I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10836

Can Gifted and Talented Education Raise the Academic Achievement of All High-Achieving Students?

Adam Booij
Ferry Haan
Erik Plug

JUNE 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10836

Can Gifted and Talented Education Raise the Academic Achievement of All High-Achieving Students?

Adam Booij
University of Amsterdam and Tinbergen Institute

Ferry Haan
University of Amsterdam

Erik Plug
University of Amsterdam, Tinbergen Institute, IZA and UCLS

JUNE 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Can Gifted and Talented Education Raise the Academic Achievement of All High-Achieving Students?*

We conduct a study under 2,400 third grade students at three large secondary comprehensive schools to evaluate a gifted and talented (GT) program with selective program admission based on past achievement. We construct three complementary estimates of the program's impact on student achievement. First, we use the fragmented GT program implementation (in different tracks at different schools) to get difference-in-differences (DD) estimates for all students above the admission cutoff. Second, we use the GT admission rule to get regression discontinuity (RD) estimates for students near the admission cutoff. And third, we combine the DD and RD designs to estimate how the program's impact varies with past achievement. We find that all participating students do better because of the GT program. Students near the admission cutoff experience a 0.2 standard deviation gain in their grade point average. Students further away from the admission cutoff experience larger gains.

JEL Classification: I22, I28

Keywords: gifted and talented education, enrichment program, secondary education, difference-in-differences, regression discontinuity designs

Corresponding author:
Adam Booij
Faculty of Economics and Business
University of Amsterdam
Roetersstraat 11
1018 WB Amsterdam
The Netherlands
E-mail: A.S.Booij@uva.nl

* The authors thank all the participating schools for their cooperation in all stages of this research project. Financial support from the Netherlands Initiative for Education Research (NRO 411-12-637) is gratefully acknowledged.

1 Introduction

Many schools provide gifted and talented (GT) programs, with program admission based on test scores or past achievement (Bhatt, 2011). In recent years, researchers began to exploit these selective admission rules in a regression discontinuity (RD) framework to identify the causal relationship between GT education programs and student achievement. While the resulting RD estimates are by design informative of the program's impact for marginal students (those who are near the admission cutoff), the same estimates are not so informative for potentially weaker and stronger students (those who are further away from the admission cutoff). Such information, however, is important to school administrators and social scientist. With credible program estimates for a broader group of students, schools could more efficiently and justly allocate resources to GT programs and researchers could better understand how to develop more effective GT programs.

In this paper we provide causal effect estimates of a GT program (with selective admission) on student achievement for a broader group of students in secondary education. The GT program we consider is an individualized pull-out program, based on ideas of (Renzulli, 1977), in which selected students can decide to replace classroom teaching (with teacher consent) for in-school time to work on self-selected projects. The GT program takes a full school year, which consists of four terms: in the first term students are selected; in the second and third term students choose and work on their projects; and in the final term students present their projects in a competition for the best project award. The GT program is selective and offered to those students who scored the highest grade point average (GPA) in the first term of the school year. About 18 percent of students pass the admission threshold and qualify for the program. We refer to these students as high achievers. About 60 percent of them decide to enter the program. We refer to these students as users. Booij et al. (2016) evaluate the impact of a comparable GT program on academic performance of the smartest students in a prestigious academic secondary school in the Netherlands and find, among other things, that those students who just made it into the GT program obtain much higher grades and follow a more science intensive curriculum. The question of how these positive program impacts generalize to other schools and students is at the center of this paper.

We conduct a study to evaluate the impact of a GT program on student achievement in three secondary schools in the Netherlands. These schools are large compre-

hensive schools that offer multiple tracks, including academic secondary education (VWO) and general secondary education (HAVO). Within each track, all students follow the same program during the first three school years. At the end of third school year, students choose their field of specialization: science, health, social sciences, or humanities. These fields strongly correlate with the choice of major in tertiary education later in life. In 2012, we introduced GT education to third-grade students in one of the two tracks within schools. Over a period of 3 years, some 2,400 third graders were involved in the study.

To measure the impact of the GT program on student achievement, we construct three different estimates. First, we use the fragmented GT program implementation (in different tracks at different schools) to get difference-in-differences (DD) estimates for all students above the admission cutoff. Second, we use the GT admission rule to get regression discontinuity (RD) estimates for students near the admission cutoff. And third, we combine the DD and RD strategies to allow for impact heterogeneity and estimate whether the program's impact is different for students with a different first term GPA ranking. Student achievement outcomes are defined in terms of grade point averages and field choice measured at the end of the third school year.

We find that the program's impact is positive for all participating students. Students near the admission cutoff experience a 0.2 standard deviation gain in their grade point average. Students further away from the admission cutoff experience even larger gains. In addition, we test for possible adverse program effects among students excluded from the program. If some of these students experience feelings of disappointment for being left out, or miss out on classroom spillovers, we may find positive program effects not because eligible students do better but because non-eligible students do worse. We find no evidence of this.

Our paper contributes to the emerging economics literature on the causal impact of GT education on student skills. The few studies on the topic apply RD designs and thus focus on students near the admission cutoff. Their results are mixed. Some studies find that high-achieving students gain academic skills (Booij et al., 2016; Card and Giuliano, 2016). Other studies find no gains for gifted students (Bui et al., 2014; Card and Giuliano, 2015). With so few program effect estimates that vary so widely, it is insightful to have one more study on the topic using credible methodologies with different data.

Our paper also contributes to the recent econometric literature on the extrap-

olation of RD estimates away from the cutoff. The studies on this topic, however, take very different approaches. We list a few representative examples. Bertanha (2016) obtains RD estimates for a more diverse group of marginal individuals by exploiting variation in cutoffs. Angrist and Rokkanen (2015) use additional control variables that are strongly related to the running variable to turn selective admission into randomized admission. If the estimated impact of the running variable, after adding the controls, gets close to zero, treatment assignment is arguably random, and corresponding RD estimates are informative for all treated individuals. Our paper most closely relates to those RD extrapolation studies that use alternative proxies for the relationship between the running variable and outcomes for treated individuals in the absence of the treatment, including functional form extrapolation based on the estimated relationship between the running variable and outcome (Angrist and Pischke, 2009), and derivative-based extrapolation based on the estimated marginal effect of the running variable on the outcome at the cutoff (Dong and Lewbel, 2015). We add to these latter studies by estimating the relationship between the running variable and outcome using students who are above the cutoff in untreated tracks in treated schools as the counterfactual students.

The remainder of the paper is organized as follows. Section 2 discusses secondary education, the secondary schools in which the GT program takes place, the features of the GT program, and the pilot design. Section 3 describes the empirical strategy and data. Section 4 presents and discusses the main empirical findings. Section 5 summarizes and concludes.

2 Context: Secondary education, GT program and study design

In this study we implement a GT program in three comprehensive secondary education schools in one particular region in the Netherlands. In the following, we provide a brief overview of secondary general education, the GT program, and how the GT program implementation is helpful in estimating the effect of GT education on student achievement.

2.1 Secondary education

Dutch secondary education offers general and vocational education. Within general secondary education, students are tracked on ability into two distinct types: academic secondary (VWO) and general secondary (HAVO). Ability is measured

through national primary exit exams taken in the final year of primary education (CITO test score). The VWO track takes 6 years and prepares students for higher education in research universities. The HAVO track, which is less advanced, takes 5 years and prepares students for education in universities of applied sciences. Within track type, students follow the same program during the first three school years and a field-specific program during the final school years, which they choose in the final term of the third year. Students can choose between science, health, social sciences, and humanities. These field choices correlate strongly with the choice of major in higher education and corresponding earnings. Many schools in secondary general education are organized as comprehensive schools that offer multiple tracks, including the VWO and HAVO tracks. Three such comprehensive schools participated in our study on GT education.

Secondary school students start school at age 12. Students are taught in classes. Within school tracks, students follow the same subjects, including languages, mathematics, history, arts, and sciences. Students are taught, tested, and graded by subject teachers. In large schools with many students and many subject teachers, students in different classes are often taught the same grade subject by different subject teachers.

2.2 The GT program

Inspired by Renzulli's notion that promising students may benefit from an enriched education program with exposure to new content, active application of own skills, and creation of a product (Renzulli, 1977, 1986), we introduced a selective GT program under third grade students in three comprehensive secondary schools. The GT program in question is an individualized pull-out program, where selected students receive the right to trade in classroom lessons (with teacher consent) for project time (spent elsewhere) to work on a project of their own choice. Teachers can deny students the right to trade in classroom lessons, for instance, when student performance is unsatisfactory. The schools provides rooms, computers, and arts and crafts facilities to help students on their projects. Participating students can choose which classroom hours to devote to their project, with a maximum of five hours per week. The GT program spans one school year. A school year is divided into four terms. In the first term of the school year students are selected. Students with a first term grade point average above a pre-determined cutoff are invited to

participate.[1] Students are graded on a 0 to 10 point scale. The cutoff grade is set at 7.5 or 8 depending on the school and year. Students are not informed about the program beforehand, nor about the cutoff value. In the second and third term of the school year students choose and develop a project topic, which can be anything. At the end of the school year, students present their projects to teachers, parents, and fellow students in a competition for the best project award. Students are supervised by their GT mentors throughout the development of the project. Mentors are instructed to let students take the lead in project development and to provide hands-off supervision aimed at having a finished project at the end of the year. Mentors and students meet every two weeks. We should note that participating students follow the same classes (when they are not working on their projects), face the same curriculum, and do the same exams as the other students.[2]

2.3 The GT study design

In order to best estimate the impact of the program on the academic performance of third graders in both tracks,.we assigned the program to third graders in one of the two tracks per school. There exist 6 possible permutations when we stratify the third grade students by school and track (*HHV-HVH-VHH-VVH-VHV-HVV*). Schools could not choose. We assigned students into program and comparison tracks randomly, by flipping a coin in front of the school management.

Before the start of the study, the three schools were informed about the details of the GT program. The third grade team leaders at the selected tracks of the three schools were instructed to coordinate the GT program. Each school formed a GT team consisting of a GT coordinator, GT mentors, and class mentors. The GT coordinator consults GT mentors and GT students about project progress, consults class mentors about student achievement, and schedules GT team meetings. GT students choose their GT mentors as supervisors. GT mentors are instructed to let students take the lead in project development and to provide hands-off supervision aimed at having a finished project at the end of the school year. GT mentors and students should meet every two weeks. The class mentors monitor the GT students

[1] Throughout the paper, we will use baseline GPA in reference to first term GPA.

[2] Booij et al. (2016) evaluate a GT program at a prestigious academic secondary school in Nijmegen. While the GT program in this study is very similar in design, it differs from the program in Nijmegen in a few respects. The selection process is based on past achievement and not on test scores. As such, the school selects high achievers. The program targets students at different ability levels. As such, the program is not elitist. Additionally, the program is exclusively for third graders.

in class. For preparation, representatives of the three GT teams visited another (academic) secondary school, that has much experience with running a comparable GT program. We further provided funds to cover some of the program costs, which include material costs for projects and opportunity costs of regular teaching time allocated to GT mentoring. The program treatment started in the school year 2012-2013 and ran for three consecutive school years. For the whole study period, one of us was present at one of the three schools to inform parents, students, and teachers about the GT program, to monitor the working of the GT program, and to check whether the GT teams and GT students behaved according to the GT program guidelines,.

Notwithstanding our continuous monitoring, some events occurred that may have compromised the setup of our study. First, one school reduced the number of school terms from 4 to 3 in the second year of the experiment. Since the first term GPA would arrive later in the school year, we decided to assign students to the GT program on the basis of GPA obtained in the last term of the year before (second grade). Second, some enthusiastic GT mentors felt that the assignment procedure was too strict and started to invite students with GPA scores under the assignment cutoff. We will therefore analyze the program's impact on student achievement in two ways. First, we provide reduced-form estimates of the impact of program assignment on student achievement. Second, we provide instrumental variable estimates of the impact of the program on student achievement using initial program assignment as an instrumental variable for program use.

3 Empirical Strategy

We apply three complementary strategies to identify the program's impact on student achievement: (i) we use the fragmented GT program implementation (in different tracks at different schools) to get difference-in-differences (DD) estimates for all students above the admission cutoff;[3] (ii) we use the GT admission rule to get regression discontinuity (RD) estimates for students near the admission cutoff; and (iii) we combine the DD and RD designs to estimate how the program's varies with

[3]In our setup, we have randomized at the school-track level. With cluster randomization involving few clusters, randomization into program and comparison tracks does not guarantee that students in program and control tracks are, on average, very similar. The DD analysis accounts for any pre-treatment differences between students in program and control tracks and therefore gives more credible estimates.

baseline achievement.

3.1 Combining DD and RD Strategies

To measure the impact of the GT program for all those students with a baseline GPA above the admission cutoff, we exploit the differentiated introduction of GT education in different tracks at different schools and estimate a standard difference-in-differences regression model on the full sample of students:

$$Y_{isty} = \beta^{DD} T_{st} Z_i + \alpha T_{st} + \gamma Z_i + \delta X_i + \mu_s + \mu_t + \mu_y + u_{isty}, \tag{1}$$

where Y_{isty} is a measure of the academic achievement of student i in track t at school s at year y (defined as the track-specific standardized GPA taken over the school year), T_{st} is a dummy variable indicating whether a student is in a treated track, Z_i is a dummy variable indicating whether a student has a baseline GPA above the admission cutoff, X_i is a vector of exogenous student characteristics including gender, age, and primary school test scores, μ_s is a set of school fixed effects, μ_t is a track fixed effect, and μ_y is a set of year fixed effects. The parameter β^{DD} measures the program's intention-to-treat impact for all program eligible students.

To measure the impact of the GT program for those students with a baseline GPA near the assignment cutoff, we exploit the discrete nature of the assignment rule and run a standard regression-discontinuity model (with a different slope on either side of the cutoff) on the restricted sample of students in treated tracks:

$$Y_{isty} = \beta^{RD} Z_i + \gamma_1 z_i + \gamma_2 Z_i z_i + \delta X_i + \mu_t + \mu_y + u_{isty}, \tag{2}$$

where z_i is the running variable measuring past achievement (defined as the GPA taken over the first term of the school year, and normalized to 0 at the admission cutoff). The track fixed effect μ_t also captures the school fixed effect because (in our setup) only one track per school is treated. The parameter β^{RD} measures the program's intention-to-treat impact for those eligible students near the admission cutoff.

To measure whether the program's impact is different for students near and further away from the admission cutoff, we combine DD and RD strategies and run

the next regression model using the full sample of students:

$$Y_{isty} = \beta_1^{DD/RD} T_{st} Z_i + \beta_2^{DD/RD} T_{st} Z_i z_i + \alpha T_{st} + \gamma z_i + \delta X_i + \mu_s + \mu_t + \mu_y + u_{isty}. \quad (3)$$

This is our preferred specification. The parameter $\beta_1^{DD/RD}$ measures the intention-to-treat effect for students near the admission cutoff. The interacted parameter $\beta_2^{DD/RD}$ measures the extent to which stronger students benefit more from the GT program.

We estimate these three regression models using ordinary least squares (OLS), with standard errors clustered at the level of the classroom-year. There are two points to note here. First, we deviate from the preferred level of clustering, which is at the treatment level. In our setup, we would then work with just 6 clusters and run the risk of underestimating the standard errors. Instead, we opt for classroom-year clusters because in large schools with many classes and subject teachers, we believe that most of the within-school-track-year variation is driven by within-classroom-year variation and not by between-classroom-year variation. Second, we have introduced strategies that provide intention-to-treat estimates. We also have information about actual program use. In Section 4 we will exploit this information. With Z_i as instrumental variable for program use, we estimate the impact of the GT program for those students who actually receive GT education using two-stage least squares (2SLS).

3.2 Data

The three participating schools gave us access to their student administration records on basic demographic characteristics, such as gender and age, primary education exit exam scores (CITO test scores), GT program assignment status, track status, field of specialization and any other school grade obtained from the day of entry until the day of leave. We sample all third year students in the school years 2012/2013, 2013/2014 and 2014/2015.

Table 1. Summary statistics

	Control tracks		Treated tracks			
A: Characteristics	mean	s.d.	mean	s.d.	difference	p-value
Male	0.48	0.50	0.46	0.50	-0.02	0.41
Age	14.41	0.41	14.33	0.40	-0.08	0.00
B: School and track						
SCHOOL 1	0.37	0.48	0.26	0.44	-0.10	0.00
SCHOOL 2	0.25	0.43	0.29	0.46	0.04	0.02
SCHOOL 3	0.38	0.49	0.44	0.50	0.06	0.00
Academic track	0.25	0.43	0.71	0.46	0.46	0.00
C: Pre-test						
raw CITO score	540.37	5.23	542.08	4.92	1.71	0.00
baseline GPA (running variable)	6.80	0.68	6.87	0.72	0.06	0.02
Z (baseline GPA above cutoff)	0.11	0.32	0.18	0.38	0.07	0.00
D: Treatment						
GT program (eligible)	0.00	0.00	0.18	0.38	0.18	0.00
GT program use	0.00	0.00	0.13	0.33	0.13	0.00
E: Outcomes						
GPA	6.49	0.69	6.60	0.74	0.12	0.00
GPA math	6.44	1.06	6.58	1.17	0.13	0.00
GPA language	6.46	0.81	6.62	0.86	0.16	0.00
GPA other	6.51	0.74	6.60	0.73	0.09	0.00
STEM	0.45	0.50	0.50	0.50	0.05	0.02
Number of classes	48		40			
Number of pupils	1,356		1,067			

For these students, we construct several measures of academic achievement: overall GPA, GPA for math, language, and other subjects, and one indicator of choosing an advanced curriculum in the final school years. Within each track, all third grade follow the same curriculum. Grades range from 1 to 10. The overall GPA variable is the mean of all subject grades that appear on the final report card issued in the last term of the year. The math variable is the final score for the standard mathematics subject.,The language variable is the mean of the final subject scores in Dutch, French, German, and English. The other subject variable is the mean of the final subject scores in all other subjects, including geography, history, arts and sciences.

At the end of the third school year, students decide on their field of specialization for their final years of secondary school. There are four fields: the science track (*NT* which stands for nature and technology), the health track (*NH* which stands for nature and health), the social sciences track (*ES* which stands for economics and society), and the humanities track (*CS* which stands for culture and society). These fields differ in many dimensions. When we order fields on math and science difficulty, we get *NT*>*NH*>*ES*>*CS* (as in Buser et al. 2014). Students must select one field. Some students select two fields (or a combination thereof). Typical field combinations are then *NT* with *NH* and *ES* with *CS*. We define third grade students as STEM students when they choose *NT* or *NH* as specialization.

Each school has 4 to 7 classes per track. With about 25 to 30 students per class, some 800 students entered third grade the year the experiment started. Over a period of 3 years, a total of 2,423 third graders were involved in our study. We assigned 1,067 students to tracks with a GT program, of which 191 students had a high enough baseline GPA to participate in the GT program. Table 1 provides sample means and standard deviations of the outcome and control variables that we study below.

4 Main Results

Suppose GT education has a substantial impact on student achievement. Under ideal experimental conditions (where GT programs are randomly introduced in different tracks in many schools), we would then see similar GPA scores for non-eligible students (with baseline GPA scores below the admission cutoff) in treated and control tracks and a sharp rise in GPA scores for eligible students (with baseline GPA scores above the admission cutoff) in treated tracks. Figure 1 shows exactly this. There, we plot differences in GPA scores between students in treated and control tracks as a function of the baseline GPA scores. Each point represents the difference in GPA means for students in treated and control tracks in bins of 0.2 GPA points. The baseline GPA scores are normalized to 0 at the admission cutoff. For non-eligible students we see that GPA differences are close to zero and statistically insignificant. For eligible students we see that the differences in GPA scores are positive and mostly statistically significant. This suggests that GT education has a beneficial impact on student achievement, regardless of previous performance.

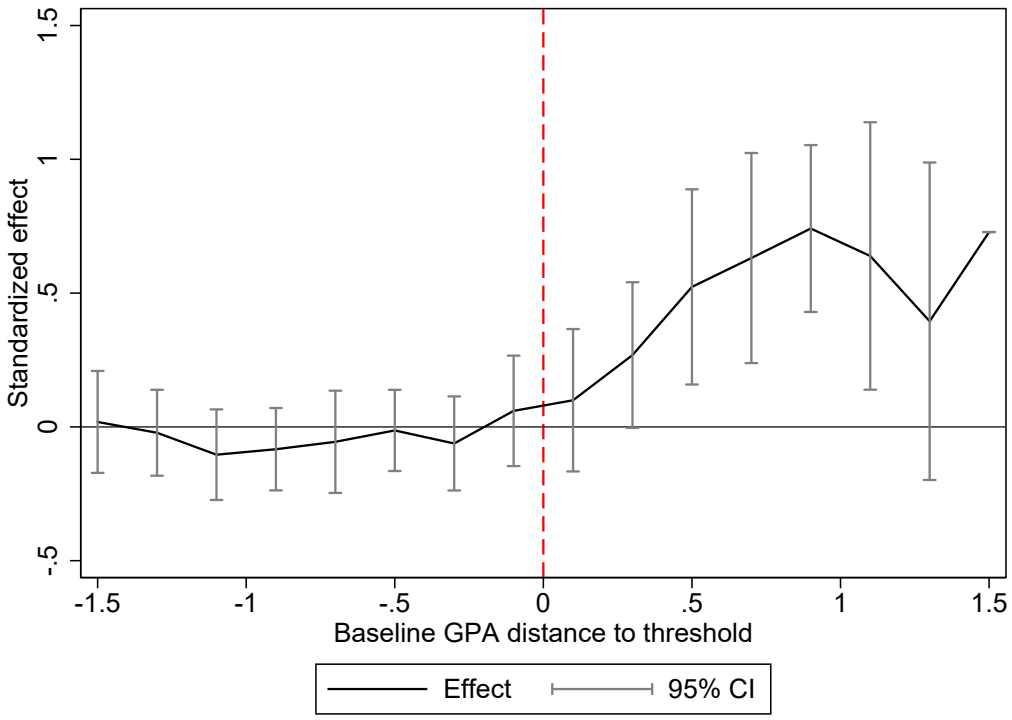


Figure 1. GPA differences in control and treated tracks
Note: The panel shows unconditional GPA differences for students in treated and control tracks by past GPA scores in symmetric bins of 0.2 GPA points around the eligibility cutoff. The baseline GPA scores are normalized to 0 at the eligibility cutoff.

4.1 ITT Results

Using the strategies described above, we can quantify the impact of GT education on student achievement. Table 2 contains the ITT estimates. All the estimates are obtained using ordinary least squares regressions.

DD Results

In columns 1 to 3 we present estimates of program eligibility on student achievement using the DD specifications in equation (1) with varying sets of control variables. In column 1 we estimate the sparsest model with program track (T), program eligibility (Z) and the interaction between the two ($T \times Z$) as the only right-hand-side variables. We find that exposure to GT education leads to a 0.23 (s.e. 0.10) standard deviation GPA gain. In columns 2 and 3 we add school, track, year dummies, and student

characteristics. The inclusion of these control variables does not affect the measured effect of being program eligible.

The causal interpretation of our DD estimates assumes that students in treated tracks, in absence of GT education, would perform similarly as students in control tracks. But can we use (potentially) eligible students in control tracks as the appropriate counterfactual students? We believe so. When we compare non-eligible students in treated and control tracks on their baseline performance, we find no structural differences in GPA scores. The treated track estimates in columns 1 to 3, which account for average GPA differences between non-eligible students in control and treated tracks, are all close to 0 and statistically insignificant. When we break up the baseline GPA into smaller segments, non-eligible students in control and treated tracks continue to perform very similarly, regardless of their baseline scores (as in Figure 1).

RD Results

In columns 4 to 6 we present the RD estimates of program eligibility on student achievement. In the RD specifications, we limit the sample to students in treated tracks and estimate equation (2) with varying sets of control variables. We find that, for students close to the admission cutoff, eligible students perform much better than non-eligible students. All three RD estimates are positive and statistically significant. Without additional covariates, we estimate GPA gains of about 0.14 (s.e. 0.06) of a standard deviation. With additional covariates, we estimate gains that are slightly larger.

The causal interpretation of our RD estimates assumes that eligible students just above the cutoff, in the absence of GT education, would perform similarly as non-eligible students just below the cutoff. Again, we can ask ourselves whether we can use non-eligible students near the cutoff as the appropriate counterfactual students. Standard tests suggest we can. Would we observe a discontinuity at the admission cutoff in program participation (suggesting that students near the cutoff are treated differently), but not in past pre-program GPA scores (suggesting that students near the cutoff are similar in pre-program characteristics), any positive (or negative) GT program effect on post-program GPA scores can be interpreted in a causal way. The RD graphs, which we show in the Appendix, confirm this. Figure B2 shows a sharp jump in the probability of GT participation (without bunching at the cutoff). Figure B3 shows no apparent jump in age, gender, and primary school

Table 2. Estimated ITT effects of the GT program on GPA

	DD			RD			DD/RD		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GT program	0.23**	0.22**	0.22**	0.14**	0.17***	0.17***	0.11*	0.10*	0.10*
	(0.10)	(0.09)	(0.09)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
GT program $\times z$				0.48***	0.54***	0.52***	0.45***	0.48***	0.47***
				(0.09)	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)
Z	1.54***	1.51***	1.47***						
	(0.08)	(0.08)	(0.07)						
z				1.16***	1.12***	1.12***	1.20***	1.20***	1.19***
				(0.05)	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)
Treated track	0.02	-0.01	-0.01				-0.02	0.04	0.04
	(0.08)	(0.06)	(0.06)				(0.06)	(0.05)	(0.05)
School, track, year dummies		✓	✓		✓	✓		✓	✓
Controls			✓			✓			✓
$\overline{y}$	0.00	0.00	0.00	0.09	0.09	0.09	0.00	0.00	0.00
$sd\,(y)$	1.00	1.00	1.00	1.03	1.03	1.03	1.00	1.00	1.00
p-value	0.022	0.024	0.019	0.024	0.006	0.006	0.061	0.074	0.066
p-value joint							0.000	0.000	0.000
ITT							0.29	0.29	0.29
s.e.(ITT)							0.05	0.05	0.05
p-value ITT							0.000	0.000	0.000
R^2	0.34	0.41	0.42	0.77	0.80	0.80	0.71	0.75	0.75
$N_{cluster}$	88	88	88	40	40	40	88	88	88
N	2423	2423	2423	1067	1067	1067	2423	2423	2423

Note: Each column represents a different OLS regression. Controls are gender, age, and CITO. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level. The reported p-value (joint) comes from a t-test (F-test) of the GT program coefficient (and interaction). The ITT is the average intent-to-treat estimate with reported standard error and p-value.

exit test scores (CITO), which are all predictors of student achievement. And Figure B4 shows a clear discontinuity at the admission cutoff in student achievement when we visually zoom in on students near the admission cutoff (panel **(b)**).

DD/RD Results

In columns 7 to 9 we combine the DD and RD approaches and present estimates that capture the impact of GT education for eligible students near and away from the admission cutoff. In column 7 we find that average student achievement jumps up by 0.10 of a standard deviation for eligible students with baseline GPA scores near the cutoff, and increases by 0.45 of a standard deviation for each additional point in baseline scores thereafter.[4] At the bottom of the table, we also report estimates of the average impact of GT education for all eligible students. Relative to eligible students in the control tracks, we find that eligible students exposed to GT education experience a 0.29 of a standard deviation gain in GPA. Including additional exogenous variables does not affect any of the program impact estimates, including the impacts we estimate for students near the cutoff, for students away from the cutoff, as well as the average impact for all students above the cutoff.

Figure 2 illustrates the sparsest model we have estimated in column 7. The blue line shows the linear relationship between current and baseline GPA scores for potentially eligible and non-eligible students in control tracks. The two black lines show the linear relationship between current and baseline GPA scores for non-eligible and eligible students in treated tracks separately. The black broken line speculates what the relationship between current and baseline GPA for eligible students in the treated tracks would be in the absence of GT education using the extrapolated relationship between current and past GPA scores for non-eligible students in treated tracks. Perhaps not surprisingly, we see that the baseline GPA scores serve as a strong predictor for current GPA scores. Average GPA scores increase steeply for all non-eligible students. We also see that the lines for non-eligible students in control and treated tracks are on top of each other.[5] Average GPA scores continue

[4]We have also estimated GT program effects for increasing bandwidth samples. In Appendix Figure B1 we report the corresponding estimates for eligible students near the cutoff using the DD/RD and RD specifications with covariates. We see that the RD estimates are most sensitive to bandwidth choice; that is, there we get increasing GT program effect estimates as the bandwidth increases. We therefore take the DD/RD estimates (for students near the cutoff) as the preferred estimates, which are slightly smaller than the corresponding RD estimates but much less sensitive to bandwidth choice.

[5]To formally test whether the relationship between current and baseline GPA is similar for non-

to increase gradually for eligible students in control tracks.[6] Average GPA scores for students in treated tracks, however, jumps at the admission cutoff and then continues to increase more steeply, particularly in comparison to potentially eligible students in control tracks.

Interestingly, the similarity in performance of non-eligible students in treated and control tracks rules out that the positive program effects come from non-eligible students doing worse. Suppose, for the moment, that non-eligible students in treated tracks score a lower GPA than they would normally score (in the absence of a GT program) because they feel frustrated for being left out or because they miss out on spillovers from those high-achieving students who left the classroom. Since we do not see that non-eligible students in treated tracks perform less than non-eligible students in control tracks, we believe that our results indicate that eligible students in treated tracks benefit from being exposed to GT education and because of that score a higher GPA at the end of the third grade.[7]

4.2 Other Outcomes

Table 3 contains additional ITT estimates of the effect of GT education on other student achievement measures. We disaggregate our GPA measure and consider the more commonly used GPA measures for math, language, and other subjects.

---

eligible students in treated and control schools, we estimate the following relationship on a sample of non-eligible students (Z_i=0)

$$Y_{isty} = \alpha_1 T_{st} + \alpha_2 T_{st}\,(z_i - \overline{z}) + \gamma_1\,(z_i - \overline{z}) + \epsilon_{isty},$$

where the forcing variable has been centered at 0 in the estimation sample. If the parameters α_1 and α_2 are zero, non-eligible students in treated and control tracks are similar in terms of academic performance. The estimates (with standard errors in parenthesis) we get for α_1, α_2 and γ_1 are 0.02 (0.06), -0.08 (0.07) and 1.24 (0.05), respectively. The estimates α_1 and α_2 are also jointly statistically insignificant (with a p-value of 0.43).

[6]We have also tested for a trend break at the cutoff for students in control tracks ($T_i = 0$). In particular, we have estimated the following relationship on a sample of untreated students

$$Y_{isty} = \gamma_1 z_i + \gamma_2 Z_i z_i + \epsilon_{isty}.$$

If the parameter γ_2 is zero, the relationship between current and baseline GPA is similar for untreated students with baseline scores below and above the cutoff. With the estimates (with standard errors in parenthesis) we get for γ_1 and γ_2, being 1.22 (s.e. 0.05) and -0.10 (s.e. 0.14) respectively, we find no evidence of a structural break for these students.

[7]Card and Giuliano (2016) also compare math and reading test scores between high-achieving students near and away from the cutoff in schools with and without a GT classroom program. While they find qualitatively similar program effect estimates for students near the admission cutoff, they also find that the positive impact of the GT program fades out for students further away from the admission cutoff.

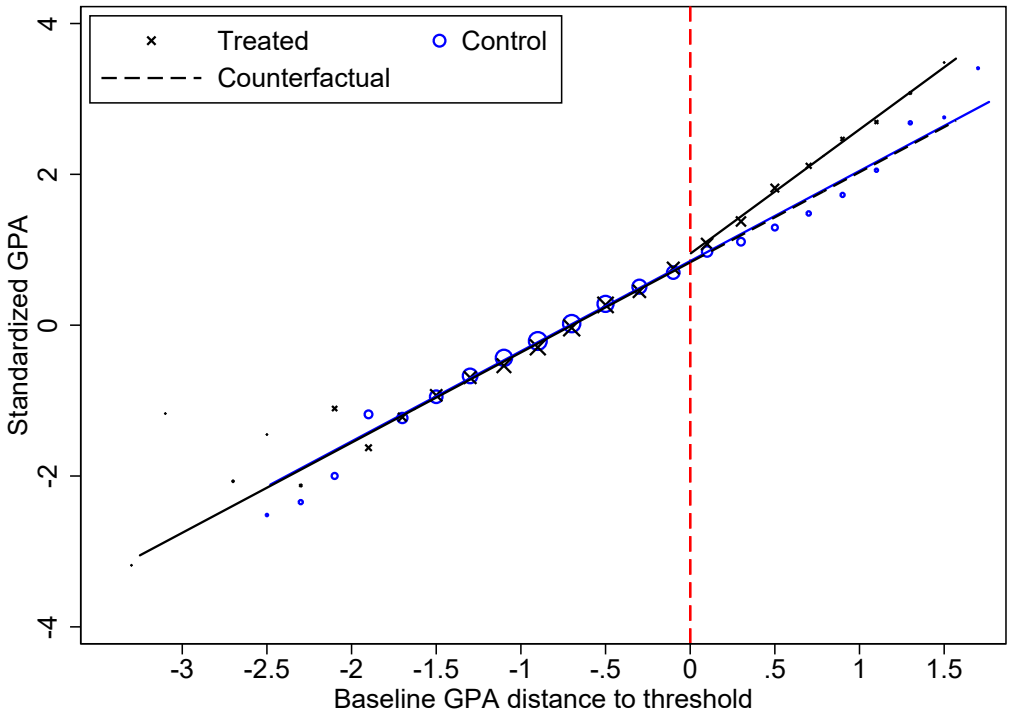


Figure 2. GPA of students in control and treated tracks
Note: The panel shows unconditional GPA scores for students in treated and control tracks by past GPA scores in bins of 0.2 GPA points. The past GPA scores are normalized to 0 at the eligibility cutoff.

We also consider field choice students make at the end of the third school year, which determines their curriculum in later school years with important implications for future graduate school choices. In particular, we estimate whether students are more likely to choose one of the more demanding science tracks (NT or NH), which we refer to as STEM choice. The model and identification strategy is the combined DD/RD model, which we estimate with and without additional covariates by OLS (as in Table 2, columns 7 and 9).

We first look at GPA in math and language separately. For eligible students near the cutoff, we find that the effects are larger for math than for language. The estimated program effects for math range from 0.12 to 0.21 standard deviations, whereas the estimates for language are close to 0.10 and statistically insignificant. For eligible students further away from the cutoff, we find that the program impacts reverse in magnitude and get significantly larger for language than for math. The

Table 3. Estimated ITT effects of the GT program on other outcomes

	Math		Language		Other		STEM	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GT program	0.21**	0.12	0.06	0.10	0.10*	0.08	0.07	0.10*
	(0.09)	(0.08)	(0.07)	(0.07)	(0.06)	(0.06)	(0.05)	(0.05)
GT program$\times z$	0.38***	0.25**	0.52***	0.63***	0.26***	0.24***	-0.02	-0.01
	(0.13)	(0.12)	(0.11)	(0.11)	(0.08)	(0.07)	(0.11)	(0.10)
z	0.83***	0.95***	1.06***	0.99***	1.11***	1.13***	0.18***	0.16***
	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)
Treated track	-0.03	0.08	0.03	0.05	-0.03	0.01	0.02	-0.03
	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.05)	(0.03)	(0.03)
School, track, year dummies		✓		✓		✓		✓
Controls		✓		✓		✓		✓
$\overline{y}$	0.00	0.00	0.00	0.00	0.00	0.00	0.47	0.47
$sd\,(y)$	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.50
p-value	0.023	0.143	0.368	0.158	0.088	0.153	0.160	0.062
p-value joint	0.000	0.007	0.000	0.000	0.000	0.001	0.237	0.060
ITT	0.36	0.21	0.26	0.34	0.20	0.17	0.06	0.10
s.e.(ITT)	0.09	0.08	0.06	0.06	0.06	0.06	0.04	0.04
p-value ITT	0.000	0.006	0.000	0.000	0.001	0.004	0.108	0.020
R^2	0.37	0.42	0.57	0.64	0.60	0.64	0.07	0.12
$N_{cluster}$	88	88	88	88	88	88	88	88
N	2423	2423	2423	2423	2423	2423	2423	2423

Note: Each column represents a different OLS regression. Controls are gender, age, and CITO. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level. The reported p-value (joint) comes from a t-test (F-test) of the GT program coefficient (and interaction). The ITT is the average intent-to-treat estimate with reported standard error and p-value.

average impact for all eligible students above the cutoff, however, are all positive, statistically significant, and similar in magnitude. Average GPA math score gains are in the order of 0.21 to 0.36 standard deviations, whereas the average GPA language score gains are in the order of 0.26 to 0.34 standard deviations (columns 1 to 4, bottom panel). When we run our regressions on GPA in other subjects, we find that the effects are overall somewhat weaker. For students near the cutoff, the program estimates for other subjects are positive but smaller than for math. For students away from the cutoff, the program estimates for other subjects are again positive but smaller than for language.

We next take a look at curriculum choice. We find that the program raised the overall likelihood that students choose a more science oriented curriculum. For eligible students near the cutoff, the estimated program impact on STEM choice is positive and ranges from 7 and 10 percentage points. Only the 10 percentage point increase is statistically significant at traditional confidence levels. These effects do not change for students further away from the cutoff. We estimate positive impacts for all eligible student at 6 to 10 percentage points (columns 5 and 6, bottom panel).

Although the estimates appear more sensitive to the inclusion of control variables than the estimates for the main outcome presented in 2, our estimates clearly indicate that the GT program has increased academic performance for all eligible students: that is, they obtain higher grades in all subjects, and choose a more science demanding curriculum.

4.3 IV Results

As noted, eligible students do not always participate in the GT program. Similarly, some non-eligible students sometimes do participate. To identify the impact of GT education on student achievement among students who actually participate in the program, we will apply a standard IV setup and use eligibility status as our instrumental variable for program participation. We take again the combined DD/RD model as our preferred model, which we estimate with and without additional covariates by 2SLS (see appendix table A1 for the corresponding 2SLS estimates for the DD and RD specifications). One complication of this model is the additional endogenous interaction term, for which we need an additional source of exogenous variation in program participation. Our experimental setup deals with this. In particular, we are able to exploit two groups of counterfactual students: non-eligible students near the admission cutoff in treated tracks, and eligible students in control

tracks.

Table 4 contains IV regression results for all outcomes, where GT program use ($\times z$) is instrumented by GT program eligibility ($\times z$). The first stage estimates that mediate these results are 0.49 (s.e. 0.06) at the cutoff and about 0.21 (s.e. 0.11) further away (interaction). This suggests that complience is about 50% the cutoff and about 0.70 percent for students with initial GPA scores one point higher than the cutoff. The first stage estimates are insensitive to the inclusion of covariates, and jointly statistically significant, with corresponding F statistics high enough for our instruments to be relevant.

In columns 1 and 2 we present the 2SLS estimates of the impact of program participation on student achievement (as measured by overall GPA scores). We find that all the impact estimates get larger, especially those for eligible students near the cutoff. According to these estimates, students who just made it into the program score about 0.22 to 0.25 standard deviation higher because of GT program participation. The positive interaction estimates of 0.57 to 0.60 indicate that stronger students benefit even more from the GT program. In the bottom of Table 3, we present additional 2SLS estimates of the effect of GT education of student achievement for all participating students who comply to their eligibility status (LATE). For those complying students, we find that GT program participation raises overall GPA with about 0.45 of a standard deviation.

In columns 3 to 8 we present results for disaggregated outcomes: math, language, and other subjects. While these estimates are more sensitive to the inclusion of covariates than the aggregate outcomes in columns 1 and 2, they concur with the ITT findings presented earlier: we find positive impacts across the board with effects of about 0.20 standard deviation at the cutoff, and an additional 0.30 for students one GPA point further away or more. The estimated program impacts for compliers are about 0.30 of a standard deviation for math and other subjects, and 0.50 for languages. These results are all statistically significant.

In columns 9 and 10 we consider STEM choice. For students near the cutoff, we find that those students who participate in the GT program are about 0.15 to 0.20 percentage points more likely to choose a more science intensive study track. These effects are sizable, and with the inclusion of additional covariates, sizable enough to be statistically significant. For students further away from the cutoff, the overall impact on STEM choice seems to fall but not in a meaningful way. The overall effect on the complier population is still substantial; that is, the average eligible student

Table 4. Estimated LATE's of the GT program

	Overall		Math		Language		Other		STEM	
	DD/RD		DD/RD		DD/RD		DD/RD		DD/RD	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GT program use	0.25*	0.22*	0.44**	0.24	0.14	0.21	0.21*	0.17	0.15	0.20*
	(0.14)	(0.13)	(0.20)	(0.17)	(0.15)	(0.15)	(0.13)	(0.12)	(0.10)	(0.11)
GT program use$\times z$	0.57***	0.60***	0.41*	0.28	0.69***	0.82***	0.31**	0.29*	-0.08	-0.07
	(0.20)	(0.20)	(0.24)	(0.19)	(0.23)	(0.27)	(0.16)	(0.15)	(0.17)	(0.17)
z	1.19***	1.19***	0.82***	0.95***	1.06***	0.99***	1.11***	1.13***	0.18***	0.16***
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)
Treated track	-0.02	0.03	-0.04	0.07	0.03	0.05	-0.03	0.01	0.01	-0.04
	(0.06)	(0.05)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.05)	(0.03)	(0.03)
School, track, year d.		✓		✓		✓		✓		✓
Controls		✓		✓		✓		✓		✓
$\overline{y}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47	0.47
$sd\,(y)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.50
p-value	0.070	0.082	0.028	0.144	0.337	0.164	0.099	0.164	0.157	0.063
p-value joint	0.000	0.000	0.001	0.012	0.000	0.000	0.001	0.005	0.254	0.074
FS F-stat level	57.2	61.7	57.2	61.7	57.2	61.7	57.2	61.7	57.2	61.7
FS F-stat slope	58.4	60.3	58.4	60.3	58.4	60.3	58.4	60.3	58.4	60.3
LATE	0.46	0.45	0.60	0.35	0.41	0.53	0.33	0.28	0.12	0.17
s.e.(LATE)	0.11	0.11	0.18	0.14	0.12	0.12	0.11	0.11	0.07	0.08
p-value LATE	0.000	0.000	0.001	0.014	0.000	0.000	0.003	0.009	0.100	0.022
R^2	0.71	0.75	0.36	0.42	0.56	0.63	0.60	0.64	0.07	0.11
$N_{cluster}$	88	88	88	88	88	88	88	88	88	88
N	2423	2423	2423	2423	2423	2423	2423	2423	2423	2423

Note: Each column represents a different IV regression where GT program use ($\times z$) is instrumented by GT program eligibility ($\times z$). The odd columns are restricted to the treated sample. Controls are gender, age, and CITO. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level. The reported p-value (joint) comes from an F-test of the GT coefficient (and interaction). The FS F-stat level(slope) is the joint first stage F-statistic of the instruments on GT program use ($\times z$). The LATE is the estimated local average treatment effect with reported standard error and p-value.

is significantly more likely to opt for a more science driven curriculum because of the GT program, with point estimates of 0.12 (s.e. 0.07) and 0.17 (s.e. 0.08) in specifications without and with control variables. With roughly 50 percent of the students choosing STEM, these impact estimates corresponds to an increase of at least 25 percent.

5 Summary and Discussion

Selective GT education programs are becoming increasingly popular in secondary education. In this paper we examine whether such programs can raise the educational performance of secondary school students in the Netherlands. In particular, we invite third grade students with a high enough GPA in the beginning of the school year to take part in a program where they can trade off classroom teaching for in-school time to work on self-selected projects. Using the fragmented program implementation in different study tracks among 2,400 third grade students in three large comprehensive schools, we provide evidence that third graders eligible for the program experienced significant GPA gains measured at the end of the third school year (average gain of about 0.30 standard deviations) and were significantly more likely to choose the science intensive track for the subsequent school years (average rise of 6-10 percentage points). Interestingly, we also find that the benefits of attending the program in terms of GPA gains were much stronger for students with higher baseline grades. When the study period of three years elapsed, the three schools decided to continue with GT education and roll out the GT program to all their first, second and third grade students.

While the GT program under study works, it is not directly clear why it works. Being above the cut-off involves many things, including being told to belong to the group of high achievers, decreasing the number of classroom hours, getting help working on a project of choice, and participating in a competition. With the data at hand, it is not possible for us to test for the mechanisms underlying our results.

Our results should, nonetheless, speak to researchers, educators and policy makers. Researchers who study selective GT programs often use RD strategies and thus identify, by design, the impact on student performance using gifted students near the admission cutoff. Their results are mixed. While some studies find that GT education works and raise the academic skills of eligible students, other studies find no gains at all. Our results suggest that these studies may have missed some of the

possible benefits of GT programs for smarter students who are further away from the cutoff.

Educators often debate GT programs. Some advocate such programs for challenging gifted students to reach their full (academic) potential. Others criticize the same programs for being elitist and unfair. Our results should appeal to both sides of the debate. GT advocates should generally like GT programs that work. GT critics should be less concerned when the GT program is not restricted to a few smart students. Our results indicate that a GT program (in which students can replace classroom hours for project hours) is beneficial for a much broader group of students. In addition, GT program costs are low. In the three schools, we calculate the average additional costs per participating student at about €200 per year. While such GT programs take away some resources from other students, it is hard to think of a substitute program with comparable gains that is cheaper.

And finally, policy makers may want to know whether to scale up such a program. Of course, advice on scaling up should generally depend on the broader generalizability of results. In another study (Booij et al., 2016), we examine the effect of a comparable GT program (already implemented since 1983) at a prestigious academic secondary school in another city. There we find results very similar to those presented here: that is, eligible students near the cutoff obtain higher grades and follow a more science intensive curriculum. In this light, we treat our results as complementary, suggesting that it is relatively easy to implement a simple individualized pull-out program that is effective for high-achieving students.

References

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D. and Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344.

Bertanha, M. (2016). Regression discontinuity design with many thresholds. *CORE Discussion Papers*.

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

Booij, A., Haan, F., and Plug, E. (2016). Enriching students pays off: Evidence from an individualized gifted and talented program in secondary education. *IZA Discussion Paper Series*, 9757.

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.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices*. *Quarterly Journal of Economics*, 129(3):1409–1447.

Card, D. and Giuliano, L. (2015). Can universal screening increase the representation of low income and minority students in gifted education? Working Paper 21519, National Bureau of Economic Research.

Card, D. and Giuliano, L. (2016). Can tracking raise the test scores of high-ability minority students? *American Economic Review*, 106(10):2783–2816.

Dong, Y. and Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 97(5):1081–1092.

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

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.

Appendices

A IV estimates for DD and RD specifications

Table A1. Estimated LATE's of the GT program

	Overall		Math		Language		Other		STEM	
	DD (1)	RD (2)	DD (3)	RD (4)	DD (5)	RD (6)	DD (7)	RD (8)	DD (9)	RD (10)
GT program use	0.39**	0.36**	0.65***	0.19	0.50**	0.29*	0.09	0.39***	0.18*	0.18*
	(0.17)	(0.15)	(0.17)	(0.19)	(0.20)	(0.16)	(0.17)	(0.15)	(0.11)	(0.11)
GT program use$\times z$		0.63***		0.24		0.83***		0.36**		-0.10
		(0.21)		(0.19)		(0.26)		(0.17)		(0.17)
Z	1.47***		0.96***		1.21***		1.46***		0.18***	
	(0.07)		(0.07)		(0.08)		(0.08)		(0.05)	
z		1.11***		0.98***		0.95***		0.99***		0.17***
		(0.05)		(0.06)		(0.06)		(0.05)		(0.03)
Treated track	-0.02		-0.01		-0.01		-0.03		-0.04	
	(0.06)		(0.08)		(0.06)		(0.06)		(0.03)	
School, track, year dummies	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$\overline{y}$	0.00	0.09	0.00	0.07	0.00	0.11	0.00	0.07	0.47	0.50
$sd\,(y)$	1.00	1.03	1.00	1.05	1.00	1.02	1.00	0.99	0.50	0.50
p-value	0.021	0.013	0.000	0.332	0.010	0.071	0.598	0.008	0.088	0.092
p-value joint	0.021	0.000	0.000	0.137	0.010	0.000	0.598	0.000	0.088	0.222
FS F-stat level		55.0		55.0		55.0		55.0		55.0
FS F-stat slope	119.9	57.7	119.9	57.7	119.9	57.7	119.9	57.7	119.9	57.7
R^2	0.42	0.78	0.18	0.48	0.41	0.64	0.36	0.69	0.09	0.09
$N_{cluster}$	88	40	88	40	88	40	88	40	88	40
N	2423	1067	2423	1067	2423	1067	2423	1067	2423	1067

Note: Each column represents a different IV regression where GT program use ($\times z$) is instrumented by GT program eligibility ($\times z$). The odd columns are restricted to the treated sample. Controls are gender, age, and CITO. Class clustered standard errors in parenthesis. */**/*** denote significance at a 10/5/1 percent confidence level. The reported p-value (joint) comes from an F-test of the GT coefficient (and interaction). The FS F-stat level(slope) is the joint first stage F-statistic of the instruments on GT program use ($\times z$).

B RD graphs

To illustrate the credibility of our RD design, we show graphs in which we plot GT program admission, GT pre-treatment cognitive test scores, and post-treatment GPAs, against baseline GPA scores. Discontinuities observed at the admission threshold in GT program admission and GPAs, but not in pre-treatment baseline GPAs, would imply that any positive (or negative) GT program effect can be interpreted in a causal way. Below we also plot RD estimates of program eligibility on post-treatment GPA on smaller samples. We have selected the smallest sample based on the formal bandwidth selection procedure of Imbens and Kalyanaraman (2012). Credible estimates are robust to wider bandwidth choices.

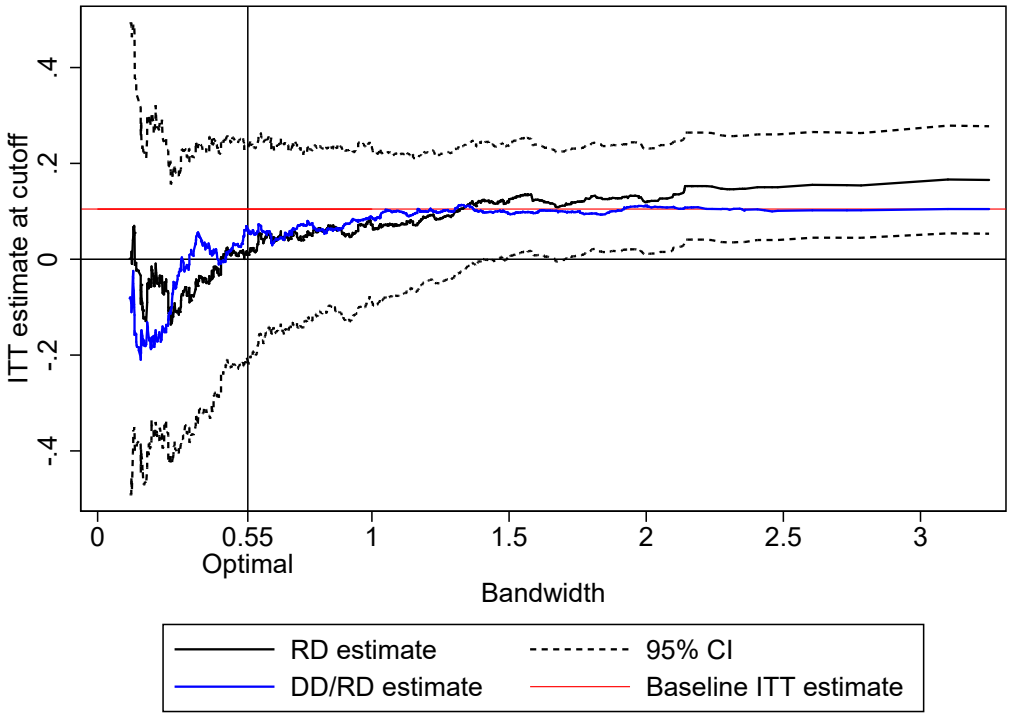
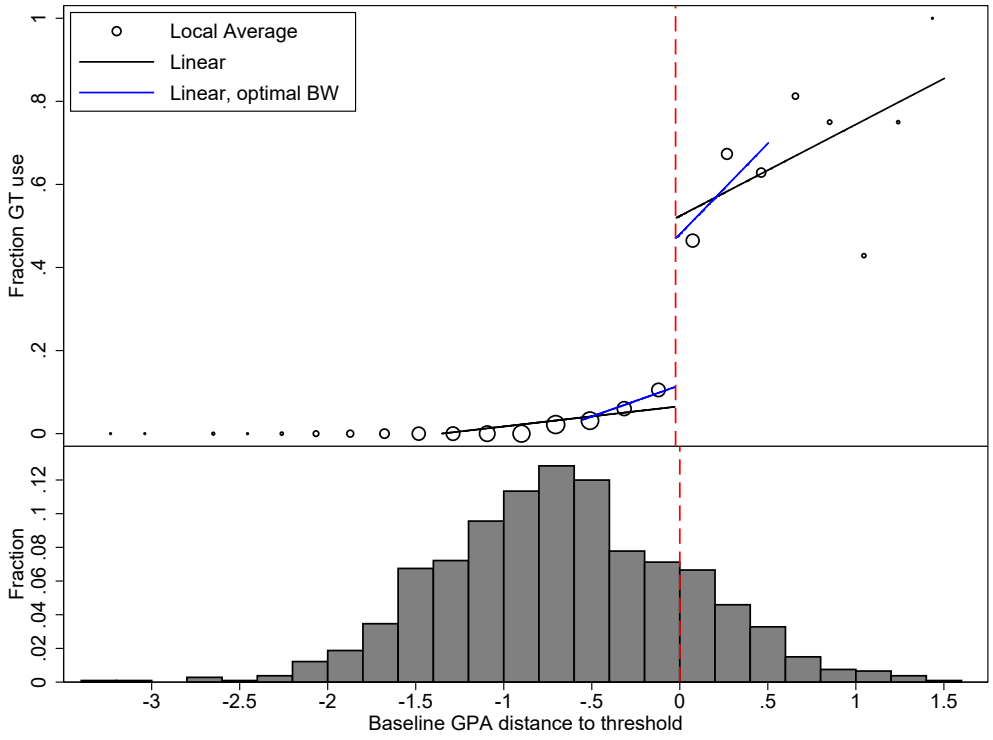


Figure B1. ITT effect estimates from linear split regression with varying bandwidth

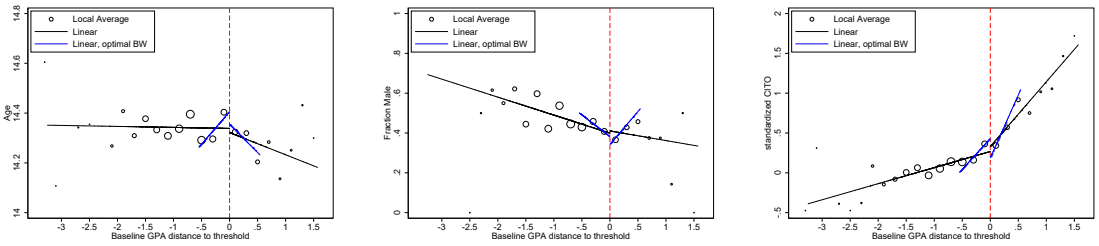
Note: The optimal bandwidth level is chosen following Imbens and Kalyanaraman (2012) using the treated sample with covariates included.



.

Figure B2. Fuzzy RD first stage effect

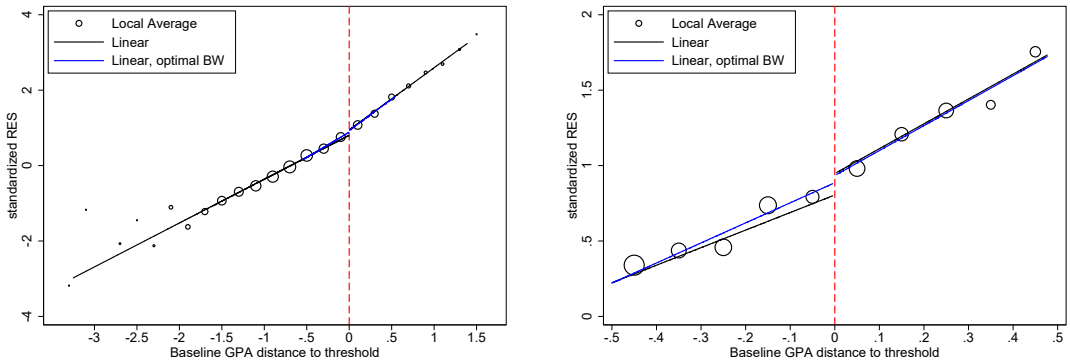
Note: The top panel shows fitted values from parametric and non-parametric first stage regressions of GT program assignment on baseline GPA scores in the treated sample, without covariates. The bottom panel shows the distribution of normalized GPA scores. The admission cutoff in this picture is normalized to 0.



(a) **(b)** **(c)**

Figure B3. Fuzzy RD reduced form effects on pre-treatment outcomes age, gender and CITO scores

Note: The left, middle, and right panels show fitted values from parametric and non-parametric reduced form regressions of pre-treatment outcomes age (coeff. -0.02, s.e. 0.05), gender (coeff. -0.01, s.e. 0.06), and CITO scores (coeff. 0.10, s.e. 0.11) on baseline GPA scores in the treated sample, without covariates. The eligibility cutoff in this picture is 0.



(a) **(b)**

Figure B4. Fuzzy RD reduced form effects on GPA scores

Note: The left panel shows fitted values from parametric and non-parametric reduced form regressions of the main outcome student GPA on baseline GPA scores, without covariates, using a sample all treated students. In the right panel we show the same results but zoom in on treated students near the cutoff.