Enriching students pays off: Evidence from an individualized gifted and talented program in secondary education*

Adam Booij, Ferry Haan, and Erik Plug

July 5, 2016

Abstract

We examine the effect of a gifted and talented program in academic secondary education. Students are assigned based on a cutoff score in a cognitive aptitude test, which we exploit in a fuzzy regression discontinuity framework to identify program effects. We find that assigned students obtain higher grades, follow a more science intensive curriculum (most notably for girls), and report stronger beliefs about their academic abilities. We also find that these positive effects persist in university, where students choose more challenging fields of study with, on average, higher returns. Together, these findings are consistent with a human capital interpretation.

JEL-codes: I22; I28

Keywords: gifted and talented education; enrichment program; secondary education; regression discontinuity

*Adam Booij (corresponding author): Amsterdam School of Economics, University of Amsterdam, The Netherlands, and Tinbergen Institute (email: Adam.Booij@uva.nl). Erik Plug: Amsterdam School of Economics, University of Amsterdam, The Netherlands, Tinbergen Institute, IZA, and UCLS (email: e.j.s.plug@uva.nl). Ferry Haan: Amsterdam School of Economics, University of Amsterdam, The Netherlands (email: f.h.g.haan@uva.nl). The authors thank Stedelijk Gymnasium Nijmegen (SGN) for access to their schoolregister, SEO Amsterdam Economics for collection and dissemination of data on starting salaries of Dutch university graduates, and DUO for data on former SGN-students in tertiary education. We further thank the GT team at SGN for their cooperation in all stages of this research project. We also thank seminar and conference participants in Amsterdam, Braga, Izmir, Ljubljana, Nijmegen, Stanford, Utrecht, and Venice for their comments and suggestions. Financial support from the Netherlands Initiative for Education Research (NRO 411-12-637) is gratefully acknowledged.

1 Introduction

Many educators are advocating targeted education to gifted students. If gifted students underperform because of an unchallenging school environment, they argue that special education programs, which are generally referred to as gifted and talented (GT) programs, can help these students to reach their full (academic) potential, possibly with positive social returns. While GT education programs become increasingly popular (e.g. Bhatt, 2011), empirical evidence in support of GT program benefits is scarce. What complicates the evaluation of GT programs is that students who receive it are, by definition, a very selective group. Any positive association between student performance and GT education is therefore not causally interpretable (Matthews et al., 2012).

In this paper we estimate the effect of a GT program implemented at a prestigious academic secondary school in the Netherlands. The GT program we consider is an individualized pull-out program, based on ideas of Renzulli (1977), in which students freely decide to replace classroom teaching for in-school time to work on self-selected projects (enrichment). The GT program is offered to gifted students. Students qualify as gifted based on a cutoff score in a standardized cognitive aptitude test that all students take at school entry. About 25 percent of students get selected, most of whom come from the top 4 percent of the nationwide ability distribution. Gifted students remain program eligible for six years, which is how long academic secondary education takes. We exploit the assignment cutoff in a fuzzy regression discontinuity (RD) design to identify GT program effects on school performance.

We find that students assigned to the program obtain higher grades, follow a more science intensive curriculum (most notably for girls), and report stronger beliefs about their academic abilities. We also find that the positive program effects persist in university, where students choose more challenging fields of study with, on average, higher returns. In addition, we test for possible adverse program effects among students excluded from the program. We find no evidence that these students experience feelings of disappointment for being left out, or miss out on classroom spillovers. Together, these results are consistent with a human capital interpretation.

Our paper relates to a small number of recent papers on the causal effect of GT education on student performance.[1] Card and Giuliano (2014) apply a fuzzy RD design to estimate the effect of a GT education program on math, reading, and writing test scores of US elementary school students. The GT program they investigate puts gifted students together in classrooms with students who performed well in previous grades, receiving a modified enriched curriculum. They find little, if any, test score gains for gifted students, but large and positive test score gains for students with high grades in previous years (which they refer to as high achievers). Bui et al. (2014) examine the effect of GT programs on math, reading, and language test scores of middle school students in Southwestern US. Since the middle schools have some autonomy over what is taught to gifted and talented students, they estimate an average program effect of GT programs that take many shapes and forms. In one application, they use a fuzzy RD design and find that students close to the GT program admission cutoff do equally well, regardless of GT program exposure. In another application, they exploit a lottery in oversubscribed middle schools that offer GT programs to identify causal program effects and find, again, that the GT program had no influence on academic performance.[2] Bhatt (2012) also looks at US middle school students but uses an instrumental variable strategy that exploits differences in GT program admission rules between schools. She finds positive test score gains, but this may (partly) reflect sorting of students to schools.

Our paper adds to this literature in several ways: it does not look at an intervention with separate gifted classes and teaching, but rather at a much simpler individualized pull-out program; it takes a longer-run perspective beyond academic secondary school and tests whether program effects persist

[1]There is much empirical work on the relationship between GT programs and student performance. See Bhatt (2011) for an overview of most of these studies, which (almost) all find positive associations between program exposure and achievement. Because these positive associations cannot make a distinction between selection and causation, their interpretation remains unclear.

[2]In a comparable experiment, Davis et al. (2010) exploit a fuzzy RD design to estimate whether GT programs can help public schools to retain gifted students. Close to the admission cutoff, they find that students who are considered gifted are more likely to stay in public schools.

in university using matched university enrollment records; also, it combines school registers with survey data on student habits and attitudes, which enable us to look at mechanisms and better understand why the GT program works.

The remainder of the paper proceeds as follows. Section 2 briefly discusses secondary education and the academic secondary school in which the GT program takes place, the GT program, and GT program assignment. Section 3 describes data, experimental design, and the assumptions needed to identify GT program effects. Section 4 presents and discusses the main empirical findings. Section 5 follows with an assessment of potential mechanisms. Section 6 summarizes and concludes.

2 The GT program at SGN

The GT program we examine in this paper was implemented at Stedelijk Gymnasium Nijmegen (SGN), an academic secondary school in the Netherlands. In this section, we provide a short outline of the Dutch secondary education system that SGN is part of, describe the GT program in more detail, and particularly focus on the assignment rules that we exploit for identification.

2.1 Academic secondary education in the Netherlands

Dutch secondary education is a tracking system that funnels pupils through one of three tracks: pre-vocational secondary education (VMBO), general secondary education (HAVO), or academic secondary education (VWO). The selection is based on teacher recommendations and CITO scores, a national secondary school admission test taken in the final year of primary education (age 11). VWO is the most advanced track, which takes 6 years (grade 1 at age 12 to grade 6 at age 18) and prepares students for university education. In the final year, students take a nationwide exam which gives access to university (conditional upon passing). The VWO track hosts approximately 20 percent of all secondary school-going students. VWO is further divided into atheneum and gymnasium schools. Gymnasium schools are the most selective and prestigious schools with an academic curriculum similar to that of atheneum schools, complemented with classical languages Latin and Greek. Gymnasium schools attract about 5 percent of all students in secondary education.

The Netherlands has 38 independent gymnasium schools, which are brought together under one foundation for, among other things, sharing experiences on how to successfully educate academically promising students. All these gymnasium schools offer comparable enrichment programs to gifted and talented students. Of these gymnasium schools, SGN gave us access to their school registers.

2.2 The GT program at SGN

In 1983 SGN was one of the first gymnasium schools to introduce a GT program. With help from the Center for the Study of Giftedness (CBO) at the Radboud University Nijmegen, SGN developed a special education program targeted at gifted students. In program design, SGN and CBO followed Renzulli's notion that students with exceptional cognitive and non-cognitive skills should receive an enriched education program with exposure to new content, active application of own skills, and creation of a product (Renzulli, 1977, 1986). The GT program has the following features.

1. The GT program is an individualized pull-out program; that is, qualified students receive the right to trade in classroom lessons for project time (spent elsewhere) to work on a project of their own choice.

2. SGN provides rooms, computers, and arts and crafts facilities to help students in their projects.

3. Participating students can choose which classroom hours to devote to their project, with a minimum of two hours each week.

4. Teachers can deny students the right to trade in a specific lesson, but they need to argue why this specific lesson can not be missed.

5. At the beginning of the school year, students choose and develop a project topic, which can be anything (within legal limits of course).[3]

6. At the end of the academic year, students present their projects to teachers, parents, and fellow students.

7. Students are supervised by specialized SGN teachers, referred to as GT coaches, throughout the development of the project.

8. GT coaches and students meet every two weeks; GT coaches provide hands-off supervision aimed at having a finished project at the end of the year.

9. Projects are not graded.

When qualified students are not working on their project, they follow the same classes, face the same curriculum, and do the same exams as the other students. Qualified student typically participate in projects in the first 4 years of school.

2.3 GT program assignment

GT program participation is exclusive to students who qualify as gifted.[4] Qualified students can choose not to participate, but they cannot undo their qualification. Students qualify on the basis of both cognitive and non-cognitive traits. CBO, on behalf of SGN, administers an intelligence test (IST test) and a motivation-to-learn test (FES test) of all students at the beginning of the first school year. The time-line is as follows: the school year starts in September; students take the IST and FES tests in October; CBO provides test results to SGN staff, after which the GT team decides upon selection in December.

[3] Examples of projects include writing a cookbook, learning Russian, designing a soccer stadium, developing software for 3D dinosaurs.

[4] We should note that SGN labels qualified students as enriched students because of their enriched curriculum. We instead follow the convention in literature and refer to qualified students as gifted students (or GT students) because of their selection on cognitive test scores.

In the selection of gifted students, IST test scores are leading. Since 1998, SGN applies a two-stage assignment procedure. In the first stage, the GT coordinator compiles a list of potentially eligible participants. This is merely a mechanical exercise; that is, all first year students are ranked on their IST scores and those students with an IST score above a certain cutoff are marked as potentially eligible. The cutoff is typically set at one standard deviation above the IST score mean, and then adjusted according to GT capacity (which depends on the number of enrolled first-year students, the number of GT coaches, and the number of gifted students from previous years).

In the second stage, the list is then used as input for the GT team (including GT coordinator, GT coaches, and class mentors) to decide upon actual assignment. While the advice of the GT coordinator is mostly followed, approximately 10 percent of students switch assignment status. Students with high IST scores do not always qualify, while students with low IST scores sometimes do. GT team members, whom we interviewed about eligibility criteria, report that assignment could change because of inadequate motivation, remaining capacity concerns or, in some years, unexpected performance on certain components of the IST test. Nonetheless, if there is a sharp increase in the probability of assignment at the IST cutoff, the two-stage assignment process mirrors that of a fuzzy regression discontinuity design which we will use to estimate GT program effects.

3 Data and design

3.1 Data

Our baseline sample is drawn from the SGN student administration. This digitalized register contains detailed information on all students enrolled at SGN. In particular, it holds student records on basic demographic characteristics, such as gender and age, primary education exit exam scores (CITO test scores), GT program assignment status, and any other school- and exam-grade obtained from the day of entry until the day of leave. CBO test scores on intelligence and motivation (IST and FES test scores) and other data on the enrichment program are stored in an (analog) archive, of which SGN provided us copies.

We use the SGN register to construct several measures of academic achievement in secondary education: grade retention, overall grade point averages (GPA) for math, languages, and other school subjects (grades 2 until 6), and three indicators of choosing an advanced curriculum in grades 5 and 6 (the number of exam subjects, the number of science subjects, and taking advanced math).[5] In the SGN register we keep those students who entered first grade somewhere between 1998 and 2010 and took the IST and FES tests. This gives us a sample of 3,127 students, of which 785 students are assigned to the GT program.

By merging our student data to the national register that keeps track of student enrollment and completion in tertiary education, we can construct several measures of academic achievement in university education: university enrollment, field of study (sciences), switching studies, and the average starting salary that corresponds to field of study.[6] Given that most of these students have just started their studies, we limit the analysis to the choice of field (at the end of) the first year at university. We are able to match field of study choices of SGN students of cohorts 1998 - 2007. The corresponding sample contains 2,438 students who have (ever) entered university.

[5]Grades range from 1 to 10. The language variable is the mean of compulsory subjects Dutch and English. The math variable is the combined score for standard mathematics (math A) and advanced mathematics (math B), using the algorithm by Leuven et al. (2010) which makes both grades comparable by adjusting them by the mean difference in scores of students that choose both.

[6]The data on starting salaries by field of study come from an annual survey held among a representative sample of recent university graduates (Berkhout et al., 2013). These starting salaries are listed (and updated) every year to inform and help secondary school students in their field of study choice (see *Elsevier* Beste Studies 2014 at http://bestestudies.elsevier.nl/). We should note that this approach ignores the returns to skills within field of study, which are likely positive for gifted students.

Table 1. Summary statistics

	Statistics			Sample size	
A: Characteristics	Range	mean	s.d.	Cohort	N
Male	{0,1}	0.54	0.50	1998	236
Age	[9.7, 14.8]	12.16	0.48	1999	264
				2000	270
B: Pre-test				2001	220
raw IST score (forcing variable)	[41, 148]	94.82	15.56	2002	214
raw FES score	[5,40]	23.17	5.36	2003	252
raw CITO score	[518,550]	547.48	2.51	2004	269
				2005	224
C: Treatment				2006	235
GT program	{0,1}	0.25	0.43	2007	254
				2008	234
D: Outcomes grades 2 - 6				2009	195
Matched record	{0,1}	0.98	0.15	2010	260
Retention	{0,1}	0.28	0.45		
GPA math	[3,9.9]	6.61	1.15		
GPA language	[4,9.5]	6.77	0.80		
GPA other	[4.4,9.3]	7.00	0.71		
E: Outcomes grades 5 - 6 [cohort '98 - '07, $N = 1771$]					
Average #subjects	[1,18]	12.60	2.05		
Average #science subjects	[0,5]	2.53	1.67		
Advanced math	{0,1}	0.63	0.48		
F: Outcomes in Higher Education [cohort '98 - '07, $N = 2438$]					
Matched record	{0,1}	0.87	0.34		
Chose science field	{0,1}	0.27	0.44		
Switched fields	{0,1}	0.06	0.23		
Predicted earnings (primary field)	[1300, 5475]	2488	523		
Number of pupils					3127

Note: Panels A, B, and C concern pre-treatment characteristics for the full sample. Panel D are average outcomes of grades 2 to 6 for tested students that were matched to the student registry (98 percent). Panel E covers subject choice outcomes in grades 5 and 6, the number of (science) subjects, and the choice of advanced math, for students that have not repeated a grade. Panel F provides outcomes in further education for a matched sample of 87 percent of students from cohorts '98 - '07.

Table 1 provides sample means and standard deviations of the outcome and control variables that we study below. Two observations follow from this table. First, the students at SGN are quite bright if we look at their CITO scores, a secondary school admission test taken at the end of primary school that measures language and comprehension skills, mathematics, world orientation (which involves geography, biology and history), and study skills. In fact, the average CITO score of 547 lies at the 85th percentile nationwide. Second, SGN classifies about 25 percent of the students as gifted and talented, the majority of whom belong to the 96th percentile of CITO nationwide (not shown in the table). Inasmuch as CITO captures ability, our gifted students therefore belong to the top 4 percent of the ability distribution.

3.2 Fuzzy RD design

The GT program setup allows us to test whether students close to the cutoff are better off when assigned to the GT program using a fuzzy regression discontinuity design. In regression models that adequately account for the influence of IST test scores, a fuzzy RD design is essentially an instrumental variable approach in which student achievement depends on GT program assignment, which is instrumented by a binary indicator for having an IST test score above the cutoff. In particular, we estimate the relationship between academic achievement, GT program assignment status, and some flexible continuous function of the IST distance to the cutoff (which is the running variable in the current fuzzy RD setup) using a two-stage least squares model where the first stage is

$$GT_{it} = \pi_1 Z_i + f(x_i) + \boldsymbol{\pi}' \boldsymbol{w}_i + \lambda_t + \upsilon_{it}, \tag{1}$$

and the second stage is

$$Y_{it} = \beta_1 GT_{it} + g(x_i) + \boldsymbol{\beta}' \boldsymbol{w}_i + \theta_t + u_{it}. \tag{2}$$

In these two equations, Y_{it} is a measure of achievement of student i who took the IST test in year t, and $\boldsymbol{w}_i$ is a vector of exogenous control variables including the student's gender, age (at the IST test), CITO, and FES test scores. GT_i is the endogenous GT program indicator, which equals 1 if a student is assigned to GT education and 0 otherwise, and Z_i is the instrumental variable, which equals 1 if the student has a test score above the cutoff and 0 otherwise. The running variable x_i is the normalized IST score, defined as the difference between the student's IST test score and the IST threshold in the given year. The functions $f(x_i)$ and $g(x_i)$ are flexible polynomials of x_i. In estimation, we model the trend relationship of the forcing variable with outcomes in six different ways: (i) linear; (ii) quadratic (which we take as our baseline model); (iii) cubic; (iv) split quadratic on either side of the cutoff; (v) zoom into ± 18 IST point range with split linear on either side; (vi) donut with ± 4 IST point range removed.[7] The parameters λ_t and θ_t are year (of test taking) fixed effects, and u_i and υ_i are the econometric error terms that may be interdependent. In estimation, the error terms are clustered at the class level.[8] The parameter of interest is β_1, which captures the causal effect of GT program assignment on student achievement among students who barely passed the assignment cutoff.

One limitation is that information on the IST cutoff is available for some years, but not for all. Porter and Yu (2015) show that in such a case, program effects can still be identified using a two-stage procedure where the cutoff is estimated first, followed by (standard) fuzzy RD in the second stage. Moreover, Porter and Yu show that the estimated cutoff in the first stage is superconsistent, meaning it does not affect the efficiency of the effect estimate in the second stage. In a large sample, the estimate and standard error of the (fuzzy) RD in the second stage will therefore be unaffected by the uncertainty induced by the first stage estimation step. In the spirit of Porter and Yu we set the cutoffs at those IST scores which best fit the jump in actual assignment.[9]

[7]We have selected the smaller sample (using a bandwidth range of 18 around the cutoff) based on the formal bandwidth selection procedure of Imbens and Lemieux (2008).

[8]We have also tried to cluster the error terms on IST score as suggested by Lee and Card (2008). In case of IST score clustering (not reported), we get somewhat smaller standard errors.

[9]The algorithm we use selects the IST threshold that maximizes the first stage R-square (without covariates). The algorithm proposed by Porter and Yu (2015) also uses information from the outcome equation. This is more efficient, but only valid under the assumption that

To see whether the imputation procedure is valid in our sample, we perform some additional tests. First, we compare estimated and realized cutoffs for the years where the original cutoffs are known. We find that the cutoffs we estimate are almost identical to the ones we observe in the SGN register (109 versus 109 in 2003; 107 versus 108/109 in 2004). Second, we compare the discontinuities in program assignment (which represent the first stage estimates in a two stage regression discontinuity design) between the years with and without known IST cutoffs. We find, again, that estimated jumps in program assignment are identical, regardless of whether SGN contains information about the IST cutoffs (0.52 in 2003 and 2004; 0.52 in the other years). And third, we run fuzzy RD regressions on samples in which we drop observations near the cutoff. We find that the so-called donut fuzzy RD effect estimates do not differ from the traditional fuzzy RD effect estimates, to which we turn below. Consequently, we do not believe that the estimation of unknown cutoffs has any impact on our results and the corresponding conclusions we draw.

3.3 RD graphs

To illustrate the working of our fuzzy RD setup, we follow common practice and show graphs in which we plot GT program admission, GT pre-treatment cognitive and non-cognitive test scores, and post-treatment GPAs in languages, math, and other subjects, against normalized IST test scores. Discontinuities observed at the IST threshold in GT program admission and GPAs, but not in pre-treatment test scores, would imply that any positive (or negative) GT program effect can be interpreted in a causal way.

Figure 1 shows the relationship between GT assignment and normalized IST scores for all the students in our sample. Each point represents the share of GT assigned students for bin-widths of 4 test score points. Each line represents fitted values of the regression result of equation (1) for different polynomial functions of $f(x)$.[10] It is clear from the figure that students with test scores just above the IST cutoff are much more likely to be assigned to GT education

a program effect exists; an assumption we are not, a priori, prepared to make.

[10]Throughout, we do not show the quadratic split specification (iv) in the graphs to prevent clutter.

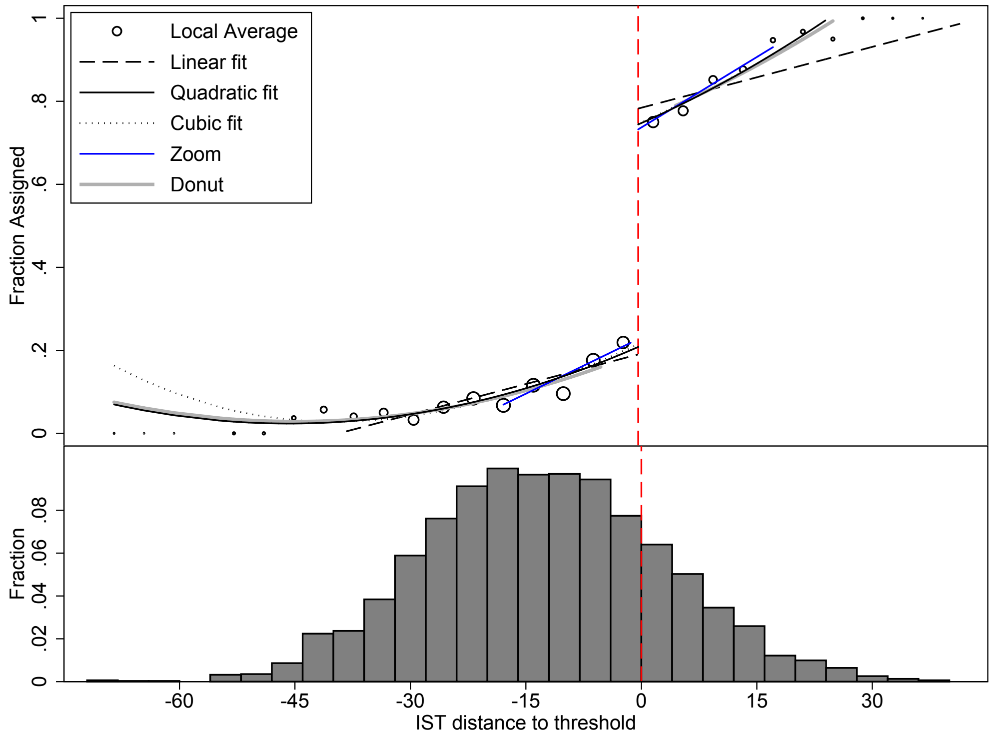

Figure 1. Fuzzy RD first stage effect
Note: The top panel shows fitted values from various parametric first stage regressions of GT program assignment on normalized IST, without covariates. The bottom panel shows the distribution of normalized IST scores. The eligibility cutoff in this picture is normalized to 0.

than those who score just below. In fact, the jump we see in GT program assignment is quite large, about 50 percent-points, and hardly changes when we narrow the sample to those students around the GT admission threshold, or look at more or less restrictive functional forms of equation (1), as discussed above. Also, the share above the cutoff is substantial, roughly 25 percent, so there seem to be a sufficient number of students on both sides of the cutoff score (Matthews et al., 2012). Both features suggest the current design has power to detect meaningful effects, without suffering from weak instrument problems.

The graphical illustration of the first stage specifications is supported by the results in Table 2. Our baseline specification, presented in column 1 shows a first stage of 0.53 (s.e. 0.03), which is highly significant and relevant (F-stat=259.2). Removing the control variables does not change anything (column 2), while restricting the function to a linear shape (column 3) gives a slightly larger jump. This latter specification does not seem to fit the data as well as the other specifications, however, that are presented in columns 4 - 7. All first stage coefficients are close to 0.53, albeit with different degrees of precision. We choose the quadratic function as our baseline specification because it seems to have an appropriate level of flexibility, while preserving power.

Figure 1 also shows the distribution of the normalized IST scores. In a traditional fuzzy RD analysis visible bunching just behind the eligibility cutoff is a potential sign of manipulation of the running variable, which can be a major problem in evaluating gifted policy (Bui et al., 2014; Card and Giuliano, 2014). In our setting, however, manipulation is unlikely given that the IST test is administered by an external organization (CBO) that does not know what cutoff the school will use in a given year, nor do the tested pupils know. Indeed, if we look at the distribution of IST scores, there is no indication of bunching at any position in the distribution, let alone the cutoffs. A McCrary (2008) test confirms that there is no statistically discernible manipulation around the eligibility cutoff (log-difference = -0.09, p-value = 0.447).

We might still worry, though, that the threshold is endogenously set for some years, which would threaten the research design. Therefore, we present Figure 2 to show that there is no apparent jump in covariates FES and CITO, the two strongest predictors of outcomes that we have. In addition, Table 2 shows some additional balancing tests with respect to age and gender (columns 8 - 11). Together, these results suggest that the school sets the cutoff more or less independent of students' observable characteristics (p-value=0.37).

4 Main results

The objective of the fuzzy regression discontinuity analysis is to estimate a local average treatment effect (LATE) that differentiates students who enroll in GT from students who do not, but are otherwise equivalent. In this section

Table 2. First stage estimates and balancing tests

	GT program status							Balancing			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Baseline Quadratic	No controls	Linear	Cubic	Quadratic Split	Zoom	Donut	Male	Age	FES	CITO
Z	0.53	0.53	0.58	0.52	0.50	0.50	0.55	-0.03	-0.02	0.07	0.09
	(0.03)***	(0.03)***	(0.02)***	(0.03)***	(0.04)***	(0.04)***	(0.04)***	(0.04)	(0.04)	(0.08)	(0.05)
Controls	✓		✓	✓	✓	✓	✓				
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
IST	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$\overline{y}$			0.25			0.34	0.21	0.54	12.16	0.00	0.00
$sd\,(y)$			0.43			0.47	0.41	0.50	0.48	1.00	0.87
p-val	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.37		
F-stat	259.2	267.2	456.7	248.5	155.6	157.8	196.7		4.26		
R^2	0.49	0.49	0.49	0.49	0.49	0.45	0.50	0.02	0.04	0.01	0.09
N	3127	3127	3127	3127	3127	1883	2649	3127	3127	3127	3127

Note: Columns 1 - 7 present regressions of GT program status on Z, controlling for normalized IST either quadratically (1,2), linearly (3), cubically (4), quadratically on both sides (5), linear on both sides zoomed into ± 18 normalized IST range (6), or quadratically with ± 4 normalized IST range removed (7). Columns 8 - 11 present separate regressions of the controls age, gender, and (standardized) FES and CITO respectively, on Z and a quadratic function of normalized IST. Cohort dummies are always included. Class clustered (robust) standard errors in parentheses in columns 1 - 7 (8 - 11). */**/*** denote significance at a 10/5/1 percent confidence level. The reported p-value in columns 1 - 7 (8 - 11) comes from an F-test testing the (joint) significance of Z.

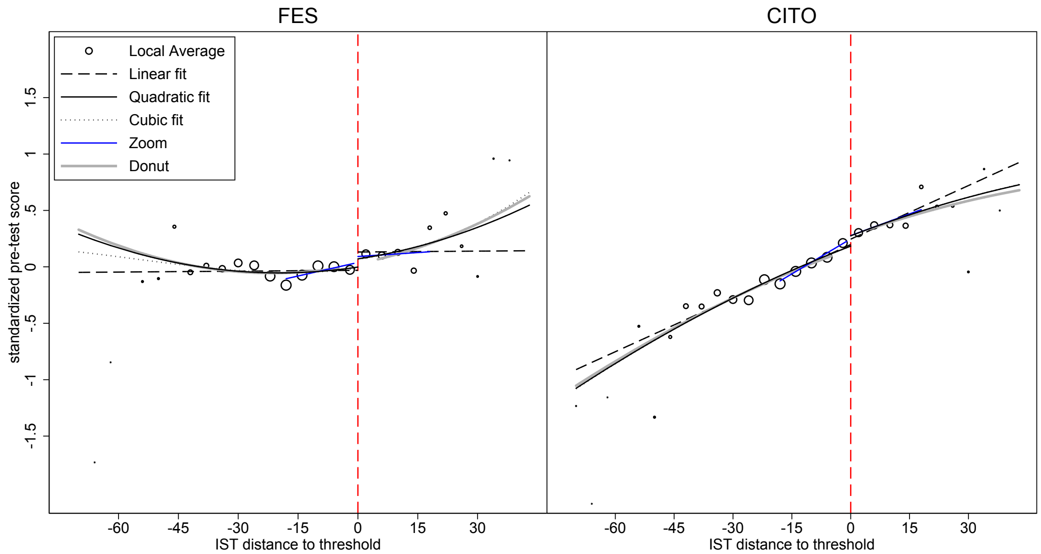

Figure 2. No discontinuities in predictors FES and CITO
Note: The left and right panels show fitted values from various parametric regressions of (standardized) outcome predictors FES and CITO respectively on normalized IST, without covariates. The eligibility cutoff in this picture is 0.

we will consider basic outcomes such as GPAs in math, language, and other school subjects, subject choices in grades 5-6, and a split by gender. Finally, for the older cohorts, we also look at field of study choices in university and the associated (starting) salary. In section 5 we consider potential mechanisms that may have led to these effects.

4.1 Results in academic secondary education

Figure 3 shows the reduced-form relationships between IST and test scores in math, language, and other subjects, averaged over grades 2 to 6. The graphs plot the local average outcomes for each IST-bin, and the fitted relationship from various (local) regression specifications discussed in Section 3.2. There is a steadily upward-sloping relationship between IST and test scores, with clear discontinuities at the entry thresholds for the GT program that are suggestive of positive program effects on student achievement.

Table 3 contains fuzzy RD estimates of the GT program impact on a variety

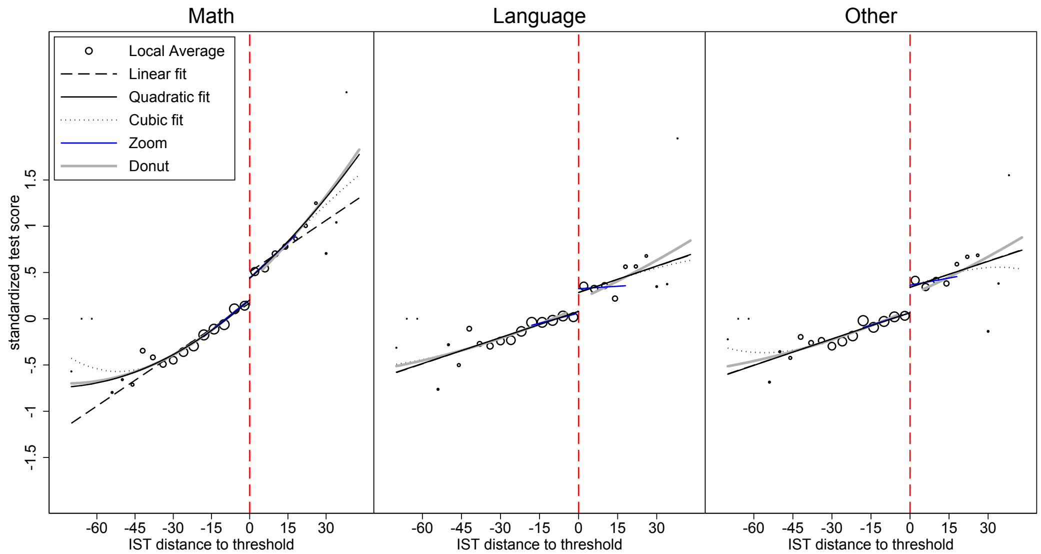

Figure 3. Reduced form effects for math, language, and other subjects
Note: The left, middle, and right panels show fitted values from various parametric regressions of (standardized) GPA for math, language, and other subjects respectively on normalized IST, without covariates. The eligibility cutoff in this picture is 0.

of academic outcomes for our baseline specification. The first two columns are meant to show that there is no sample selection with respect to a tested student being matched to his administrative record or having retained a class. The insignificant differences we find mean that the other estimates presented in columns 3 to 8 can arguably be interpreted in a causal way, and do not reflect selection.

We find that students do better in math, language, and other subjects once they are assigned to the GT program. The estimated effects show that GT program participation raises cumulative grade point averages in math, language, and other subjects with 0.38SD (s.e. 0.14), 0.30SD (s.e. 0.14), and 0.44SD (s.e. 0.15), respectively. Overall, these program effects are substantial and comparable to what Card and Giuliano (2014) find for high achievers, but not for gifted students. One possible explanation for this is that the gifted students in our study are comparable in that they were high achievers too (recall that SGN only admits students with high CITO scores).

In the last two years of academic secondary school, most gifted students stop participating in projects. Instead, they may opt for a more challenging curriculum. In columns 6 to 8 we look at student's subject choices. The GT program does not necessarily lead to a larger number of exam subjects (#sub); the GT program effect estimate in column (6) is positive but does not enter significantly. The GT program does, however, affect the way in which students compose their curriculum in their final years of school. We find that the curriculum of GT students contains more science related exam subjects; the estimates indicate that number of science subjects increases by 0.76 (column 7, s.e. 0.27), and the likelihood that a student chooses advanced math increases by 0.18 (column 8, s.e. 0.08). These results together suggest that the GT program has increased academic performance across the board: students obtain higher grades in all subjects, and choose subjects with higher levels of abstraction. Note that the control variables are all significant predictors of outcomes, most notably the CITO test scores. Balance with respect to these variables, therefore, strengthens the credibility of the design.

Table A1 on page 34 reports the GT program effect estimates for math using three other specifications: quadratic-split estimation on the full sample; linear-split estimation on the smaller sample; and quadratic estimation on the donut sample. We find that the estimates are quite stable across the three different specifications and comparable to those GT program effect estimate obtained using our baseline specification.[11]

Table 4 reports the estimates of the impact of the GT program for boys and girls separately. Although this sample split comes at the cost of precision, two interesting patterns emerge. First, we find traditional gender differences in GPA gains. If we look at columns 3 to 5, it seems that the positive program effects for math are mostly driven by boys (0.53SD, s.e. 0.18), while the positive program effects for language are mostly driven by girls (0.36SD, s.e. 0.20). The program effects we find for other subjects are most comparable to the positive program effects we find for math for boys and language for girls,

[11]Figure A1 also plots GT program effect estimates using linear-split estimation for increasing bandwidth samples. We see that the corresponding estimates are insensitive to bandwidth choice and, again, very similar to GT program effect estimate we get when we regress our baseline specification on the full sample.

Table 3. Estimated effects of the GT program on academic secondary school outcomes

	Sample selection		GPA			Grades 5 - 6		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Matched	Retention	Math	Language	Other	#sub	#sc	Adv. Math
GT program	-0.01	-0.06	0.38	0.30	0.44	0.22	0.76	0.18
	(0.02)	(0.05)	(0.14)***	(0.14)**	(0.15)***	(0.24)	(0.27)***	(0.08)**
Male	-0.01	0.10	-0.02	-0.44	-0.35	-0.25	0.56	0.20
	(0.01)	(0.02)***	(0.03)	(0.04)***	(0.04)***	(0.08)***	(0.09)***	(0.03)***
Age	-0.00	0.02	-0.09	-0.06	0.01	-0.11	-0.30	-0.08
	(0.00)	-0.02	(0.03)***	(0.03)*	(0.03)	-0.08	(0.08)***	(0.02)***
std FES	0.00	-0.00	0.04	0.04	0.06	0.04	0.10	0.02
	(0.00)	(0.01)	(0.02)**	(0.02)***	(0.01)***	(0.04)	(0.04)**	-0.01
std CITO	0.02	-0.11	0.19	0.28	0.24	0.17	0.13	0.04
	(0.01)***	(0.01)***	(0.02)***	(0.02)***	(0.02)***	(0.06)***	(0.05)**	(0.01)***
Cohort	✓	✓	✓	✓	✓	✓	✓	✓
Quadratic IST	✓	✓	✓	✓	✓	✓	✓	✓
$\overline{y}$	0.98	0.28	0.00	0.00	0.00	12.60	2.53	0.63
$sd\,(y)$	0.15	0.44	1.00	1.00	1.00	1.56	1.65	0.48
p-value	0.549	0.244	0.007	0.040	0.004	0.348	0.005	0.017
FS F-stat	259.2	262.4	262.4	262.4	262.4	133.2	133.2	133.2
R^2	0.06	0.09	0.21	0.16	0.15	0.02	0.10	0.12
N	3127	3056	3056	3056	3056	1771	1771	1771

Note: Each column represents an instrumental variable regression where GT status is instrumented by Z, and controls for gender, age, FES, CITO, cohort (dummies), and a quadratic function of normalized IST are included. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level. The reported p-value comes from a t-test of the GT program coefficient. The FS F-stat is the first stage F-statistic testing the significance of Z.

probably reflecting traditional gender differences in school curriculum, with more science-related subjects for boys and language-related fields for girls. Second, we find that the program encourages girls to opt for a more math and science intensive curriculum. If we look at columns 7 and 8, we see that the program effects on subject choice mostly come from girls who take significantly more math and science subjects.

4.2 Results in university

In Table 5 we test whether the positive program effects are longer lasting and carry over to student choices in university. In column 1 we find there is no selectivity with respect to the records that we were able to match to the university registry (87 percent). It further implies that the decision to enroll in university is not affected by GT education. In column 2 we turn to field of study and find that the inclination to choose a science field increases with 14 percent-points, which is not statistically significant but nonetheless consistent with the more abstract subject choices students make in high school in grades 5 and 6. In column 3 we show results for switching fields in the first year. If gifted students make more informed field of study choices because they had more opportunities to learn different topics and visit university, we expect less switching among gifted students. This is not what we see, at least not for boys. Perhaps they find it difficult to commit to one particular field of study. Most interestingly, in column 4 we find that the GT program increases expected wages by 9 percent (s.e 0.04). The exact mechanism behind this result can be found in Table A2 on page 35. There, we show that boys are less likely to choose the least rewarding field of study (Arts and Behavioral sciences) whereas girls are more likely to choose the most rewarding field of study (Medicine and Healthcare). Hence, it seems that gifted students, once exposed to GT education, become more ambitious in their subject choice at university, at least from an expected wage perspective.

Table 4. Estimated effects of the GT program on academic secondary school outcomes by gender

	Sample selection		GPA			Grades 5 - 6		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Matched	Retention	Math	Language	Other	#sub	#sc	Adv.Math
Baseline ($N = 3127$)								
GT program	-0.01	-0.06	0.38	0.30	0.44	0.22	0.76	0.18
	(0.02)	(0.05)	(0.14)***	(0.14)**	(0.15)***	(0.24)	(0.27)***	(0.08)**
Female ($N = 1434$)								
GT program	0.02	-0.05	0.23	0.36	0.35	0.39	1.15	0.24
	(0.02)	(0.07)	(0.21)	(0.20)*	(0.21)*	(0.42)	(0.41)***	(0.12)*
Male ($N = 1693$)								
GT program	-0.04	-0.10	0.53	0.28	0.53	0.05	0.50	0.16
	(0.03)	(0.08)	(0.18)***	(0.19)	(0.19)***	(0.30)	(0.38)	(0.10)

Note: Each estimate comes from a separate instrumental variable regression where GT status is instrumented by Z, and controls for gender, age, FES, CITO, cohort (dummies), and a quadratic function of normalized IST are included. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level.

Table 5. Estimated effects of the GT program on long term outcomes by gender

	Matched	Sc. field	FY switch	$log\,(w)$
	(1)	(2)	(3)	(4)
Baseline ($N = 2438$)				
GT program	-0.02	0.14	0.07	0.09
	(0.06)	(0.09)	(0.05)	(0.04)**
Female ($N = 1135$)				
GT program	0.01	0.11	-0.02	0.06
	(0.06)	(0.09)	(0.05)	(0.05)
Male ($N = 1303$)				
GT program	-0.02	0.18	0.17	0.12
	(0.09)	(0.14)	(0.08)**	(0.05)**

Note: Each estimate comes from a separate instrumental variable regression where GT status is instrumented by Z, and controls for gender, age, FES, CITO, cohort (dummies), and a quadratic function of normalized IST are included. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level.

5 Program use and mechanisms

To better understand where GT program benefits come from, we conducted an online student survey. In this section we describe the survey, discuss program use, and analyze various channels that could potentially lead to a positive program impact.

5.1 The online survey

In May 2014, we conducted an online survey in which we asked current students at SGN questions about their behavior and attitudes regarding school. To enhance response, the survey was made smartphone compatible, leading to 452 responses of students from grades 3 to 6, of which 100 are GT participants. The total response rate was just below 50 percent (452/911) and, important for our purposes, not selective with respect to GT status at the margin (p-value = 0.73).

We asked several questions to gifted students concerning program use. Table 6 summarizes the answers (complemented with SGN register information on project type). We see that program participation rates are high. About 95 percent of all students assigned to the GT program actively participate and work on projects. We also see that GT students spend, on average, about 3.5 hours per week on their project. Girls more frequently skip language classes than boys, and the reverse holds for math. Most students (roughly 66 percent) report they skip classes they consider easy. When we look at project choice, we see that girls are more likely to choose language related projects, whereas boys more frequently choose math/science related projects.

We have also asked how students spend their project time. While most students report to devote all (37 percent), or almost all (38 percent), of their project time to the project, about 25 percent of the students indicate that they regularly use their project time for other purposes (homework, test preparation, or some form of leisure activity). Among all students that squander project time (including those who report to do this only once in a while), test preparation appears the primary reason for misuse (not reported).

Table 6. Usage of the GT program

	Female		Male			
	Mean	s.d.	Mean	s.d.	Difference	*p*-value
Program use†	0.94	0.24	0.96	0.18	0.03	0.18
Substituted hours	3.41	1.81	3.62	2.15	0.21	0.61
all to project time	0.32	0.47	0.39	0.49	0.07	0.49
almost all to project time	0.44	0.50	0.35	0.48	-0.09	0.38
regularly for other use	0.24	0.43	0.26	0.44	0.02	0.81
Substituted subject						
Math	0.50	0.51	0.59	0.50	0.09	0.39
Language	0.98	0.14	0.88	0.33	-0.10***	0.00
Other	0.96	0.20	0.94	0.24	-0.02	0.34
Substituted subject, why?						
Easy	0.68	0.47	0.65	0.48	-0.02	0.80
Dislike	0.26	0.45	0.26	0.44	-0.01	0.94
Other	0.06	0.24	0.09	0.29	0.03	0.56
Substituted to project type†						
Math	0.14	0.34	0.45	0.50	0.31***	0.00
Language	0.30	0.46	0.20	0.40	-0.10***	0.01
Other	0.57	0.50	0.36	0.48	-0.21***	0.00
N	34		66			

Note: Survey sample of assigned students, excluding 10 cases (9 percent) due to item non-response. All variables are binary except for *hours*, which runs from 0 to 12 hours per week. P-values from independent mean comparison t-tests. */**/*** denote significance at a 10/5/1 percent confidence level. † Registry sample of 512 treated students from cohorts '98-'05.

Interestingly, we also see that the gender differences in project subjects coincide with the gender differences in project effect sizes: boys choose math topics and improve mostly there, while girls more frequently choose language projects and improve there. This supports the idea that GT students learn from GT education and gain academic skills. Researchers in gifted education have reported that approximately 40 to 50 percent of traditional classroom content and skill instruction at a given grade level is redundant for gifted students (e.g. Reis and Purcell, 1993). The gifted students seem to realize this and skip the easy classes to work on project subjects that, on average, seem related to the classes they missed. Hence, GT students may have improved their grades because of the skills they acquired from working on their projects.

5.2 Mechanisms that benefit GT students

Several mechanisms predict that students improve their grades in response to the GT program; among these are increased student effort, improved academic skills (and beliefs thereof), enhanced psychological traits, and differential treatment of parents and teachers. In our survey we ask all students about these features; as a result, we can identify causal channels, as in the previous section, by comparing the (standardized) answers of students who were tested just below and just above the assignment cutoff. Table 7 shows fuzzy RD regression results for different (standardized) measures using three different specifications (baseline, linear, and linear split).

The first mechanism runs through effort. If GT education demands students to work harder we expect that GT students get higher grades. We measure effort by asking students how much time per week they spend on homework and test preparation. The estimates indicate that GT students do not spend more time on homework (and test preparation) than non-GT students; on the contrary, we find that students who spend less time in class also report to spend less time learning the core curriculum.

The second mechanism asserts that students accumulate more academic skills because of GT education. We have already presented some evidence in favor of academic skill gains: GT students opting for a more academic and demanding school curriculum, and parallel gender patterns in program effect sizes and project subject choices (including the unobserved skills they likely develop there). Here we test whether GT education helps students to become more independent and responsible, which we view as academic skills. In theory, we expect that an individualized GT program with features such as class skipping, project development, and project management, builds responsibility and independence. In the survey we ask students whether they see themselves as a responsible and independent student on a 0-100 (slider) scale. It seems that the GT program enhances responsibility, but leaves independence unaf-

Table 7. Estimated mechanisms of the GT program

Mechanism variable (standardized)	GT coefficient given specification		
	Baseline (1)	Linear (2)	Linear split (3)
Effort			
Time learning	-0.59(0.27)**	-0.49(0.21)**	-0.59(0.25)**
Skills			
Responsibility	0.32(0.40)	0.27(0.32)	0.26(0.35)
Independence	-0.02(0.40)	-0.07(0.32)	-0.08(0.37)
Psychological traits			
Motivation	-0.49(0.35)	-0.44(0.28)	-0.54(0.33)*
Confidence	-0.31(0.33)	-0.15(0.30)	-0.26(0.32)
Academic esteem	0.53(0.26)**	0.40(0.22)*	0.46(0.24)*
Parents & Teachers			
Parent stimulus	0.29(0.31)	0.19(0.25)	0.23(0.27)
Teacher stimulus	0.29(0.27)	0.06(0.22)	0.22(0.25)
Exam scores†			
Math	0.50(0.20)*	0.35(0.14)*	0.38(0.18)*
Language	0.31(0.19)	0.14(0.14)	0.23(0.17)
Other	0.75(0.18)***	0.49(0.14)***	0.67(0.16)***
N	452	452	452

Note: Each estimate comes from a separate instrumental variable regression where GT status is instrumented by Z, and controls for gender, age, FES, CITO and cohort (dummies) are included. Normalized IST is controlled for quadratically (1), linearly (2), or linearly on both sides (3). Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level. The mechanism variables are all standardized and constructed by taking the mean of two or more (standardized) items with a 0-100 (slider) scale. The average Cronbach's alpha of these measures is 0.65. The skill and academic esteem variables are single item. † Registry sample of 1643 students from cohorts '98-'07.

fected. The estimates are stable between specifications, but come with large standard errors, so the results for responsibility serve, at most, as suggestive evidence of academic skill accumulation.

The third mechanism links the GT program to a broader set of psychological traits and characteristics. Many psychological factors may contribute to student achievement, with intrinsic motivation and self-confidence probably the most important ones. In our survey, we measure intrinsic motivation using six slider questions asking students whether they are eager to learn, eager to go to school, easily bored, keen on high grades, and receptive learners at school.[12] There is no clear evidence that GT education fosters motivation. We find insignificant estimates that systematically point in the opposite direction. Similarly, there is no clear evidence that GT education builds self-confidence either, as captured by the self-confidence slider measure.

Another psychological factor that could lead to better grades is improved academic esteem (e.g. Bong and Skaalvik, 2003). If students are told that they are gifted, they may hold stronger beliefs about their academic abilities and improve their academic performance. As a proxy for academic esteem we have asked students whether they think of themselves as a good learner. Our regressions show positive, significant, and substantial program effects on self-assessed measures of academic esteem, which suggests that labeling students as gifted can be one of the channels at work.

The fourth and last mechanism involves behavioral changes of parents and teachers. If parents and teachers know that some of the children are gifted, they may treat these children differently. Students were asked six questions on whether they were helped, encouraged, or pushed by their parents and teachers. We find some weak evidence in favor of treatment differentials; that is, all the estimates are similarly positive. The standard errors are too large to draw any firm conclusions, however.[13]

[12] We take the average over the six motivation items. The boredom question was phrased negatively, and thus reversed. The reliability of the intrinsic motivation measure is 0.71, as captured by Cronbach's alpha.

[13] One concern related to treatment differentials is that some parents and teachers may push their children and students too much, possibly with adverse effects. When we run separate fuzzy RD regressions for helping and encouraging parents and teachers versus

Another aspect of changed teacher behavior may go through grading practices. To see whether teachers assess gifted children more favorably, we look at nationwide subject-specific exams that students take in their final year (6th grade). These exam grades are externally validated. If teachers assign too high grades to GT students, we should find smaller test score gains in tests that are checked by external teachers who are unaware about the students' giftedness status. When we run our fuzzy RD effect regressions on standardized exam scores for math, language, and other subjects, we find effect estimates that are as large, if not larger than the effect estimates we report in Table 3. Hence, we do not think that GT students do better because of favorable grading practices.

5.3 Mechanisms that hurt non-GT students

Identification of GT program effects must assume that GT education itself does not weaken the performance of those students who are not gifted; otherwise, our fuzzy RD design is contaminated and differences in grades observed at the assignment cutoff can no longer be interpreted as GT program benefits for gifted students. Such contamination may arise through disappointment of being left-out, and through missed classroom spillovers. We discuss each contamination mechanism in turn.

First, students who are left out may feel disappointed and frustrated and as a consequence perform less than they are able to. We find no evidence to support such a disappointment mechanism. All students are informed about their IST and FES test scores. If we assume that non-gifted students near the admission cutoff are also the most disappointed and frustrated students, we should see poorer performance of non-gifted students just below the admission cutoff. Figure 3, presented earlier, shows quite clearly that there is no dip in grades just below the cutoff. In addition, we have asked all left-out students how disappointed they were for not being selected for the program. Over 80 percent of all students report that they were not disappointed. Of those who

pushing parents and teachers, we find program estimates (not reported here) that are all similarly positive.

report feelings of disappointment, only 5 students expressed to be seriously so.

Second, there is a large empirical literature on spillover effects in schools documenting that peers matter, either through their characteristics (e.g. Sacerdote, 2011), or their number (e.g. Krueger and Whitmore, 2001). In our context, the GT program enforces gifted and talented students to spend some time outside their regular classroom reducing possible positive spillover effects to those who stay.[14] If GT peers matter, we expect their impact to increase in their number. In Table A4 on page 37 we regress grades on the fraction of gifted students in the class (columns 3 to 5). The point estimates are all small and not statistically significant. These estimates can be interpreted in a causal way if gifted pupils are randomly assigned to classes. Columns 1 and 2 of the same table suggest this is the case; the classroom fraction of gifted students is unrelated to FES scores, CITO scores, and other student characteristics. Hence, spillover effects are not likely.

6 Summary and discussion

In this paper we use a fuzzy regression discontinuity design to estimate the effect of a gifted education program on math, language, and other subject grades of academic secondary school students in the Netherlands. All the fuzzy RD estimates consistently show positive program effects on student achievement; that is, the marginal student who is barely admitted to the GT program gets at least 0.30 standard deviations higher grades for math, language, and other subjects, follows a more math and science intensive curriculum, and, after secondary school graduation, chooses a more challenging field of study in university with, on average, 9 percent higher returns.

We further use student surveys to better understand why the program works. The program did not encourage students to work harder, boost their self-confidence, or raise their motivation to learn. The program did, however, improve the students' academic esteem. Additionally, students exposed to the

[14]It is also possible that, as the GT students leave, the class become more homogeneous. This might help teachers tailor their teaching to the benefit of the remaining group of non-GT students (e.g. Duflo et al., 2011). Hence, the effect of having gifted peers sometimes leave the class is a priori ambiguous.

program substitute easy courses in favor of complex projects that are, on average, related to the classes they missed; male students work more on math and science related projects, and female students work more on language related projects; and male students experience the largest gains in math grades, and female students in language grades. These findings together are all suggestive of a conventional human capital channel; that is, GT education builds academic skills (and beliefs thereof), which in turn leads to higher grades and average starting salaries.

The GT program we evaluate runs in comparable form at many prestigious academic secondary schools in the Netherlands. An important question is whether the benefits of such a program outweigh its costs. The GT program we look at is a low-cost program. SGN spends about €100k on GT education every year, out of a total budget of €10m. Given the number of gifted students, this would translate into program costs of about €300 per gifted student per year. The GT program is also a beneficial program. Given that GT education increased the average starting salary that corresponds to the field of study choice, rather than the actual starting salary, any comparison between the program costs and benefits can only be speculative. Nonetheless, our most conservative calculations suggest that the labor market benefits of GT education are far greater than its costs.

References

Berkhout, E., Prins, J., and van der Werf, S. (2013). Studie & Werk. Technical report, Stichting Economisch Onderzoek.

Bhatt, R. (2011). A review of gifted and talented education in the united states. *Education Finance and Policy*, 6(4):557–582.

Bhatt, R. (2012). The impacts of gifted and talented education. Technical report, Georgia State University.

Bong, M. and Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15(1):1–40.

Bui, S. A., Craig, S. G., and Imberman, S. A. (2014). Is gifted education a bright idea? Assessing the impact of gifted and talented programs on students. *American Economic Journal: Economic Policy*, 6(3):30 – 62.

Card, D. and Giuliano, L. (2014). Does gifted education work? For which students? Technical Report 20453, National Bureau of Economic Research.

Davis, B., Engberg, J., Epple, D. N., Sieg, H., and Zimmer, R. (2010). Evaluating the gifted program of an urban school district using a modified regression discontinuity design. Working Paper 16414, National Bureau of Economic Research.

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 – 1774.

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

Krueger, A. B. and Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *Economic Journal*, 111(468):1–28.

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

Leuven, E., Oosterbeek, H., and van der Klaauw, B. (2010). The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association*, 8(6):1243 – 1265.

Matthews, M. S., Peters, S. J., and Housand, A. M. (2012). Regression discontinuity design in gifted and talented education research. *Gifted Child Quarterly*, 56(2):105 – 112.

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

Porter, J. and Yu, P. (2015). Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics*, 189(1):132–147.

Reis, S. M. and Purcell, J. H. (1993). An analysis of content elimination and strategies used by elementary classroom teachers in the curriculum compacting process. *Journal for the Education of the Gifted*, 16(2):147–170.

Renzulli, J. S. (1977). *The enrichment triad model: A guide for developing defensible programs for the gifted and talented.* Creative Learning Press Mansfield Center, CT.

Renzulli, J. S. (1986). *The Three Ring Conception of Giftedness: A Developmental Model for Creative Productivity.*, volume Conceptions of giftedness, pages 53–92. Cambridge University Press, New York.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In Hanushek, E., Machin, S., and Woessmann, L., editors, *Handbook of the Economics of Education*, volume 3, chapter 3, pages 249 – 277. Elsevier Science Publishers B.V.

Appendix A Sensitivity analysis

Table A1 on the next page.

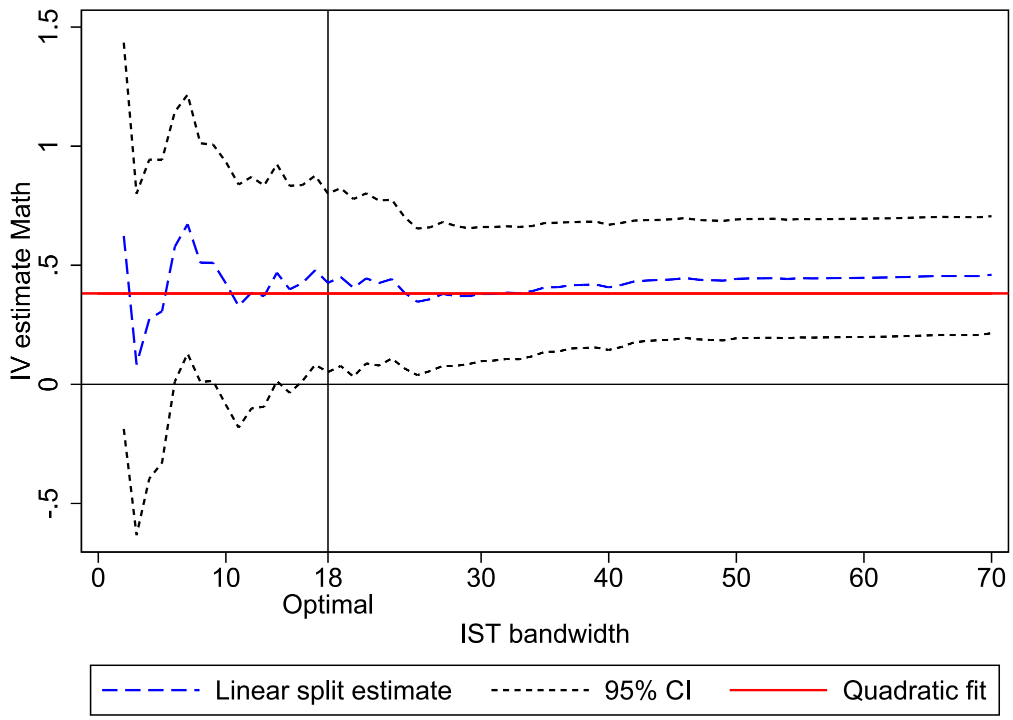

Figure A1. Estimated math effect from linear split regression with varying bandwidth

Table A1. Estimated effects of the GT program on various high-school outcomes using alternative specifications

	Sample selection		GPA			Grades 5 - 6		
	(1)	(2)	(3)	(4)	(5)	(7)	(8)	(9)
	Matched	Retention	Math	Languages	Other	#sub	#sc	Ad. Math
Baseline								
GT program	-0.01	-0.06	0.38	0.30	0.44	0.22	0.76	0.18
	(0.02)	(0.05)	(0.14)***	(0.14)**	(0.15)***	(0.24)	(0.27)***	(0.08)**
Quadratic Split								
GT program	-0.03	-0.08	0.32	0.33	0.40	0.04	0.89	0.17
	(0.02)	(0.07)	(0.20)	(0.19)*	(0.21)*	(0.36)	(0.38)**	(0.11)
Zoom								
GT program	-0.01	-0.09	0.43	0.44	0.55	-0.14	0.71	0.17
	(0.02)	(0.07)	(0.19)**	(0.19)**	(0.20)***	(0.36)	(0.37)*	(0.11)
Donut								
GT program	-0.00	-0.07	0.27	0.07	0.21	0.28	0.62	0.11
	(0.02)	(0.07)	(0.19)	(0.20)	(0.22)	(0.30)	(0.32)**	(0.09)

Note: Each estimate comes from a separate instrumental variable regression where GT status is instrumented by Z, and controls for gender, age, FES, CITO, and cohort (dummies) are included. The baseline model includes a quadratic function of normalized IST. The split model includes separate quadratic functions of normalized IST on both sides. The zoom model includes separate linear functions of normalized IST on both sides zoomed into ± 18 normalized IST range. The donut model includes a quadratic function of normalized IST, with ± 4 normalized IST range removed. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level.

Appendix B HE field of study choice

Table A2. Results on HE field of study choice

			Effect of *GT program*		
Chosen Field of study	Wage	Share	Total	Male	Female
Earth and environment	€ 2392	0.06	0.01	0.05	-0.01
			(0.05)	(0.08)	(0.05)
Economics and business	€ 2752	0.14	-0.04	-0.04	-0.05
			(0.06)	(0.10)	(0.06)
Science and informatics	€ 2730	0.08	0.08	0.11	0.06
			(0.06)	(0.11)	(0.05)
Behavioral and societal sciences	€ 1969	0.09	-0.05	-0.10	0.00
			(0.05)	(0.05)**	(0.08)
Healthcare	€ 2974	0.17	0.08	0.00	0.17
			(0.07)	(0.10)	(0.10)*
The Arts	€ 1909	0.09	-0.11	-0.20	-0.03
			(0.06)*	(0.10)**	(0.07)
Educational and pedagogy	€ 2045	0.03	-0.05	-0.03	-0.07
			(0.03)	(0.03)	(0.05)
Law and governance	€ 2544	0.12	0.04	0.10	-0.02
			(0.06)	(0.08)	(0.09)
Language and communications	€ 2115	0.09	-0.01	0.07	-0.11
			(0.05)	(0.07)	(0.08)
Technical	€ 2520	0.12	0.04	0.02	0.06
			(0.07)	(0.12)	(0.07)

Note: All effect estimates come form separate instrumental variable regressions of a field of study dummy on GT status, instrumented by Z, and controls for gender, age, FES, CITO, cohort (dummies), and a quadratic function of normalized IST included. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level.

Appendix C Contamination

Table A3. Disappointment not admitted to the *GT program*

	Female		Male			
	Mean	s.d.	Mean	s.d.	Difference	p-value
Disappointed	0.13	0.03	0.25	0.04	0.11	0.02**
N	180		118			

Note: Sample of control students, excluding 44 cases (13 percent) due to item non-response. Disappointment is a binary indicating that a student is either 'Yes, a little bit disappointed' (48 cases) or 'Yes, very much disappointment' (5 cases). P-values from independent mean comparison t-tests. */**/*** denote significance at a 10/5/1 percent confidence level.

Table A4. Estimated spillover effects of fraction GT among students below the cutoff

	Balancing		GPA		
	Frac.GT in class	Av.$CITO$ in class	Math	Language	Other
	(1)	(2)	(3)	(4)	(5)
Frac. GT			-0.06	-0.09	0.09
			(0.24)	(0.18)	(0.21)
Male	-0.00	-0.00	-0.03	-0.42***	-0.36***
	(0.00)	(0.01)	(0.04)	(0.04)	(0.05)
Age	-0.00	0.00	-0.13**	-0.11**	-0.05
	(0.00)	(0.01)	(0.04)	(0.04)	(0.04)
std FES	0.00	-0.00	0.05*	0.05*	0.07***
	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)
std CITO	-0.00	-0.00	0.17***	0.27***	0.23***
	(0.00)	(0.00)	(0.02)	(0.03)	(0.03)
Cohort	✓	✓	✓	✓	✓
Quadratic IST	✓	✓	✓	✓	✓
$\overline{y}$	0.24	0.00	-0.18	-0.10	-0.11
$sd\,(y)$	0.10	0.16	0.91	0.97	0.95
p-value	0.34	0.44	0.80	0.61	0.67
N	2285	2285	2246	2246	2246

Note: Sample of students before the cut-off. Each column comes from a separate regression with controls for gender, age, FES, CITO, cohort (dummies), and a quadratic function of normalized IST included. The p-value in columns 1 and 2 are from an F-test on the joint significance of the controls. The p-value in columns 3 - 5 come from a t-test on the coefficient of *Fraq. GT*. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level.