015)
Missing grades 1st term	0.013	0.023	-0.010**
	(0.112)	(0.150)	(0.003)
Math grade 1st term	2.178	3.543	-1.365**
	(0.658)	(1.126)	(0.018)
Norwegian grade 1st term	2.767	3.792	-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 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 (* $p<0.10$, ** $p<0.05$).

However, while the participating students are low-performing, how "the lowest-performing ten percent" should be interpreted is less clear. We see this from Figure 1, which shows how the share of students in the program schools that participates in the program varies over the municipality-specific distribution of first term GPA.[3]

[3]Figure A.1 in the Appendix shows the 1st 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 participating conditional on the percentile rank point. OThe vertical line indicates the 10 percent lowest-scoring pupils in each municipality. Also added is a fit estimated with a local linear regression, with a bandwidth of 2 percentile rank points, weighted using the Epanechnikov kernel.

Less than half of the target group, the 10 percent lowest-scoring students in each municipality, actually participates in the first year of 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 seperately below and above the 10th percentile reveals no difference. There is no clear discontinuity either way.

There are several reasons why we, in spite of the clear instruction from the Ministry, do not see a clear discontinuity in participation around the 10th percentile. First, "performance" is not straightforward to measure. 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 Math and Norwegian grades could be given more important, as some coordinators of the programs report.

Second, some students already receive different kinds of special education. The Ministry explicitly states that in such cases participation in the program is only relevant when it is considered to be a better than the alternative. Given that these students already have an individually adapted curriculum and teaching, this is unlikely to be the case. About 11 percent of 10th grade students have such individual programs. While we are unable to identify these, it seems reasonable that such student are overrepresented among the low-performers.[4] Thus, this may explain a large share of the "missing" intensive training participants below the 10th percentile. On the other hand, we observe that about 10 percent of the students in the program schools participate. If some low-performing students are not considered for participation, this means that the schools will need to include higher-performing students. With different shares of individual program-students at different schools, this can give rise to different participation thresholds.

Finally, schools or municipalities may choose not to have a fixed threshold for participation. Ultimately, the schools or municipalities were in charge of recruiting students to the program. There is anecdotal evidence that the selection of students for participation was based on the effect the teachers expected a given student would have from participation.

To conclude, the selection of students was done in different was at different schools, or in different municipalities. Some schools or municipalities chose students in a way that produced no discontinuities in the probability of participation, or in such a way that participating and non-participating students with similar first term GPA are systematically different. Other schools and municipalities are likely to have assigned students according to a local cutoff, unknown to us. In the next section we detail how we identify such cutoffs and proceed to use them in RDD estimation of the program's effect.

4 Empirical strategy

The challenge in estimating the causal effect of the intensive training program is that participation is endogenous. In Table 1 we saw that participants are different in many observable characteristics. If we simply compare students who attend with those who don't the estimated effect of the program will likely be heavily downward biased. To avoid this the main identification strategy in this thesis relies on the directive from the Ministry of Education stating that the bottom 10 per cent of students should participate in the program. However,

[4] While we do not have data on individual programs, the number of subjects a student receives a grade in may be a proxy. Studying this, we find that there are students with reduced 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 shown in the previous section, the municipalities/schools had some room for maneuvering with the implementation of the program, but there are also some that follow the Ministry's recommendation. For the municipalities that construct average mid-term grades across all subjects, and assign some share of the lowest scoring to treatment, the difference between those just assigned to treatment and those just not assigned may indeed be small, and in the limit nonexistent, in all factors determining the outcomes of interest. If participation was mandatory a comparison of students just below with those just below would give an unbiased estimate of the effect of the program. Participation in the program was voluntary, however, but as most that received an offer participated there is a clear difference in probability for participation across the cutoff that we can use as an instrument. This identification strategy is known in the literature as a "fuzzy" regression discontinuity (FRD) design. In section 4.1 we discuss the identification in further detail.

Some municipalities and schools seem to use a rule-based assignment to treatment, although not necessarily the 10th pecentile. In section 4.2 we describe the search algorithm we apply to find these rules.

4.1 The effects of the intensive training program

The effect of some treatment on an outcome y for student i can conceptually be found by the difference in potential outcomes. Let $y_i(1)$ be the outcome of interest under treatment, and $y_i(0)$ the outcome in the absence of treatment. The causal effect for student i is then $y_i(1) - y_i(0)$. "The fundamental problem of causal analysis", coined by Holland (1986), is that we cannot observe one student in both states at the same time and thus we are left to estimate *average* causal effects, either on one student over time, or on some population of students. This study aims to reveal an average effect for the sub-population of students around the cutoff that is induced to participate by the instrument, here being below the cutoff, known in the literature as "compliers"(Angrist et al., 1996). First consider, however, all individuals close to the cutoff. Following(Hahn et al., 2001), we use the potential framework to illustrate the necessary conditions for identification in the fuzzy design. We start with a flexible model for the observed outcome that allows for heterogeneity in treatment effects

$$y_i = y_i(0) + d_i\beta_i, \tag{1}$$

where $\beta_i \equiv y_i(1) - y_i(0)$. Now let let $d_i = 1$ if student i participated. It can be considered a random variable given the individuals' GPA, g_i. The conditional probability of receiving treatment can then be defined as $E[d_i \mid g_i = g] = Pr[d_i = 1 \mid g_i = g]$. This conditional

probability has to be discontinuous at some cutoff c for the GPA, $g_0 = c$. This necessary discontinuity required for the evaluation design can be defined :

$$d^- \equiv lim_{\epsilon \uparrow 0} Pr[d_i \mid g_i = c + \epsilon] \neq lim_{\epsilon \downarrow 0} Pr[d_i \mid g_i = c + \epsilon] \equiv d^+ \tag{2}$$

In words, the probability of treatment conditional on the individual's GPA must be different when moving along g towards the cutoff from below (defined d^-) and above the cutoff (d^+), respectively.

Now the main identifying assumption, that the potential outcomes of individuals close to the cutoff are similar, can be formalized:

Assumption (A1): $E[y_i(0) \mid g_i = g]$ and $E[y_i(1) \mid g_i = g]$ is continuous at $g_0 = c$.

Lee (2008)provides an argument for situations where assumption A1 is realistic, even in the presence of maximizing individuals, possibly preferring one side of the cutoff to the other; Lee (2008) argues that as long as there is an element of chance determining the assignment variable, they cannot *exactly* self-select into their preferred side of the cutoff. Adapted to this evaluation, if there is a random element partly determining the students' GPA, call it e_i , then even with a systematic part determined by say the innate ability and effort, denoted a_i, it will still be random whether students with similar a_i fall on one side of the cutoff or the other. We thus imagine that GPA can be decomposed into a systematic and a stochastic element: $g_i = a_i + e_i$. In the words of Lee and Lemieux (2010) as long as individuals cannot "... *precisely* "sort" around the discontinuity threshold", the assumption of continuity of potential outcomes at the cutoff is realistic. The very nice feature of assumption A1 is that it predicts that students just below and above they cutoff should have the same baseline characteristics, and it thus leads to similar tests to the one conducted between control and treated students in a randomized experiment Lee (2008). Such tests will be presented below to discuss the plausibility of the assumption. Returning briefly now to this context, theoretically from the students' perspective it seems plausible that there is a stochastic element to the first term average grade, after all it depends on grading in several courses on multiple tests by different teachers. However, what we regard as a greater threat is whether teachers, and especially the class head teacher, might have this complete control to "sort" students below or above the cutoff, perhaps based on perceived gains from the program. In addition to potentially biasing the effect estimates, this could also be undetectable in the balance tests if these characteristics influencing the teacher's sorting are uncorrelated with observed characteristics.

Now we can first look at a local intention to treat (ITT) parameter, which is closely related to the global ITT parameter found in randomized controlled trials with partial compliance in randomized by looking at the difference at the cutoff c.

$$\beta^{ITT} = lim_{\epsilon \uparrow 0} E[y_i \mid g_i = c + \epsilon] - lim_{\epsilon \downarrow 0} E[y_i \mid g_i = c + \epsilon] \equiv y^- - y^+ \tag{3}$$

With heterogeneous effects of the program, and without further assumptions, we can only find an average effect for that group of students that are induced to participate by the instrument, the so-called "compliers". This thus excludes students that would get into the program regardless of their first term grades, as well as 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 don't. With maximizing students one would expect the compliers to perceive their gains from treatment to be higher. Angrist et al. (1996) shows that if we can also assume monotonicity, that there are no students who would participate if they were above the cutoff, but not when below, we can identify the local average treatment effect (LATE) for the compliers as the ratio of the local ITT and the difference in treatment probability:

$$\beta^{LATE} = \frac{y^- - y^+}{d^- - d^+} = E[y_i(1) - y_i(0) \mid \text{student } i \text{ is a complier}, g_i = c] \tag{4}$$

Treatment may also affect the untreated students who do not receive intensive training (the students scoring above the 10th percentile). Treatment spillovers to untreated students may arise for example if schools reallocate teachers, or if different organization of classes or increased motivation of the participating students affects the amount of disruptions in classes that the non-participating students experience. Thus, in (4) above, y^+ can also be (causally) affected by the program, relative to a counterfactual situation where the program is not introduced in the school. If this is the case we can still consistently estimate individual-level treatment effects; the effect on a marginal individual's outcomes of being assigned to treatment, relative to not being assigned to treatment but still being in a program school. We discuss this in more detail below, when discussing DiD estimation.

4.2 Searching for cutoffs and "strict" implementation

As we know which students participate in the program, as well as the grades for the whole distribution of students, we can implement an algorithm that finds the percentile that best explains actual assignment to the program. There are five different combinations of courses c that the local program administrators have reported that have been used to assign students. As grades differ across municipalities and this was level of selection mandated by the Ministry we calculate these distributions specific to each municipality m for each cohort t and calculate the 1st through the 35th percentile, denoted the nth percentile, for each of these distributions.

Now for all the municipalities where there are participating students we iterate over all the relevant combinations of percentiles and find the percentile that maximize the coefficient of determination, R^2, in estimating the following model with OLS:

$$d_i = \gamma_0 + \gamma_1 Target_{icmtn} + u_i, \tag{5}$$

where $Target_{icmtn} = 1[gpa_{icmtn} \leq Percentile_{cmtn}]$

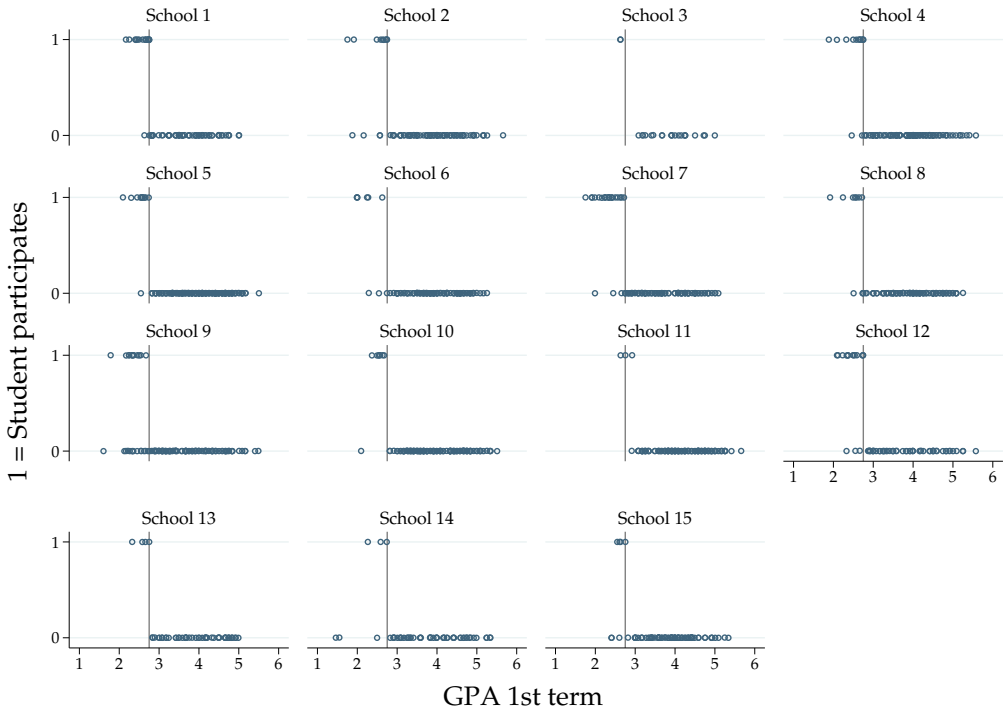
The $Percentile^*_{cmtn}$that is found to maximize the R^2 is then saved and forms the municipality and year specific cutoff. We also use the R^2 from the best regression to categorize the municipality in how strict they seem to be. If the municipality is perfectly strict, where all students in the nth percentile participate, the model would perfectly explain the variation and thus yield an R^2 of 1. We categorize municipalities with an coefficient larger than 0. 5 as "quite strict" in Figure A.4.

The same procedure of iterations is repeated at the school level with the same municipality-specific percentiles. The idea is that there could be certain "strict" schools, even though not all schools participating in the municipality is strict. We categorize the schools in the same way in Figure A.4.

We find that the combinations of subjects and percentiles that best explain program participation differ substantially across schools and municipalities. From Figure A.2 in the Appendix, we see that the percentiles varies from the 5th to the 30th in the first cohort. only looking at the munici although and especially for the municipalities most fall in the more narrow range from 10 to 25. How well we are able to explain assignment also varies, cf. Figure A.3, but is on overall rather low: Most schools have a share of explained variation (R^2) smaller than 0.6 and most municipalities smaller than 0.4. When we categorize units by the share of variation explained and show program participation by the best forcing variable in Figure A.4 we see clear differences in the extent to which participation changes discontinuously at the cutoff.

However, we find that there are few municipalities with reasonably strict assignment. This group is dominated by the municipality of Stavanger. This matches well with reports that assignment in Stavanger was particularly strict, as well as what we see when we plot the individual students for schools in Stavanger in Figure 2. Assignment seem to adhere to a strict cutoff. In all schools the same municipality-specific cutoff at the 11th percentile made from the average of all grades predicts participation very well, with the exception of school 9. To avoid introducing heterogeneity from different municipalities with different administrations when it does not provide a notable increase in the amount of data, we will use Stavanger as our main estimation sample excluding school 9.

Figure 2: Assignment in Stavanger, 1st cohort



Notes. Use cutoff from algorithm: 11th percentile, or GPA of 2.75

Concerning the results from searching for school-specific cutoff, we do find some schools that seem to have a reasonably strict assignment, as shown in Figure A.3. However, iterating over a large number of units and specification, we can expect to find some spurious cutoffs: Schools where assignments have not really been strict, but where the students participating in the program (and possibly selected for participation based on their expected performance) happen to be clustered around a certain value of a possible forcing variable. Given the relatively low share of schools with seemingly strict assignment, it is possible that a large share of these are such spurious schools. Also, balancing checks for this sample (to be discussed in section 4.4) suggest that the assumption A1 may not be satisfied. Thus, we will not use this sample to estimate effects of the program.

4.3 Estimation

The parameters derived in section 4.1, the intention to treat (ITT) and local average treatment (LATE) for the compliers, are the difference of the limits across the cutoff. In practice we need to use observations away from the discontinuity in the estimations. The main challenge in practice is therefore how to model the functional form in terms of order of polynomials, the relative weight given to each observation, and choosing the interval from which

to draw the data. The nonparametric regressions presented in Figure 5 below suggests that a linear model is a good approximation to the underlying data. We will thus estimate linear regressions allowing the slope to differ at each side of discontinuity. We estimate these locally, restricting the sample to where the standardized GPA (cutoff set to zero), still denoted g_{itm}, is less than or equal to the absolute value of the bandwidth b, i.e. $-b \leq g_{itm} \leq b$. To show the sensitivity to the choice of samples we will present estimates for four different bandwidths, from a quarter of a grade point on each side of the cutoff to one and a half grade points. In all models we use a triangular kernel to weight the observations, in effect 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 we follow the recommendation of Lee and Card (2008) and cluster the sampling errors on these discrete values of the assignment variable. Estimation is done using the Stata procedure rd (Nichols (2007)).

With the model choices we can express participation in the program, d_i, in terms of the first term GPA, g_{itm}, and the indicator variable for being in the target group. If a student has a standardized first term GPA less or equal to zero he or she is in the target group, i.e. $Target_{itm} = 1[g_{itm} \leq 0]$. In the preferred sample we only have one municipality, Stavanger, but this specification allows for the municipality and year specific cutoffs we search for in section 4.2. As mentioned above, for Stavanger the cutoff for the first cohort is found to be at the 11th percentile.

$$d_i = \mu_{j0} + \mu_{j1} Target_{itm} + \mu_{j2} g_{itm} + \mu_{j3} g_{itm} \cdot Target_{itm} + u_{ji} \tag{6}$$

The estimate for the coefficient μ_{j1} is the sample analog to the denominator in Equation 4, the difference in probability of participation for student i. We here allow this probability to differ for the different j outcomes studied, as the population comprise of the individuals comprise those with non-missing values for each of the outcomes.

The outcomes can similarly be expressed

$$y_{ji} = \alpha_{j0} + \alpha_{j1} Target_{itm} + \alpha_{j2} g_{itm} + \alpha_{j3} g_{itm} \cdot Target_{itm} + v_i, \tag{7}$$

where the coefficient on $Target$ in this equation is the estimate for the intention to treat, the numerator in the wald estimand in Eqution 4.

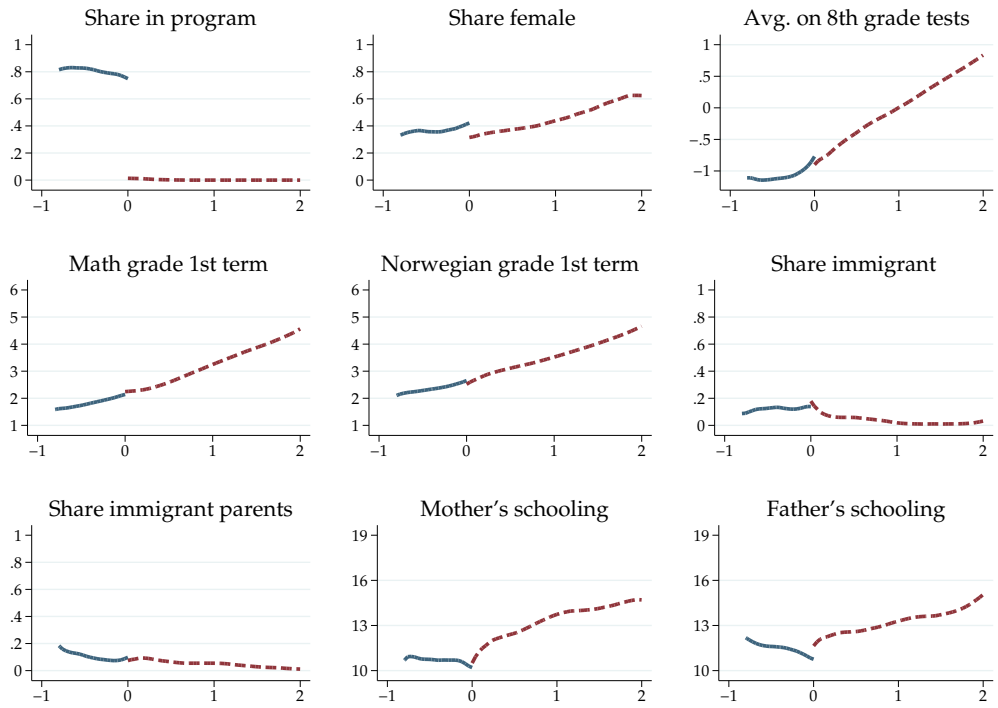
Finally we estimate the ratio, the LATE parameter, with the two-stage least squares estimator with the following structural equation. Participation is instrumented with target group membership.

$$y_{ji} = \beta_{j0} + \beta_{j1} d_{ji} + \beta_{j2} g_{itm} + \beta_{j3} g_{itm} \cdot Target_{itm} + \varepsilon_i, \tag{8}$$

4.4 Assessing the identifying assumptions

In order to estimate to causal effects we need the probability of participation to change discontinuously at the cutoff, while the $y(0)$ and $y(1)$ do not (assumption A2 in section 4.1). The change in treatment can be estimated directly, as described in the previous subsection. Whether outcomes given treatment/no treatment changes is fundamentally untestable, but they will be if there is indeed local randomization around the cutoff. That is, in some interval around the cutoff students (or teachers) do not choose whether a student is above the cutoff (and thus not participates) or below the cutoff (and thus participates). As discussed in Section 4.1 above it is to ulikely that students could precisely manipulate their position relative to the cutoff andthey had no way of knowing what the cutoff would actually be in the first year of the program. It is more conceivable that the teachers could manipulate the first term GPA of their students, to shift them in and out of treatment. First term grades are entirely set by the students' teachers, who know the students well, and may have a preference for whether the student should participate. However, at least in Stavanger where assignment seems to be based strictly on a municipality-specific cutoff, and a large number of students are spread over several schools, teachers cannot reasonably know the cutoff when setting their students' grades.Still, with a selection of baseline covariates this local randomization can be tested, much in the same way as the global randomization is tested in papers using data from randomized controlled trialsLee and Lemieux (2010). Following the suggestions in Lee and Lemieux (2010) we present all results and balance checks for four different bandwidths of the assignment variable, the first term GPA. These bandwidths are a quarter of an average grade-point, half a point, one point and one and a half grade-points.

Figure 3: Balancing tests: Composition of student characteristics around discontinuity



Notes.

Figure 3 shows how program participation and student characteristics change around the cutoff in our estimation sample. First, there is a clear discontinuity in program participation, which drops from a stable level just below 80 percent to zero. This means that the first first requirement is satisfied. Furthermore, student performance, measured by performance on standardized test in 8th grade, and the first term grades in the specific subjects Math and Norwegian (which make up part of the forcing variable first term GPA) show no sign of discontinuities. On the other hand there seem to be some indication of differences in the student composition with respect to gender and parental education. However, from Table 2, where we present rd estimates of the changes in the different characteristics for different bandwidths, we see that all changes are far from being significant. Thus, what may look like systematic differences in Figure 3 is likely to only be random variation.

Table 2: Balancing tests: Composition of student characteristics around discontinuity

	(1) .25	(2) .50	(3) 1.00	(4) 1.50
Share in program	0.742**	0.736**	0.744**	0.758**
	(0.088)	(0.070)	(0.057)	(0.053)
Share female	0.187	0.107	0.078	0.091
	(0.187)	(0.119)	(0.086)	(0.075)
Avg. on 8th grade tests	-0.287	0.131	0.006	-0.072
	(0.398)	(0.203)	(0.130)	(0.110)
Math grade 1st term	-0.073	-0.101	0.035	0.100
	(0.259)	(0.149)	(0.106)	(0.091)
Norwegian grade 1st term	0.082	0.134	-0.040	-0.054
	(0.235)	(0.145)	(0.103)	(0.090)
Share immigrant	-0.124	-0.040	0.034	0.032
	(0.153)	(0.089)	(0.058)	(0.049)
Share immigrant parents	0.149	0.025	-0.020	-0.013
	(0.106)	(0.068)	(0.051)	(0.044)
Mother's schooling	-1.857	-0.298	-0.886	-0.905
	(1.919)	(1.214)	(0.837)	(0.734)
Father's schooling	-2.119	-0.885	-1.232*	-1.119*
	(1.392)	(0.888)	(0.664)	(0.589)
Observations	171	311	608	919
Wald test of joint significance, all but 'Share in program'	7.595	3.922	6.155	9.128
p-value Wald test	0.474	0.864	0.630	0.332

Notes. ; stars indicate the significance level (* p<0.10, ** p<0.05).

A test for manipulation of the assignment variable using the density, following McCrary (2008), looks somewhat more suspicious also for the preferred sample.[5] There are patterns that resemble manipulation with large drops after the cutoff found in section 4.2. However, further inspection shows that these drops are not particular to the cutoffs, rather they appear at regular intervals. This potential issue of "heaping" in the assignment variable is shown to induce bias in the estimates in (Almond et al., 2010). In Barreca et al. (2012) they study this potential issue of "heaping" in the assignment variable further and propose tests. In this particular case the heaping has a straightforward explanation. 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 when the distribution is shown with high resolution. Students with 12 grades could be systematically different, they are for one more likely to *not* be defined as special needs, but in Figure A.5

[5]Results available upon request.

there is no indication of any systematic difference.

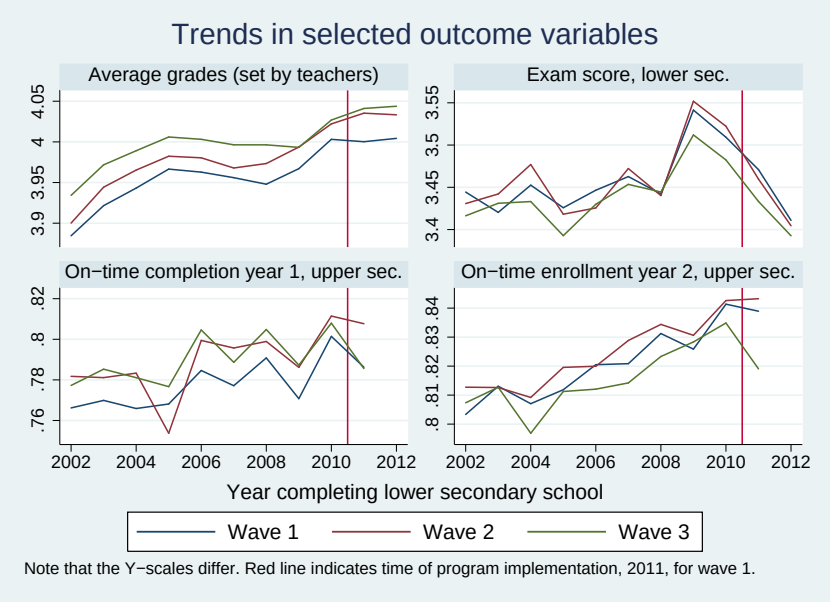
We have also performed balancing checks for the schools found to be possibly strict, cf. section 4.2. These checks are presented in Figure A.6 and Table A.1. As opposed to what was the case for the Stavanger sample, we here find some indication of significant differences in students' characteristics around the cutoff. This may be an indication that the cutoffs to some extent are endogenous with respect to the teachers' expectations about the effect of the program on the individual student. In any case, the finding suggest a possible violation of assumption A1, such that we cannot draw credible causal inference from this sample.

4.5 Difference-in-differences (DiD) estimation

An alternative approach to estimating the effects of the program is a difference-in-differences (DiD) estimation at the school level, exploiting the fact that the program was implemented over three years. We can then compare how the students' outcomes evolves in the schools where the program was offered early to those schools where it was offered later. This could produce an intention to treat effect at the school level. Ideally one would like this timing to be random, however, this was not the case. In Eielsen et al. (2013) we discuss the selection of schools and municipalities for participation in more detail.

Still, as long as any differences in average outcomes between schools are stable we may eliminate these through difference-in-differences estimation. Figure 4 show the evolution of several outcome variable for students graduating from schools in the first, second and third wave. There seem to be a persistent difference in levels, particularly for average teacher grades, however, the pre-program evolution (left of the vertical line) are roughly similar.

Figure 4: Trends in selected outcome variables



To improve on the DiD-estimates it is possible to construct a comparison group of students in untreated schools using matching techniques. The idea of matching is to find comparable schools based on observable pre-determined characteristics such as school size, teacher density or pre-treatment results. In practise we have estimated propensity scores (i.e., a probabilities of participation) and balanced all covariates through full Mahalanobis matching, this was done with the Stata module psmatch2 (Leuven and Sianesi, 2003). As the selection of schools into treatment was done either at the county or the municipality level, and is mostly a black box for us as researchers, this was a data driven exercise. A successful specification will provide common support i.e. that the densities of the predicted probabilities of participation for both non-participants and participants are overlapping. Based on this we select a sample and make the strong assumption that after conditioning on the selected covariates, receiving treatment is "as if" random. This assumption cannot be tested directly, but comparable to conventional DiD analysis we look at whether the conditional trends of pre-treatment outcomes are similar as a way of validating the approach.

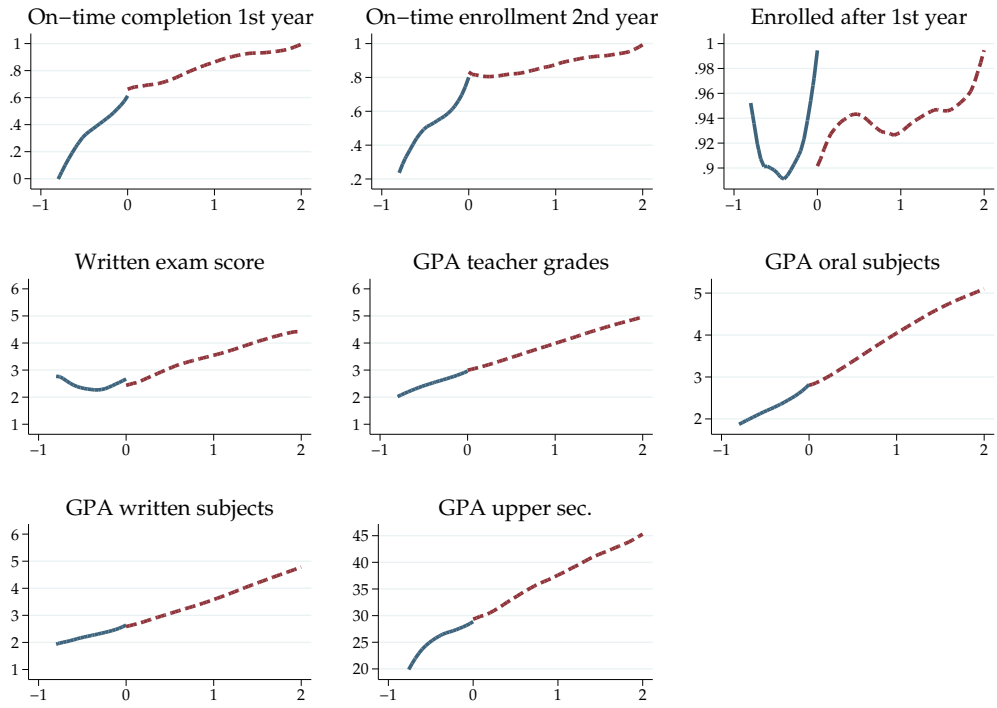
With these samples of comparable treated and non-treated schools we do a regular DiD analysis with additional control variables to estimate the effect of the program and argue that the estimates for the program effect is causal unless there are important unobservable variables that change both over time and between schools that determine, in part, both our outcomes and program participation. In this setup with non-experimental data that is as close as it is possible to get however, and we can only argue whether or not there exist plausible such omitted variables and discuss theoretically which way we would expect this to

bias our estimate. However, the matching-DiD did not provide notably different results from DiD on the full sample, see (Eielsen et al., 2013) for details.

5 Results

In this section we present the estimated effect of the program. Figure 5 shows how the outcomes of interest change around the discontinuity. None of the outcomes show any clear effect. All outcomes seem to vary continuously, with a possible exception of the share enrolled after first year. However, there is little variation in this variable, and we see that, relatively to the scale of the figure, this share is also volatile away from the discontinuity.

Figure 5: ITT estimates



Notes.

Looking at the RD estimates presented in Table 3, where we present ITT and LATE estimates with standard errors, we see that all estimates are far from being significant. The results in Table 3 are estimated using a bandwidth of half a average grade point. Table A.2 in the Appendix shows estimates for four different bandwidths, still none are significant.

Table 3: ITT and LATE estimates

	(1) Obs. in bwidth count	(2) ITT b/se	(3) LATE b/se
On-time completion 1st year	311	-0.049	-0.066
		(0.120)	(0.164)
On-time enrollment 2nd year	311	-0.032	-0.043
		(0.098)	(0.133)
Enrolled after 1st year	311	0.093	0.126
		(0.064)	(0.087)
Written exam score	295	0.235	0.323
		(0.199)	(0.270)
GPA teacher grades	310	-0.035	-0.048
		(0.067)	(0.091)
GPA oral subjects	306	0.014	0.018
		(0.100)	(0.135)
GPA written subjects	299	0.053	0.072
		(0.104)	(0.142)
GPA upper sec.	275	-0.462	-0.637
		(1.916)	(2.650)

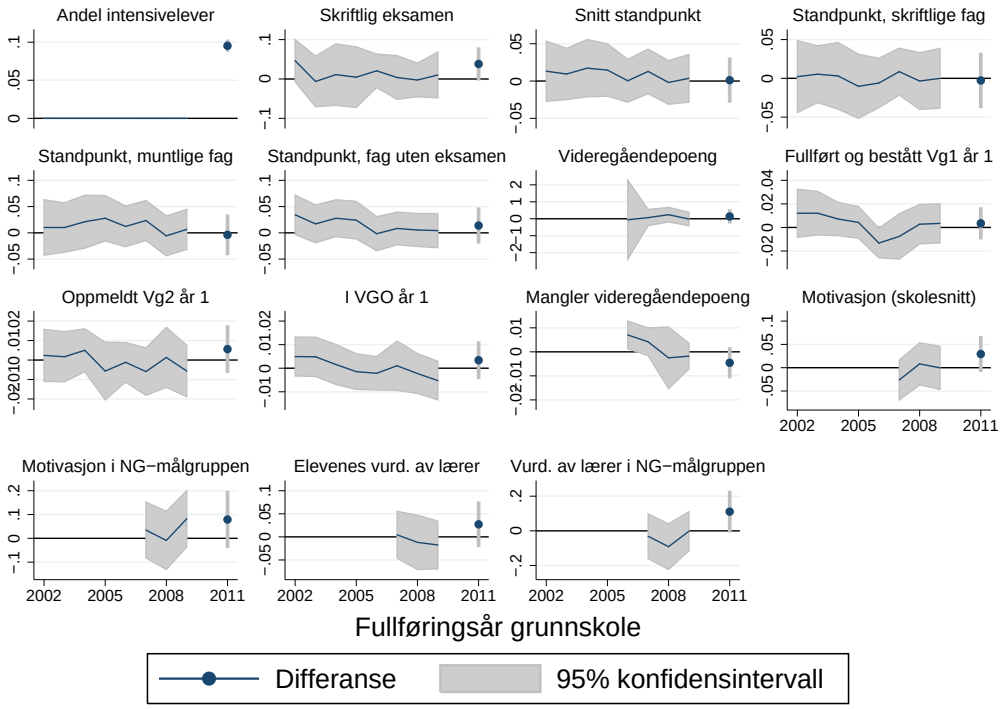
Notes. The bandwidth is half a grade point; stars indicate the significance level (* $p<0.10$, ** $p<0.05$).

While we do not find any effect on any of the outcomes studied, we are not able to rule out substantial effects. For example, in Table 3, the standard error of the estimated effect of completion of first year of upper secondary is over 16 percentage points and the standard error on written exam score is .27 grade points (about 1/4 of a standard deviation). Thus, any effect would need to be very large in order for us to be able to reject the null.

5.1 Supplementary results from DiD estimation

To supplement the RD estimation we have also undertaken DiD estimation. While the RD effect is an individual-level effect, corresponding to the effect we would expect on a marginal student in a program school being assigned to the program, the DiD estimate captures the school-level average effects. Thus, while the RD effect is the difference between the direct effect on a (marginal) treated student and the spill-overs (if any) on a (marginal) untreated student, the DiD estimate will capture an average of all kinds of effects, direct effects and spillovers, across all students.

Figure 6: DiD-estimates



Notes.

In Figure 6 (taken from (Eielsen et al., 2013)) we present differences in average outcomes between schools in wave one and wave two/three. All differences are relative to the 2010 difference, i.e., the difference in the last year before the introduction of the program. From Figure 6 we see that there is a clear change in the share of students participating, as expected a 10 percentage point increase. Furthermore, there are no signs of significant pre-reform differences, suggesting that the identifying assumption underlying the use of DiD may indeed be satisfied.

However, we do not find any effect on any outcome variable using DiD either. As we use a much larger sample (all students in all schools) the DiD estimates are much more precise than the RD estimates. However, most students are at most affected through spillovers. If we were to assume that any effect of the program was a direct effect on the participating student, we could find estimate this average treatment effect by scaling with the change in program participation. This would give ATE estimates with limited precision.

We have tried to estimate DiD for different parts of the first term GPA distribution, in order to investigate if there is any effect on groups of students with a larger change in the share of participants. The reduced samples gives less precise ITT estimates, but as the greater change in participation may enable us to estimate ATEs more precisely. However, we

still do not find any evidence of any effects, details are provided in (Eielsen et al., 2013).[6]

6 Conclusion

We have shown how a search over possible definitions and values of the forcing variable has successfully recovered the first term GPA threshold for participation in the program. However, the same procedure applied to all schools produced a sample that does not allow credible estimation of effects of the program, possibly because we have found a large share of schools that by chance seem to follow a strict assignment rule, without actually doing so. Also, estimation on the Stavanger sample provided little information on effects. No effect estimates are significantly different from zero, but the results are very imprecise, and thus hard to interpret.

We have complemented the RD analysis with a DiD analysis, using the gradual introduction of the program to study differences between schools. The RD and DiD analyses are not directly comparable, neither in terms of effect estimated, estimation sample nor the assumptions we must make to estimate credible effects. However, they are consistent in the sense that none of the estimates suggest a significant effect, but all the estimates suffer from lack of precision. Furthermore, we are not able to study the outcomes explicitly targeted by the program. The participating students have not yet had the time to complete upper secondary, and grades are not a perfect measure of basic skills. Cortes and Goodman (ming) study an intensive training program and find that there are effects on graduation, in spite of lacking immediate effects on performance.

However, it would be unsurprising if the program had at best small effects. Firstly, there is both theory and evidence suggesting that early interventions focusing are more effective than later ((Carneiro and Heckman, 2003; Heckman and Cunha, 2007)), and towards the end of compulsory school may be too late to make a large impact. This is especially the case as the program studied is of limited extent compared to programs that have proven effective. For example, it does not increase the amount of instruction (as opposed to Cortes and Goodman (ming)), but rather changes the group size and pedagogy. Still, Machin and McNally (2008) is an example where only changing the pedagogy is found to be effective. Another reason for the lack of precise results is the likely difference in treatment between schools.(Sletten et al., 2011) reports substantial variance in the group size in which the trainings took place, and the differing practice implementation the assignment rule could also be an indication of more

[6]As seen in sections 3 and 4.2 participation is not very well predicted by first term GPA except in Stavanger. Thus, while the change in the share participating increases, the change is still much smaller than in the RD estimations. Stavanger alone constitute a too small sample to estimate DiD with any precision.

fundamental differences at both the municipality and the school level. With this potentially substantial treatment heterogeneity there could be both effective and ineffective versions of the program canceling each other out.

Still, the limited extent 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 may be economically relevant. With the current data, we are not able to identify such small effects. However, as the program has now be going on for three years more data will eventually become available. Assuming that the program effect (if any) does not change over time, a larger data set will allow more precise effect estimation.

Several years of data also opens for extensions of the procedure we apply to search for participation thresholds. If having a strict cutoff is persistent over time (not necessarily the same cutoff) and spurious cutoffs are random events that are not persistent, we may be able to more precisely identify schools with strict cutoffs. This would in turn allow a larger sample and more precise RD estimation, as well as a further evaluation of the general merit of our search procedure.

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., 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.

Cortes, K. E. and Goodman, J. S. (forthcoming). Ability-tracking, instructional time and better pedagogy: The effect of double-dose algebra on student achievement. *American Economic Review: Papers & Proceedings*.

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

Eielsen, G., Kirkebøen, L. J., Leuven, E., Rønning, M., and Raaum, O. (2013). Effektevaluering av intensivoppæringen i overgangsprosjektet, ny giv. Technical report, Report 54/2013, Statistics Norway.

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

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

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.

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics*, 116(3):777–803.

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.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48:281–355.

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.

Leuven, E. and Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components, Boston College Department of Economics.

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.

Nichols, A. (2007). RD: Stata module for regression discontinuity estimation. Statistical Software Components, Boston College Department of Economics.

OECD (2013). Education at glance 2013.

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.

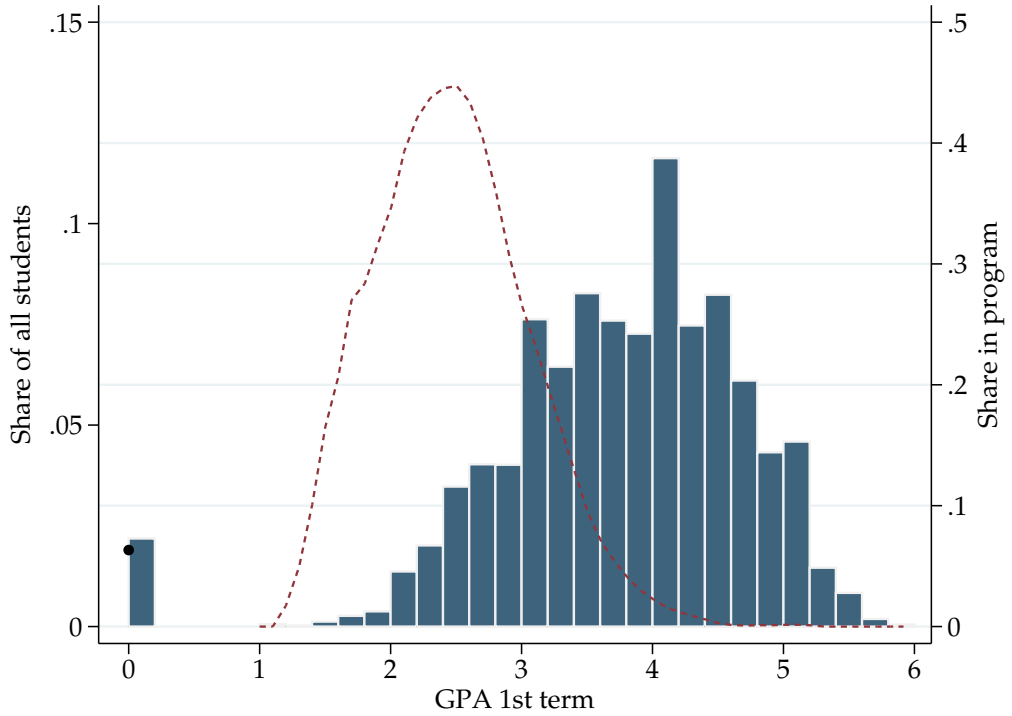
Sletten, M. A., Bakken, A., and Haakestad, H. (2011). Ny start med ny giv? kartlegging av intensivopplæringen i regi av ny giv-prosjektet skoleåret 2010/11.

Utdanningsdirektoratet (2013). Gjennomf/oringsbarometeret 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.

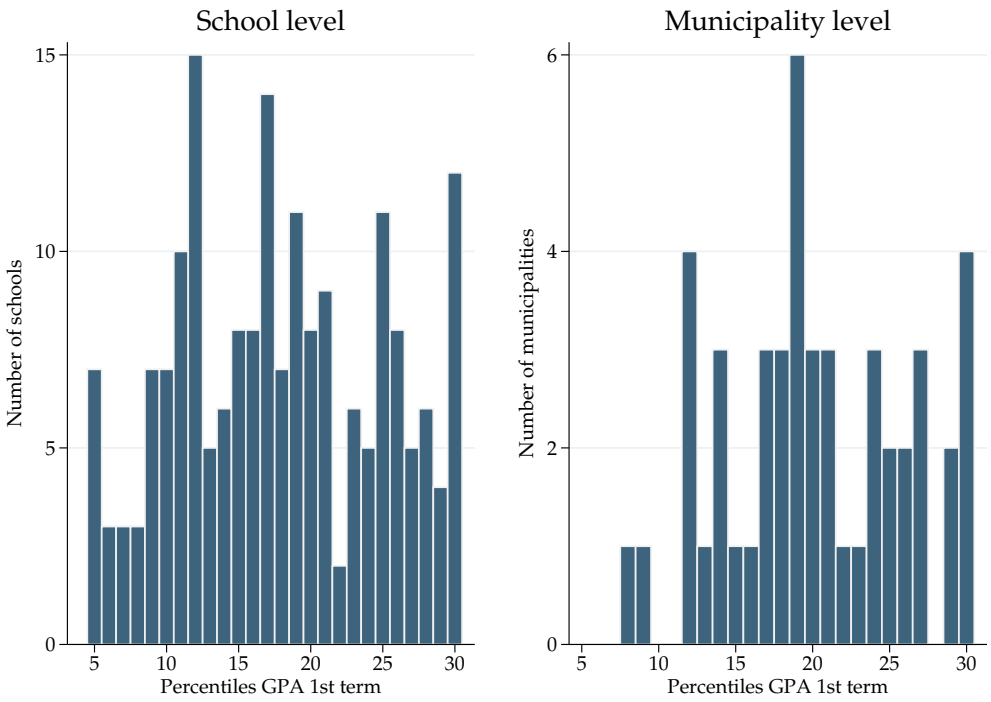
Appendix

Figure A.1: Pupils in wave 1 schools



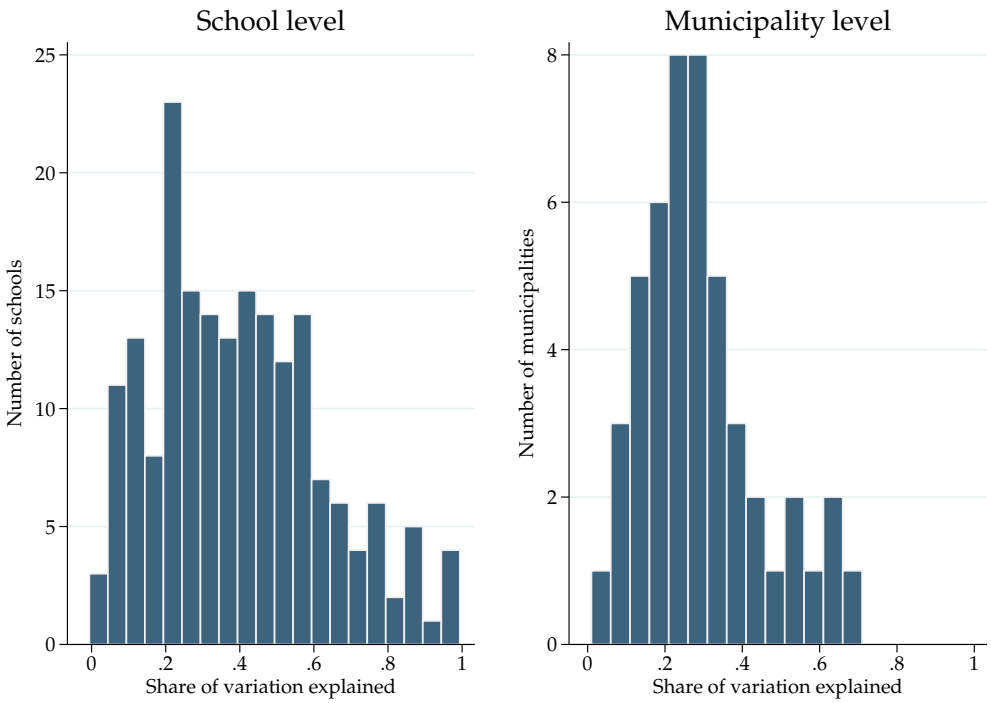
Notes. The density shows the distribution of first term GPA for students participating in the intensive training, while the histogram show the distribution of other the other students in the participating schools.

Figure A.2: Effective percentiles used



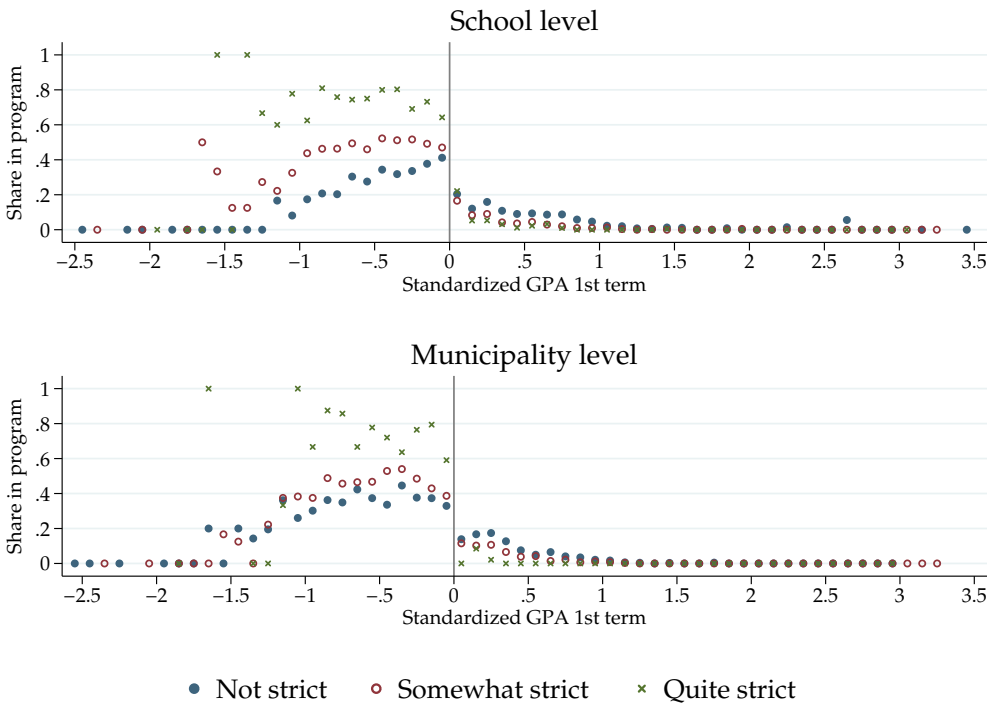
Notes.

Figure A.3: Degree of strict assignment



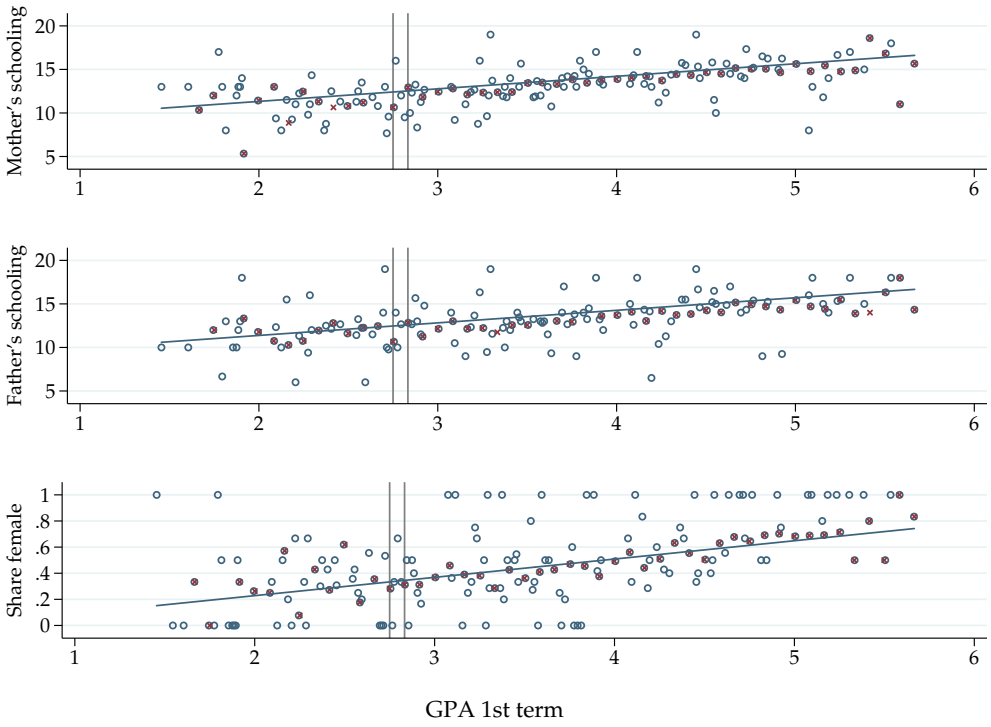
Notes.

Figure A.4: Probability of participation by "strictness"



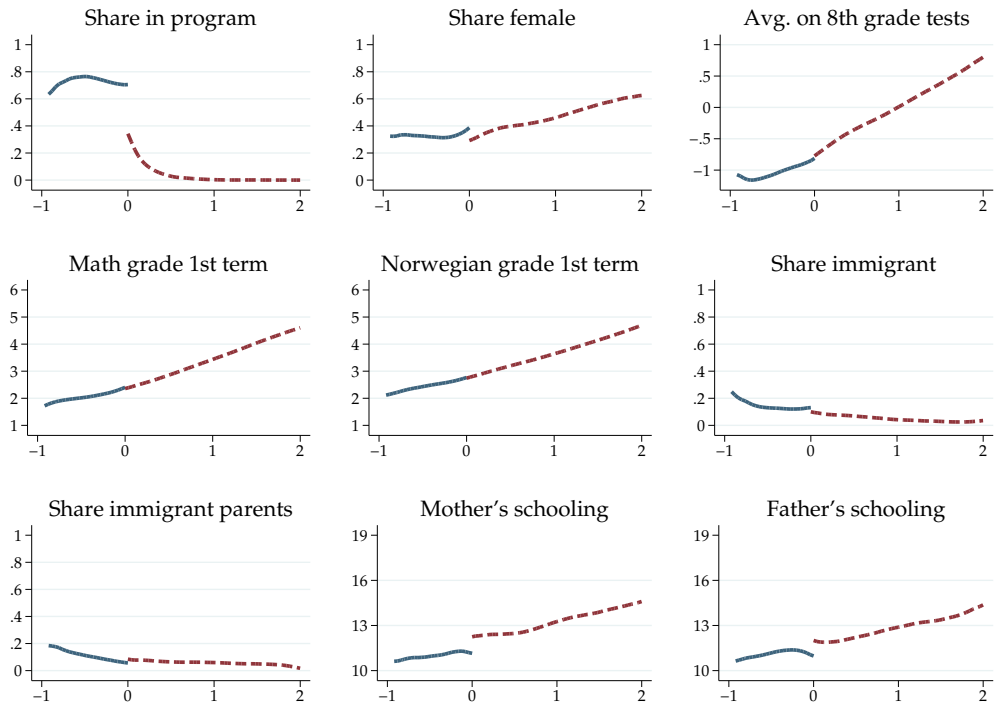
Notes.

Figure A.5: Covariates vs. assignment



Notes.

Figure A.6: Balancing tests: Composition of student characteristics around discontinuity



Notes. Sample of "strict" schools as found with search procedure.

Table A.1: Composition around cutoffs, school cutoffs

	(1)	(2)	(3)	(4)
	.25	.50	1.00	1.50
Share in program	0.170**	0.363**	0.502**	0.568**
	(0.085)	(0.059)	(0.041)	(0.035)
Share female	0.074	0.095	0.027	0.024
	(0.086)	(0.062)	(0.045)	(0.039)
Avg. on 8th grade tests	0.044	-0.039	-0.096	-0.120**
	(0.134)	(0.098)	(0.070)	(0.061)
Math grade 1st term	0.149	0.045	0.023	0.030
	(0.138)	(0.097)	(0.069)	(0.059)
Norwegian grade 1st term	0.069	0.023	-0.016	-0.011
	(0.106)	(0.080)	(0.058)	(0.051)
Share immigrant	0.036	0.031	0.030	0.026
	(0.055)	(0.042)	(0.030)	(0.026)
Share immigrant parents	-0.040	-0.028	-0.027	-0.023
	(0.056)	(0.035)	(0.025)	(0.021)
Mother's schooling	-1.788**	-1.098**	-0.970**	-0.724**
	(0.672)	(0.528)	(0.394)	(0.348)
Father's schooling	-1.192*	-1.031**	-0.480	-0.322
	(0.656)	(0.508)	(0.380)	(0.338)
Observations	568	1097	2109	3081
Wald test of joint significance, all but 'Share in program'	13.762	12.475	12.332	12.055
p-value Wald test	0.088	0.131	0.137	0.149

Notes. ; stars indicate the significance level (* p<0.10, ** p<0.05).

Table A.2: ITT estimates with different bandwidths

	(1)	(2)	(3)	(4)
	.25	.50	1.00	1.50
On-time completion 1st year	-0.104	-0.049	-0.046	-0.038
	(0.210)	(0.121)	(0.085)	(0.073)
On-time enrollment 2nd year	-0.005	-0.032	-0.054	-0.038
	(0.192)	(0.098)	(0.073)	(0.064)
Enrolled after 1st year	0.250	0.093	0.036	0.029
	(0.153)	(0.064)	(0.039)	(0.033)
Written exam score	0.015	0.235	0.159	0.034
	(0.351)	(0.200)	(0.149)	(0.133)
GPA teacher grades	-0.036	-0.035	-0.017	-0.011
	(0.110)	(0.068)	(0.049)	(0.042)
GPA oral subjects	-0.048	0.014	0.032	0.029
	(0.162)	(0.100)	(0.074)	(0.064)
GPA written subjects	-0.036	0.053	0.028	0.040
	(0.174)	(0.105)	(0.072)	(0.061)
GPA upper sec.	-1.150	-0.462	0.135	-0.089
	(3.719)	(1.927)	(1.317)	(1.130)
Observations	171	311	608	919
Wald test of joint significance	3.254	4.378	3.264	2.135
p-value Wald test	0.917	0.822	0.917	0.977

Notes. ; stars indicate the significance level (* p<0.10, ** p<0.05).