CESifo WORKING PAPERS

10212
2023

January 2023

Long-Term Effects of Grade Retention

Simon ter Meulen

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10212*

Long-Term Effects of Grade Retention

Abstract

Grade retention offers students a chance to catch up with unmastered material but also leads to less labor-market experience by delaying graduation and labor-market entry. This is the first paper to quantify this trade-off, using an exit exam cutoff of Dutch academic secondary schools, where failing implies grade retention. I find no impact of retaining on final educational attainment, although retained students are later to graduate. Grade retention does lead to annual earnings loss at age 28 of 3000 euro (8.5%) due to reduced labor-market experience. Overall, grade retention is of no benefit for students around the cutoff.

JEL-Codes: I210, I230, I260.

Keywords: grade retention, secondary education, higher education degrees, earnings loss.

Simon ter Meulen
ifo Institute – Leibniz Institute for Economic Research
at the University of Munich
Poschingerstraße 5
Germany – 81679 Munich
termeulen@ifo.de

This version: November 2022
I am grateful to Hessel Oosterbeek and Bas van der Klaauw for their guidance and advice throughout this project. I further thank Julius Ilciukas, Peter Fredriksson, Mikael Lindahl, Magne Mogstad, José Montalbán Castilla, Oskar Nordström Skans, Umut Özek, Erik Plug, Jenifer Ruiz-Valenzuela, Ștefania Simion, Giuseppe Sorrenti, Leonard Treuren, Dinand Webbink, Andrej Werner, and Ludger Woessmann for their valuable comments; as well as the seminar participants at the 33rd EALE conference, the EEA-ESEM 2021 conference, the 34th ESPE conference, the Erasmus University Rotterdam, ifo Center for the Economics of Education, the third HELM workshop, the Helsinki Labour Institute for Economic Research, the Paris School of Economics, the Masaryk University, the University of Gothenburg, the University of Stavanger, the University of Lancaster, and the University of Uppsala. The non-public micro data used in this paper are available via remote access to the Microdata services of Statistics Netherlands (CBS). All remaining errors are my own.

1 Introduction

Grade retention is a popular remedial practice throughout the world, affecting millions of students each year.[1] Moreover, grade retention has profound labor-market consequences for those who have to retain, as it delays labor-market entry and potentially increases students' human capital by extending learning by one year. Recently, some seminal and essential papers have made headway in studying the human capital effects of grade retention by studying the effects of grade retention on various educational outcomes (Jacob and Lefgren (2004); Jacob and Lefgren (2009); Manacorda (2012); Fruehwirth et al. (2016); Schwerdt et al. (2017); Figlio and Özek (2020); Landaud and Maurin (2020)). I contribute to this emerging literature by studying the effects of grade retention on labor-market outcomes and final (tertiary) educational attainment, thereby quantifying the central trade-off of grade retention.

In theory, grade retention can have opposing impacts on retainers' human capital. In the short run, grade retention gives students the chance to master the educational content they failed to learn in the previous year, which should lead to increased educational attainment in the long run. However, grade retention also delays study progression by one year. This delay implies later graduation and later entry into the labor market and, therefore, less labor-market experience and lower wages. Consequently, in terms of earnings, it is an empirical question whether the remedial effect of grade retention is larger than the effect of the delay.

There are several identification challenges that have hampered previous work. First, teachers only retain low-performing students, and so in most contexts, retained and promoted students cannot be readily compared.[2] Second, the grade retention effects on test scores are confounded by age because retained students arrive one year later, and thus one year older, at any subsequent test. This age-at-test problem makes identification of grade retention effects on the most natural student performance measure, test scores, hard to disentangle from age effects.[3] Third, students can retain at multiple moments during their educational career, making it likely that a share of the initial control group is treated at a later point.[4]

In this paper, I deal with these identification challenges in the following way. First, I exploit the exogenous variation created by an exit exam cutoff in a regression discontinuity design (RDD) to deal with the selective nature of grade retention. Second, I focus on outcome variables that are free from age-at-test problems, such as educational attainment and earnings. Third, I address the sequential retention problem by using the last year of secondary school, as

[1] Using PISA 2018 data, I estimate that the average percentage of 15-year-old students that have ever been retained across OECD countries is 10.6%. Figure A1 in the Appendix plots this ever-retained rate at age 15 for each of the OECD countries.

[2] Allen et al. (2009) provides an overview of older work on grade retention.

[3] If we measure an increase in test scores of the retained on a conventional subsequent test, we cannot identify if this is due to grade retention (re-doing the material) or due to the increase in age (an additional year of maturation), or due to a regression to the mean effect.

[4] Sequential retention, which is a form of dynamic treatment assignment, can be especially problematic in static evaluations of grade retention as students of different abilities retain at different stages of their educational career, see Fruehwirth et al. (2016).

there is no retention after the final year. As a result, I can simultaneously account for selection, age-at-test, and sequential retention.

I use the graduation rules of Dutch academic secondary school to identify the exit exam cutoff. Dutch academic secondary school prepares students for university, which only require students to pass the secondary school exit exam. Students in secondary schools are normally retained on a discretionary basis, but in the final year of the academic track, multiple exit exams determine whether they fail and retain. The national graduation rules state that if the students fail three or more major course exit exams, they do not graduate and retain. Students with two failed major courses, and no others, can retake one of their failed exams and graduate upon passing it. This implies that the difference between two and three failed major courses contains identifying variation. The third-lowest grade among the major courses captures this variation. Exam grades run from 1 to 10, and a 5.5 is the lowest passing grade. As such, a 5.5 as the third-lowest major course grade indicates two failed major courses, and a 5.4 indicates three failed major courses. I use this 5.5 cutoff on the third-lowest major grade for identification in the regression discontinuity design.

Using Dutch registry data, I find that failing to graduate entails spending an additional year in secondary education, implying that failing to graduate is the same as retaining. Students do not drop out of secondary education due to failing, nor do retaining students switch to lower ability tracks. After doing the additional year, retaining students increase their GPAs, but this could be due to an age-at-test effect or due to regression toward the mean. Still, a naive observer might think that grade retention is a powerful way to increase performance, because they do not observe the long-term outcomes. I do not find that more students use anti-depression medication due to failing, which is an indication that the psychological shock of failing and retaining is limited.

One of the two main findings is that grade retention does not affect educational attainment. Still, retainers are less likely to enroll in university and more likely to enroll in the less academical professional colleges. But, promoted students are more likely to switch midway from university to these professional colleges. As a result, there is no effect of grade retention on the probability of getting a university or college degree at the age of 28, the last age for which I can measure educational attainment for all cohorts. Nonetheless, grade retention does lead to an average delay in graduating from higher education by about six months.

The other main finding is that grade retention causes a large earnings loss. Grade retainers earn about 3,000 euro, or 8.5%, less than promoted students at the age of 28. There are similar effects for the age range of 22 to 30, the last age I observe. Early in their twenties, the difference in earnings is mainly due to a difference in hours. At this point, some promoted students have already started to work where a larger fraction of retainers is still in education. Later in their twenties, the difference in earnings is mainly due to a lower hourly wage for the retained students.

The lower earnings at age 28 are likely due to the lower level of experience at that age.

Grade retainers have the same earnings as promoted students in the first year after higher education, and similarly, earn the same conditional on experience, indicating that the earnings trajectories of retainers are not affected. At the age of 28, retained students acquire on average five fewer months of work experience than promoted students, reflecting the later graduation date. The earnings losses at 28 are consistent with a five-month setback in the estimated students' earnings-experience profiles. In short, grade retention delays students in their labor-market trajectory by five months but does not affect their trajectories.

Although the RDD identifies effects at the cutoff, the identified effects apply to a wider group of students. First, the exogeneity test by Bertanha and Imbens (2020) shows that the observed outcomes are similar for treated compliers and always takers, and for non-treated compliers and never takers. Second, using the extrapolation method of Dong and Lewbel (2015), I find that the effects of grade retention for educational attainment and earnings are very similar for inframarginal students and marginal students. These results indicate that I would find similar treatment effects if the cutoff was marginally lower, and I would re-estimate the same RDD.

This study contributes to the small but growing literature on the causal effects of grade retention on retainers. Jacob and Lefgren (2004) and Schwerdt et al. (2017) document an increase in material mastered after grade retention in grade 3 among American students using an RDD. Both studies deal with the age-at-test problem by using vertically scaled tests.[5] However, in these papers, students who initially just pass, are likely to be retained in one of the following years. Jacob and Lefgren (2004) do not deal with this subsequent retention problem, but Schwerdt et al. (2017) use bounds to show that subsequent retention likely explains a substantial part of the initial effects' fade-out. Ferreira Sequeda et al. (2018) and Landaud and Maurin (2020) show that the short-run positive effects are not limited to early grade retention. Similar to this paper, they study students at the end of a (post-)secondary school educational program and do not have subsequent retention. However, they do not deal with the age-at-test problem and focus on short-run effects, which is in their case interesting in itself. Compared to these four studies, this paper shows the effect of grade retention in the long run, and accounts for the age-at-test problem and the subsequent retention problem.[6]

Some other studies exploit secondary school exit exams to investigate ability tracking and the signaling value of a degree. Diamond and Persson (2016) and Machin et al. (2020) use

[5] Vertically scaled tests use a similar exam question for all grades and item response theory to create a score free from age concerns.

[6] Another branch of this literature shows that grade retention can increase dropout (Eren et al., 2018; Jacob & Lefgren, 2009; Manacorda, 2012). I do not find such effects. A reason could be that people can be refused unemployment benefits without a secondary school degree until the age of 27. Another set of papers such as Fruehwirth et al. (2016), Gary-Bobo et al. (2016), Cockx et al. (2019) use structure to overcome selection and sequential retention problem. However, the primary outcome variable of these studies are conventional test scores, which makes the identified effects a bundle of the effects of grade retention, effects of age-at-test, and/or, the effect of being in different grade levels. Furthermore, Eide and Showalter (2001) also study long-term outcomes but with school-starting age instruments. Papers such as Oosterbeek et al. (2021) show that school-starting age has a direct effect on these long-term outcomes. Lastly, Figlio and Özek (2020) show that early grade retention coupled with instructional support substantially improves the English skills of non-native English learners.

secondary school exit exams to look at long(er)-term outcomes. However, both studies are, in essence, tracking studies. The exams they study determine whether or not a student will do vocational or more academic higher education.[7] Another set of exit exam studies, such as Clark and Martorell (2014), investigate signaling theories of education. They find that graduating in itself has no impact on the earnings of American high school students at the margin. Jepsen, Mueser, and Troske (2016) find a similar result for students that marginally pass the American GED. There is no exit exam paper that identifies the long-term effects of grade retention.

This paper continues as follows. The next section details the setting in which this research takes place. Besides discussing contextual details, it also shows how the graduation rules imply as-good-as-random variation that can be used for identification. Section 3 discusses the data, and section 4 shows how the research design uses the as-good-as-random variation. The results are in section 5, and the last section concludes and discusses policy implications.

2 Institutional context, graduation rules, and the identifying variation

2.1 The Dutch education system

The focus of this paper is the final year of the academic track of secondary school in the Netherlands. Figure 1 shows the later stages of the educational system in the Netherlands of which secondary school is part. Besides the academic ability track, there are vocational and college ability tracks in secondary school. These tracks vary in how academically demanding they are, with the academic track as the most demanding track. Additionally, the vocational track only takes four years, the college track takes five years, and the academic track takes six years. Students that are above the compulsory schooling age of 18 can choose to transfer to transition centers (ROC).[8]

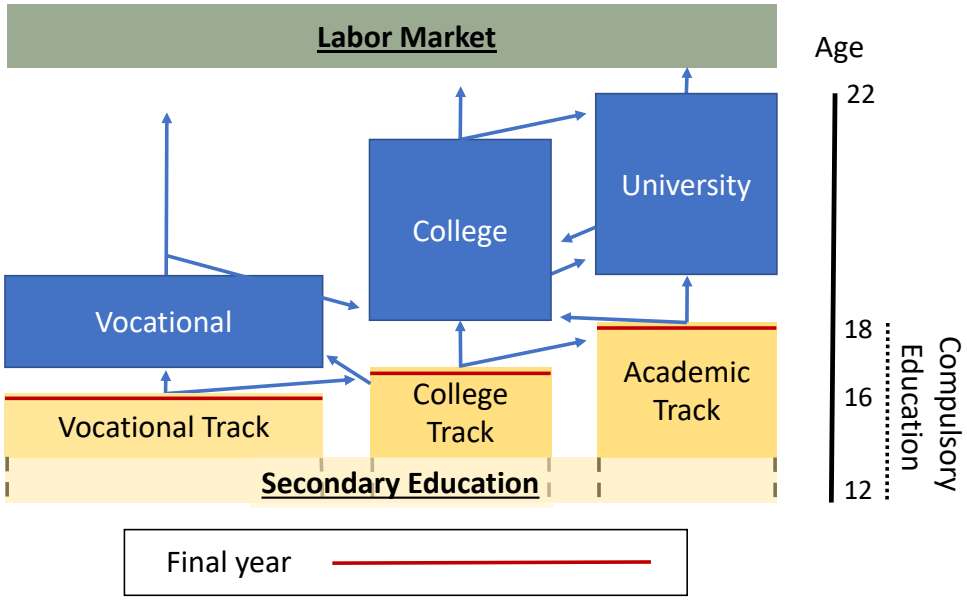
The focus is on the last year of the academic track because enrolled students need to pass this year to continue their education. The lack of switching options ensures that failing the exit exams implies grade retention. Students who fail in the two other tracks can continue in less academically demanding tertiary education instead of retaining the year. For instance, upon failing the last year of the college track, a student can enter vocational professional training without further examination. In these cases, failing does not imply grade retention. As the interest of this study is the effect of grade retention, the college and vocational tracks are excluded from the analysis.[9]

[7] Ability tracking happens in the Netherlands at the end of primary education instead of at the end of lower secondary education.

[8] In appendix H, I use the 2008 expansion of transition centers ("Rutte-regeling") to credibly show that there are no treatment differences between those retaining in transition centers and those retaining in their old secondary schools.

[9] Students who first obtain a college track secondary school degree before entering academic track can enter

Figure 1: Educational System of the Netherlands



Notes. The solid vertical bar on the right indicates students' age at nominal study duration. Arrows represent possible transitions.

Besides the ability track, students pick a major that prepares them for their field of study in university. There are four majors: science, health, social science, and humanities. A major implies picking four major courses.[10] Besides the major courses, students in the academic track take courses in four languages, literature, civics, artistic education, history, general physics, and one free-choice course.[11]. In 2009 there was a re-organization of the major system, including a reform of the promotion rules, limiting the cohorts this study uses.

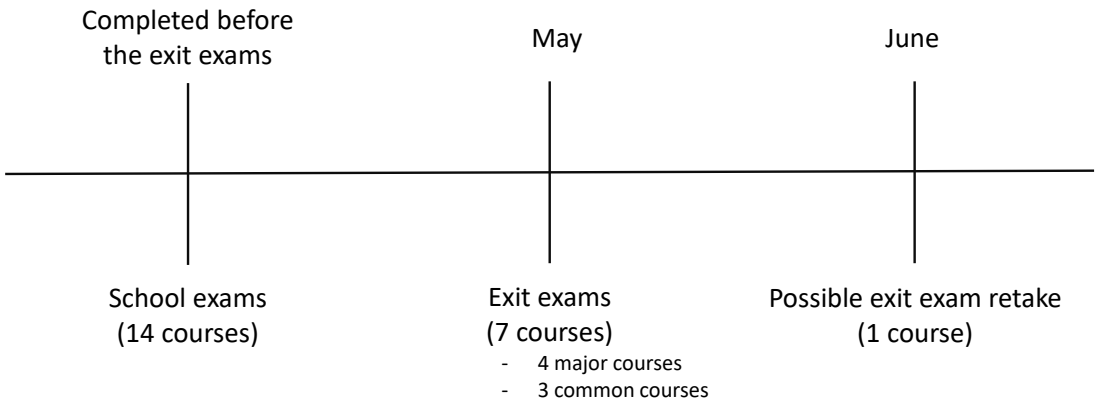
Figure 2 shows the timeline of the exams in the final year of the academic track. Students first finish their 14 school exams before they take the exit exams in May. Initially failed students can retake one ext exam in June, but only if they meet the graduation requirement upon passing that retake. Students take an exit exam in their four major courses, Dutch, English, and their elective course. The grade on the exit exam counts for 50%, the school grades count for the other 50%. The final grade in the courses without an exit exam is 100% based on the school grades. The school grades are entered into the central system WOLF before the exit exams take place (Cornelisz et al., 2019). Grades are between 1 and 10, with 5.5 being the lowest passing grade.

into professional college in case of failing that academic track. Therefore, these students are also omitted from this study

[10]For instance, students of the health major take classes in biology, physics, chemistry, and mathematics.

[11]Students in the academic track also have the option to additionally study Greek and Latin

Figure 2: Time-line of the last year of academic secondary school



Notes. All exit exam courses also have a school exam. Not all school exam courses have an exit exam. The retake only concerns the exit exam part of the course.

2.2 Grading practice

A major concern might be grade manipulation, as it would imply that passing or failing is not random. Fortunately, the exit exam grading infrastructure makes it hard to manipulate the exit exam grades. The first measure against manipulation is that the course teacher and an independent external evaluator both grade the exams. However, this is not the only reason why we can expect little manipulation. Manipulation is also hard because the conversion factor of correct answers to grades is not known at the time of grading. The teacher and the external evaluator both grade the exam and agree on the number of correct answers. Afterward, the correct answers are uploaded to a central platform before being converted into grades.[12] Therefore, the graders do not know how many extra correct answers they need to give to a student in order for him to pass. Additionally, manipulation is hard because it is not clear to the graders if a specific exam is crucial for graduating. When the teacher and external evaluator are grading an exam, they do not know the student's performance in other exams. Therefore, the graders do not know if their exam is pivotal to the student's graduation chances and so whether or not they need to help the student to pass.

The retake grades are another margin that dictates grade retention, however this margin is likely manipulated by the graders. There are two reasons why this is likely. First, the retake uses the same conversion factor as the May exams, and therefore the conversion factor is known to the graders at the time of grading the retake exam. As such, graders know exactly by how much it needs to manipulate the score in order to let a student pass. Second, at the time of the retake, it is salient that the student needs to pass the specific exam in order to graduate. If the student does not pass that exam, the student will fail and retain the year. Therefore, it will be apparent to graders if a student needs help. As retake grades are likely manipulated, I

[12] In 2006, only the first five alphabetically ranked students were uploaded. In theory, in 2006, teachers could have changed the grades that were not uploaded yet. This is unlikely as there is no difference in the distribution of grades between 2006, 2007, and 2008.

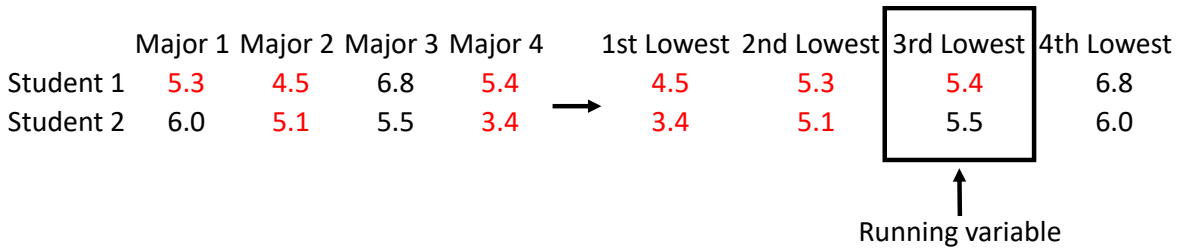
will only focus on the initial take grades. Section 4.2 confirms that the retake grades are indeed manipulated, whereas the initial exit exams in May are not.

2.3 Graduation rules and the identifying variation

The regression discontinuity design uses the good-as-random-variation between students who are allowed to do a retake or not. This variation, based on the students' initial exit exam grades, is created by the graduation rules. These rules are as follows: a student can fail only one course from its four major courses and only two courses overall. Moreover, a student can only do one retake. As an example, these rules imply that given no fails outside of its major, three failed major courses equals certain grade retention, but two failed major courses imply a retake and a possible pass. The difference between two and three failed major courses is captured by the third lowest major grade.

Figure 3 illustrates why the third lowest grade contains the identifying variation for students without non-major fails. The left panel of Figure 3 illustrates a situation with two students, one has two failed major courses and the other three, and both have no other fails. The first identification step is to order the major courses from lowest to highest, which is done in the right panel. If the third lowest grade is a just-fail (5.4), such as student 1, the first two grades are lower-equal and so are also fails, hence likely three failed major courses. On the other hand, if the third-lowest grade is a just-pass (5.5), such as student 2, the student will (likely) have two failed courses. Therefore, the third-lowest grade captures the identifying variation for students without non-major fails and is a viable running variable with 5.5 as the cutoff. For these students, this is the running variable and cutoff in the regression discontinuity design.

Figure 3: Illustration of how the running variable is located



Notes. This figure illustrates the construction of the RDD running variable for two students that have no failing grades outside of their major. The left panel shows the major grades of the students, where the failing grades (a grade less than 5.5) are indicated in red. Student 1 has three failing grades, and student 2 has two failing grades. The right panel shows the first construction step: the ordering of the grades from smallest to largest. The black box clarifies that the third-lowest grade is the marginal grade between two and three failing grades. For these students, the third-lowest grade is the RDD running variable, as two failing grades qualifies a student for a retake and three failing grades imply certain grade retention.

Which of the major grades contains the identifying variation depends on the number of failed non-major courses a student has. This is due to the requirement that students can

Table 1: Marginal grade

(1)	(2)	(3)	(4)
Retake margin	Number of failed non-major cources	Maximum number of failed major course in order to qualify for the retake	Running variable (marginal grade)
a.	0/1	2	3rd lowest grade
b.	2	1	2nd lowest grade
c.	3	0	1st lowest grade
d.	4+	Certain grade retention	-

Notes. This table shows how the running variable (the marginal grade) in the last column depends on the number of failed non-major courses, which are depicted in the second column. The number of failed non-major courses implies the maximum number of failed major courses a student can have in order to qualify for the retake. These, in turn, imply which lowest grade is the marginal grade for qualifying for the retake

only fail two courses overall. Table 1 summarizes the four different situations. The situation captured by Margin a. is as discussed above, if a student has one or zero outside non-major fails, then the third-lowest major grade contains the identifying variation and is the running variable for these students. With two failed non-major courses, Margin b., a student can only have one failed major grade in order to be admitted to the retake. In such a case, three fails overall is a retake, whereas four overall fails imply definite grade retention. In this situation, the second-lowest grade among the major grades is the difference between three and four overall fails, and therefore contains the identifying variation. Margin c., a student has three non-major fails, implies that a student cannot have any major fails in order to be allowed to do the retake. In this situation, the lowest major grade contains the identifying variation and is the running variable for these students. At last, students with four non-major fails will definitely retain and therefore do not contribute to the regression.[13]

The running variable of the RDD analysis is the major grade that contains the identifying variation and is referred to as the marginal grade. Which major grade is picked is identified by the retake margins, and these margins enter the regression as fixed effects. Still, 95.5% of students belong to Margin a., and has the third-lowest major grade as running variable.

[13]There are hardly any students with four non-major fails.

3 Data

3.1 Data sources

This paper uses data from Statistics Netherlands. Statistics Netherlands gathers administrative records from various government agencies that can be merged using unique identifiers. Data on enrollment and degrees are from the ministry of education. Degree and enrollment data are used for the reimbursement of schools and are therefore highly audited and reliable. The enrollment data contain information on the secondary educational track, major and school, and on the higher education type and field of study. The central exam and school grades are available from 2006 onward. Nonetheless, as stated before, in 2009, there was a reform of the course structure, limiting the cohorts to 2006, 2007, and 2008. Moreover, I limit the sample to students that do the central exam for the first time.[14] I can do this because I have exam registrations from 2003 onward.

The earnings from (self-)employment are from the Dutch tax authority and are available from 2011 to 2019. As students are on average 17 in the final year, I have earnings data for all three cohorts up to age 28, but I have data up to age 30 for the 2006 cohort. To compare annual earnings observed in different years, I covert the earnings into 2015 prices.

Years of experience measures the number of years of earnings in excess of full-time minimum wage. Many students do part-time work next to their studies. These experiences, for example, working in a bar, likely do not contribute to their later life earnings. Especially, workers in industries with collective bargaining agreements (such as Ph.D.'s) have a pre-specified plan where earnings increase with the number of years of relevant experience. To separate relevant experience from non-relevant experience, I count the years a person earns more than full-time minimum wage, as part-time student work likely does not pay more than this amount.

Unique to the grade retention literature, I can match students to parents. I use the de-aged income of the family in the final year of secondary school to divide families into low (33%), middle (33%), and high (33%) income groups. Moreover, I divide families into university families or not, based on whether at least one of the parents has a university degree. Finally, if at least one of the parents is born in a non-western country, I classify the family as non-western.

3.2 Summary statistics and the distribution of the running variable

Table 2 shows the summary statistics for all first-time final year academic track secondary school students in the years 2006, 2007, and 2008. The table splits the summary statistics by retaining status in the first four columns and shows the difference between retained and promoted in the last two.

Panel A shows that retainers are from a slightly less-off background. They are somewhat less likely to be from a family of high income and similarly less likely to come from a family

[14] This includes students who have done lower ability track central exams.

Table 2: Summary statistics

	Retained (R)		Promoted (P)		Difference (R-P)	
	Mean	SD	Mean	SD	Mean	SE
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Background characteristics						
High Income Family	0.270	(0.444)	0.342	(0.474)	-0.072***	(0.006)
University Family	0.144	(0.351)	0.200	(0.400)	-0.055***	(0.005)
Female	0.531	(0.499)	0.548	(0.498)	-0.011**	(0.006)
Non-western descent	0.207	(0.405)	0.090	(0.287)	0.117***	(0.004)
Age at test	17.45	(0.60)	17.23	(0.51)	0.23***	(0.01)
Panel B: Baseline						
GPA (at first try)	4.93	(1.27)	6.46	(0.77)	-1.52***	(0.01)
Marginal grade	-0.13	(0.64)	1.32	(0.76)	1.45***	(0.01)
Panel C: Outcomes						
Aca. track secondary sch. diploma	0.974	(0.158)	1	(-)	-0.024***	(0.001)
Anti-depression medication	0.022	(0.147)	0.009	(0.094)	0.013***	(0.001)
University enrollment	0.695	(0.460)	0.871	(0.335)	-0.176***	(0.040)
University degree	0.477	(0.500)	0.745	(0.436)	-0.268***	(0.005)
Earnings at 28	31,307	(22,906)	41,373	(35,015)	-10,066***	(402)
Starting earnings	29,001	(11,128)	32,073	(26,426)	-3,072***	(327)
Experience at 28	3.244	(1.794)	3.644	(8.023)	-0.685***	(0.024)
Observations	7,912		86,866			

Notes. The table only includes students that do the final year for the first time, a total of 94,778 students. Retained implies all students that ever have to re-do the final year of secondary school. Promoted indicates those that can continue education. High income implies that the student is part of a family with a joint income in the top 33 percentile. University Family implies that at least one of the parents has a university degree. Non-western descent implies that one of the parents is born in a non-western country. Anti-depression medication usage is measured in the year after the exit exams. Earnings at 28 include zero earnings. Starting earnings only include those that have started. Earnings at 28 and Starting earnings are in 2015 prices. Experience at 28 measures the years of earnings in excess of full-time minimum wage. * p<0.1, **p<0.05, and *** p<0.01.

with at least one parent with an academic degree. Moreover, they are more likely to be of non-western descent. The retainers are also less likely female but are slightly older than the non-retainers at the time of the central exam.

Panel B shows the performance of the students in the final year. None surprisingly, retained students score a lower GPA than promoted students. Still, GPA is not part of the graduation rules. But, similarly, the retained students also score a lot lower on the running variable.

Panel C shows that retainers generally have worse educational outcomes than promoted students, even after retention. Retained students have slightly fewer academic track secondary

school degrees at the age of 28 (11 years after they had to retain). In the year after the final exam, retained students are twice as likely than promoted students to use anti-depression medication. University enrollment of grade retainers is much lower than for promoted students. Similarly, there are larger differences in terms of final higher education degrees. More promoted students as retained students have a university degree. Still, the overall level is relatively high, reflecting the high ability population of the academic track.

The final set of outcomes in Panel C of Table 2 are on the labor market, which shows more considerable differences between retained and promoted students. Earnings at 28 are about 10,000 euro lower for retained students than for promoted students. However, the difference in starting earnings between retained students and promoted students is about 7000 euro smaller. At the age of 28, promoted students have about 8.4 months (0.7 years) more experience.

The differences in outcomes presented in panel C can be due to retaining but can also be due to the differences in background characteristics between the two groups. Therefore, if we want to learn about the effects of grade retention, we can not readily compare the average outcomes of promoted and retained students.

4 Research design and its validity

4.1 Research design

This paper uses the exogeneity created by the graduation rules to credibly identify the effects of grade retention on education and labor-market outcomes. Specifically, it uses the admittance policy to the retake exam in a regression discontinuity design. As explained in the institutional section, the running variable ($Grade_i$) contains the identifying variation between definite grade retention and a retake (and so a likely pass). First, I use this variation in the following reduced-form model:

$$y_i = \gamma_0 + \gamma_1 Below_i + \gamma_2 Grade_i + \gamma_3 Grade_i \times Below_i + \gamma_4 X_i + u_i$$

with outcome variable y_i for individual i. $Below_i = I(Grade_i < c)$ is a dummy variable that is equal to 1 when $Grade_i$ is lower than the cutoff. In the text $Grade_i$ is referred to as the marginal grade. X_i are the covariates that include dummy variables for cohort year, student major, and the identifying margin. I take care of any non-linear effects of $Grade_i$ on y_i by only using a small bandwidth around the cutoff. In the robustness section, I also estimate regressions with a quadratic running variable. I cluster the errors u_i by secondary school.[15] The identifying

[15] $Grade_i$ is discrete with 0.05 intervals. In such a case, Lee and Lemieux (2010) recommend clustering the error term by the intervals of the running variable. This seems intuitive as the rounding of grades will create a clustered error structure. However, Kolesár and Rothe (2018) and simulations by the author show that clustering on the running variable leads to too high rejection probabilities. If the running variable is monotone increasing (or decreasing), then clustering will mechanically lead to negative intraclass correlations between bins

assumption of this model is that the potential outcomes, conditional on $Grade_i$, $Grade_i \times Below_i$ and X_i, are continuous at the cutoff. If this is the case, γ_1 identifies the ITT.

Second, I use the identifying variation for fuzzy regression discontinuity design regressions. Fuzzy RDD's are a form of instrumental variable (IV) regressions and will be referred to as such. These regressions use the cutoff for admittance to the retake as an instrument for retaining. The first stage of the IV:

$$Retain_i = \alpha_0 + \alpha_1 Below_i + \alpha_2 Grade_i + \alpha_3 Grade_i \times Below_i + \alpha_4 X_i + v_i$$

where $Retain_i$ is a dummy variable that is equal to 1 if a student ever re-does the academic track of secondary school. The second stage is:

$$y_i = \beta_0 + \beta_1 Retain_i + \beta_2 Grade_i + \beta_3 Grade_i \times Below_i + \beta_4 X_i + e_i$$

β_1 identifies the ATE of grade retention for the compliers. The compliers are the students that pass when they are admitted to the retake, and that retain when they are not. Similar to γ_1, it is a necessary requirement for the identification of β_1 that the conditional potential outcomes are continuous at the cutoff. Still, the additional assumption of the IV approach is that all effect of scoring below the threshold is due to grade retention. For instance, failing should not lead to ability track switches or drop out.[16] The rest of the paper will both show reduced form and instrumental variable regressions.

4.2 Validity

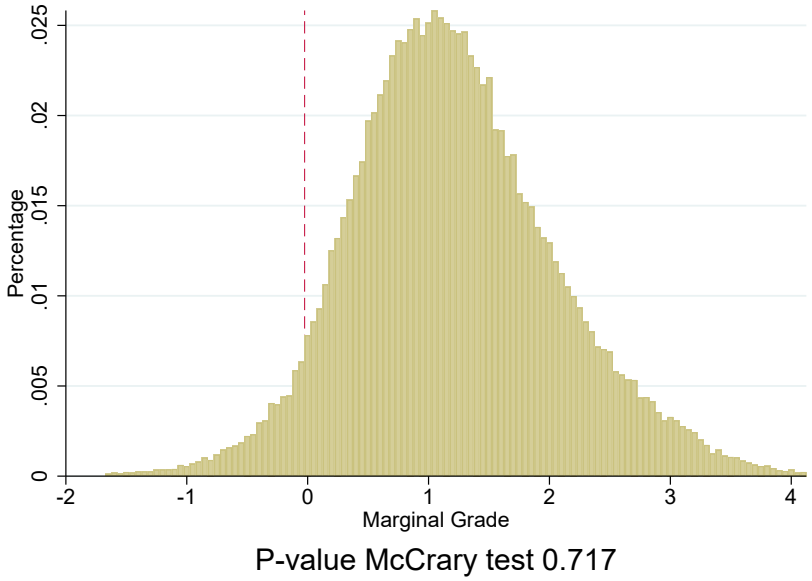
In this paper, the RDD continuity assumption translates into a requirement on the absence of manipulation by the graders, as such grade manipulation could lead to systematic sorting of students. In case of systematic sorting, we would be unsure if the jump in the outcome variables at the cutoff is due to treatment or due to sorting.

Figure 4 shows the distribution of the running variable. The running variable is constructed as explained in Section 2. I center the running variable around zero such that a value above zero implies a retake and a possible pass, and a value of the running variable below zero

(running variable intervals). Negative intraclass correlations imply that clustering reduces the standard error. This is not only true for simulations, the standard errors of nearly all regressions in this paper decrease when I cluster on the running variable. As a result, clustering on the school level, which hardly affects the standard error, is conservative.

[16] Eren et al. (2018) worry about the independent psychological or investment effect of failing on future outcomes. An "insufficient label", such as grade retention status, can lead to lower educational investment if it updates the priors of a student about the returns to investing in education. As such, it can be a co-founder of the grade retention effect. However, it is impossible to construct grade retention (or any other remedial intervention) without such an "insufficient label". Therefore, I view the "insufficient label" as an integral part of the treatment.

Figure 4: Density of the marginal grade



Notes. Each bar has an interval of 0.05 grade. Grades are normalized to be 0 at a 5.5, the passing grade. Each bar represents the percentage of students scoring the particular grade. The McCrary test p-value is estimated with the method of Cattaneo et al. (2020) and is printed below each graph. The figure is based on the cohorts take the central exam from 2006 to 2008, a total of 94,778 students.

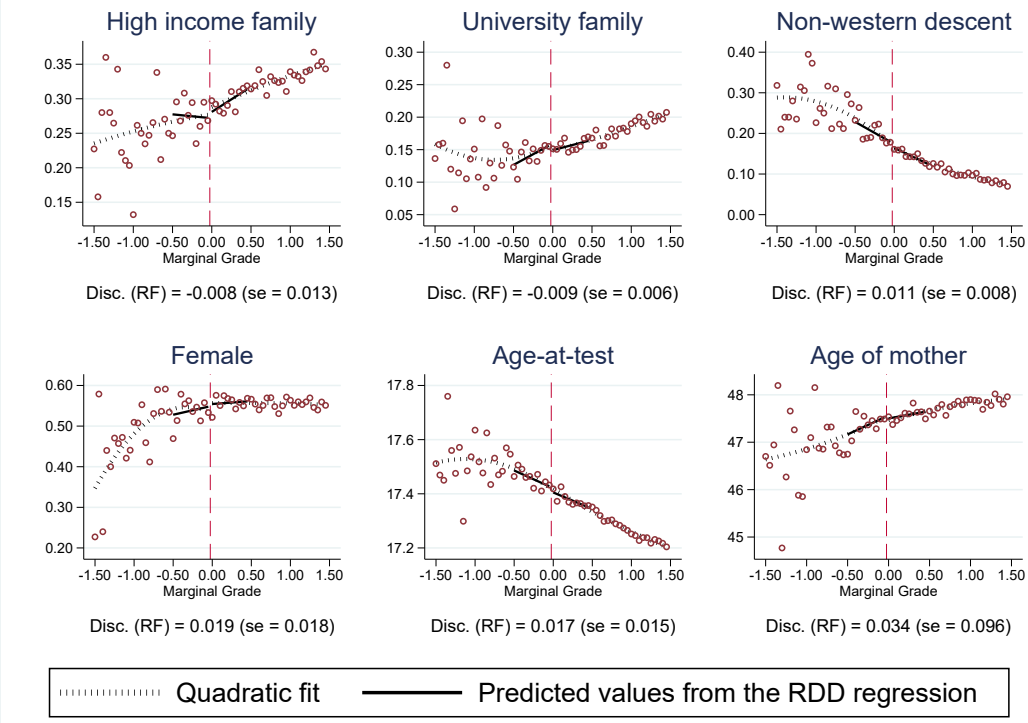
implies definite grade retention. As explained, the running variable consists out of different major grades, depending on which grade contains the identifying variation. Figure B1 in the Appendix shows the different major grades out of which the running variable is composed.

Figure 4 shows that the running variable is likely not manipulated, as it does not show any bunching around the cutoff. The McCrary test, as calculated by the method of Cattaneo et al. (2020), confirms the visual pattern as it does not reject no manipulation. The appendix Figure B2 shows the grade densities of the courses out of which the running variable is selected.

The distribution of the *retake* grades does show heavy bunching around the cutoff. The distribution is shown in Figure C1 in the Appendix. The McCrary test confirms the visual pattern. Bunching of grades means that the grading teachers decided to manipulate some students' grades such that they do not have to retain. Section 2 explains how and why this can happen. Manipulation might be justified in itself, but it is likely non-random in nature, and therefore I cannot use these grades in an RDD setting.

A more direct way to test the continuity assumption is to test the continuity of observed pre-determined characteristics. Figure 5 shows regressions where pre-determined characteristics are used as an outcome variable. Figure 5 shows that these characteristics run continuously through the cutoff and show no evidence of sorting. Additionally, the six pre-determined characteristics cannot together jointly predict the treatment. The p-value of the F-test for a joint prediction test is 0.7535. Pre-determined characteristics are unrelated to the instrument.

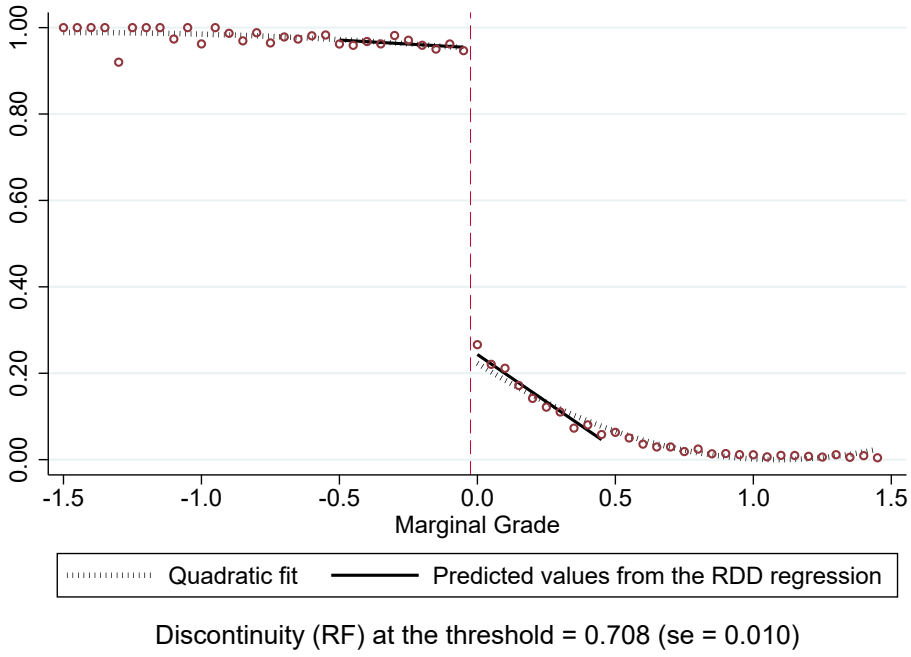
Figure 5: Balancing of pre-determined characteristics



Notes. Each panel shows an RDD plot of a pre-determined characteristics as a dependent variable. A high income family is a family with earnings in the top 33% bracket in the final year of academic secondary school. A university family, is a family of which at least one parent has a university degree. A student is of non-western descent if at least one of his parents is born in a non-western country. Age-at-test is the age at the time of taking the exit exams measured in months. The mother's age is the age of the mother in the month of the exit exams. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. Scoring below 0 implies certain grade retention, and scoring above 0 implies a retake and so possible passing. Each red circle is the average for a bin of 0.05 of a grade. The solid black lines are the predicted values from an RDD regression. The dashed line is a second-degree polynomial estimated separately, left and right of the cutoff. The reduced form RDD estimates (RF) are printed below each graph and are based on the grades that fall within a 0.5 bandwidth, below and above, the passing grade. The scatter plot uses 62,774 observations. The RDD estimate uses 15,965 observations.

Figure 6 shows a graphical representation of the first stage. Just scoring below the cutoff implies an increase in the probability of retaining by 71.1 percentage points compared to scoring just above. The effect is not a 100 percentage points because students that score above the cutoff merely qualify for the retake and can still fail this retake. Nonetheless, it seems that the vast majority passes. Some students that do not qualify for the retake still manage to progress into higher education. There are two reasons why we have these never-takers. First of all, for a select group of students it is not clear what grades are major courses and which are non-major courses, or even extra courses. This leads to slight miss-specification of the central exam grades set and so of the running variable.[17] Another reason is that if a student fails based on a 100% school grade course, it can apply for a special summer treatment and still pass.[18]

Figure 6: First stage: the probability of ever retaining the last year of academic secondary education

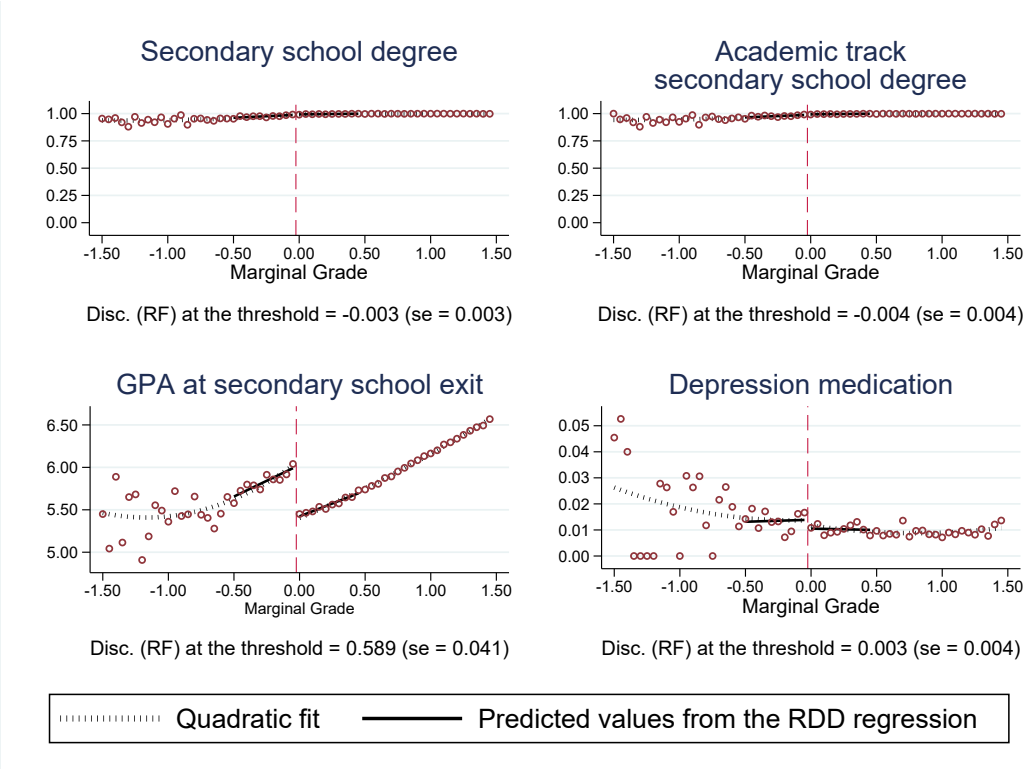


Notes. The figure shows the first stage of the IV: whether a student has to re-do the final year of secondary school at any point in time after the central exam and can only enter higher education at least a year later. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. Scoring below 0 implies failing, and scoring above 0 implies a retake and so possible passing. Each red circle is the average for a bin of 0.05 of a grade. The solid black lines are the predicted values from the RDD regression. The dashed line is a second-degree polynomial estimated separately, left and right of the cutoff. The RDD estimate printed below the graph is based on the grades that fall within a 0.5 bandwidth, below and above, the passing grade. The scatter plot uses 62,774 observations. The RDD estimate uses 15,965 observations.

[17]This misspecification only happens for students in the humanities major. In contrary to the other majors, they have multiple options in choosing major courses. Especially, if they take more courses as required, they can ex-post shift around courses in and out of their major such that they can pass.

[18]In Dutch the name for this treatment is the "October regeling".

Figure 7: Short-run outcomes



Notes. The upper left panel shows the percentage of students that have a secondary school diploma at the age of 28. The upper right panel shows the percentage of students that have academic track secondary school diploma. The lower left panel shows the grade point average (GPA) after the retaining students retook the central exam. The GPAs of students that retain in the transition centers are missing. The GPAs left of the cutoff are from students that re-do the last year of secondary school at their old school. Right of the cutoff shows the grades of all promoted students. The lower left panel shows the percentage of students that take doctor-prescribed anti-depression medicine. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. Scoring below 0 implies certain grade retention, and scoring above 0 implies a retake and so possible passing. Each red circle is the average for a bin of 0.05 of a grade. The solid black lines are the predicted values from the RDD regression. The dashed line is a second-degree polynomial estimated separately, left and right of the cutoff. The RDD estimate printed below the graph is based on the grades that fall within a 0.5 bandwidth, below and above, the passing grade. The scatter plot uses 62,774 observations. The RDD estimates use 15,965 observations.

5 The effects of grade retention

Grade retention can affect multiple outcomes in the short and long run. In the short run, there might be an effect on secondary school diplomas, GPA, and students' mental health. In the long run, the remedial effect of grade retention could increase educational attainment. Retaining students should graduate later and so enter the labor market at a later point, leading to a lower level experience at any age. The impact on earnings can be positive if grade retention increases human capital (through educational attainment) or negative if the effect due to lower experience outweighs the educational attainment effect.

5.1 Short-run outcomes

Secondary school graduation is not linked to grade retention, as the upper left panel of Figure 7 shows that nearly everybody obtains a secondary school diploma. The upper right panel of

Figure 7 shows that not everybody obtains an academic track secondary school diploma. Still, the cutoff is not related to the probability of getting such a diploma. The two panels together show that the consequence of scoring too low is retaining and not switching to a less demanding ability track or dropout.[19]

The GPA of students that re-do the final year in their old secondary school increases by 0.84 points or 22 percentiles, as shown in the lower left panel of Figure 7. Unfortunately, only the grades of students that retain in their old school are available. Nonetheless, it shows that grade retention must look powerful to experienced teachers.

The lower right panel of Figure 7 investigates mental health drug prescriptions in the year after the first exit exam. The cutoff is not related to the percentage of students that get the anti-depression drug prescribed. Thus, it does not seem that retention has an impact on mental health.

5.2 Long-run educational outcomes

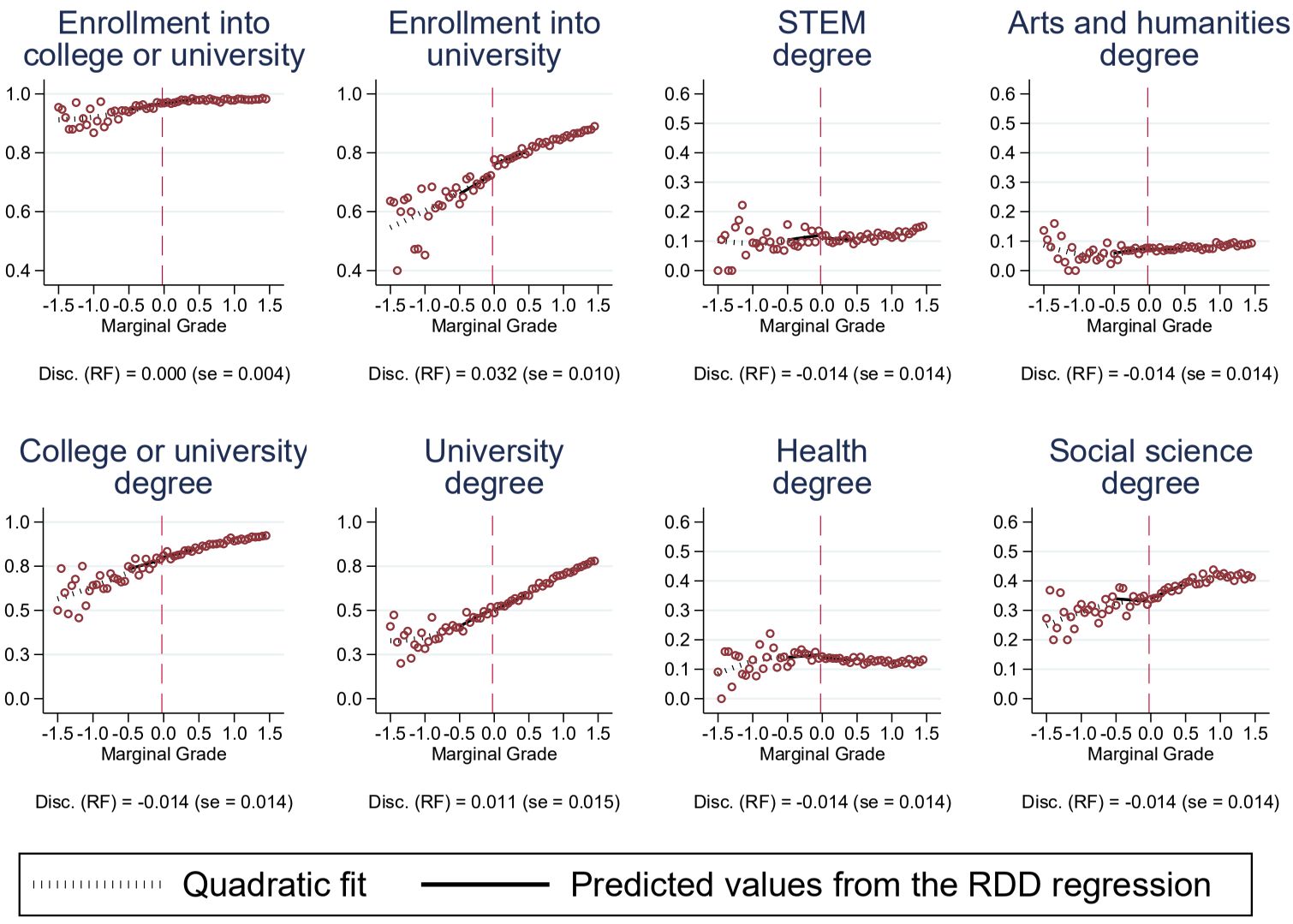
The first available post-secondary school outcome is the enrollment into college and university. The upper left panel in the first column of Figure 8 shows that enrollment into college and university is near-universal. Moreover, there are no differences between grade retainers and promoted students. There are differences between promoted and retained in the probability of entering university, as shown in the upper panel of the second column. Grade retainers have a 4.5 percent point lower probability (IV estimate) of ever going to university.

Despite the difference in university enrollment, there is no difference in educational university degrees between retained and promoted, nor is there a difference in terms of college or university degrees. The lower panels of the first and second column of Figure 8 respectively plot the college or university degrees and the university degrees. Both these panels use educational attainment at age 28, the oldest age for which we observe degree data for all cohorts.

Table 3 confirms that grade retention does not lead to differences in degree level. Table 3 is split up into two parts. The upper part shows the regression estimates for whether or not a student receives a university degree. The lower part shows the regression estimates for whether or not a student gets a college or university degree. The first four columns of Table 3 show estimates of parametric regressions that differ in bandwidths. The last two columns show the estimates of non-parametric regressions that use a triangular kernel. Column (6) displays the results of the data-driven non-parametric regression that is estimated using the method of

[19] Figure E1 in the Appendix shows what students do in the year after their first exit exams. The upper left panel of Figure E1 is very similar to the first stage, indicating that most students immediately re-do the last year of secondary school but, the lower left panel shows that some (still) re-do the last year of secondary school two years after failing. The second column of Figure E1 shows that hardly anybody switches to vocational education, consistent with the absent impact of grade retention on academic diplomas. The third column indicates that the majority of promoted students start higher education directly after secondary school. The lower part of column three shows that most retained students join their promoted counterparts a year later. Finally, the last column of Figure E1 shows that some students take a gap year. In the first year after their first exit exam, more promoted students take a gap year than retained students. But, in the second year, this is reversed.

Figure 8: Enrollment, degrees, and field of study



Notes. The upper left panel shows the probability of ever enrolling in college or university, the lower right panel shows the probability of obtaining a college or university degree at the age of 28. The second column shows the probability of ever attending a university in the upper panel and getting a university degree in the lower panel. The third column shows the probability of graduating with a STEM degree, college and university degrees combined in the upper panel, and the probability of graduating with a Health degree in the lower panel. Health degrees include medical school degrees. The last column shows the probability of graduating with a humanity or social science degree in the upper and lower panels. The cohort of students considered did their (first) final year in 2006, 2007, or 2008. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. Scoring below 0 implies failing, and scoring above 0 implies a retake and so possible passing. Each red circle is the average for a bin of 0.05 of a grade. The solid black lines are the predicted values from the RDD regression. The dashed line is a second-degree polynomial estimated separately, left and right of the cutoff. The reduced form RDD estimate printed below the graph is based on the grades that fall within a 0.5 bandwidth, below and above, the passing grade. The scatter plots use 62,774 observations. The RDD estimates use 15,965 observations.

Table 3: Obtained higher education degrees at age 28 - IV estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Fraction with a university degree			
Retain	0.016	0.015	0.018	0.038	0.018	0.033	0.021
	(0.020)	(0.025)	(0.031)	(0.037)	(0.025)	(0.036)	(0.031)
Obs.	26,468	15,669	11,731	6,995	15,669	11,731	15,669
				Fraction with a college or university degree			
Retain	-0.003	-0.020	-0.012	-0.029	-0.014	-0.017	-0.017
	(0.016)	(0.021)	(0.027)	(0.032)	(0.020)	(0.030)	(0.025)
Obs.	26,468	15,669	9,988	6,995	15,669	9,988	15,669
Band-width	+/- 0.75	+/- 0.50	Opt.	+/- 0.25	+/- 0.5	Opt	+/- 0.50
Pre-det. Char.					✓		
Est. Tech.	Par.	Par.	Par.	Par.	Par.	N-Par.	N-Par.

Notes. All columns include major, year, and cutoff margin fixed effects. Opt. implies using the bandwidth chosen by the mean-squared error optimal bandwidth algorithm. Par. implies parametric or OLS regressions. Non-P. implies non-parametric regressions that use a triangular kernel. The non-parametric regression in column (6) is estimated using the data-driven methods of Calonico, Cattaneo and Tituinik, 2014. The standard errors are displayed between the brackets and are clustered on school basis. * p<0.1, **p<0.05, and *** p<0.01.

Calonico, Cattaneo, and Titiunik (2014). This regression first finds the optimal grade point bandwidth by trading-off regression bias with an increase of variance and afterward estimates non-parametric local linear regressions using the triangular kernel. The estimates are stable over all the columns, but the standard error is bigger for the smaller bandwidths, indicating a decrease in power. In column (5), I include pre-determined characteristics into the regression model. The inclusion of these characteristics does not change the estimates in a meaningful way compared to column (2), again validating my approach.[20]

It could be that 11 years is not long enough after the exit exams to show a difference in degrees. This is unlikely, as Figure F2 in the Appendix shows that there are no differences in enrollment between retainers and promoted students in 2019. Overall, it seems that there are no differences in educational degree level between retainers and non-retainers at the margin.

Grade retention also has no impact on the field of study in tertiary education. The upper panels of the third and fourth column in Figure 8 show the impact of grade retention on

[20]Further robustness checks are provided in Appendix I.

graduating with a stem degree or a arts and humanities degree. There is no impact on both. The lower panels of these columns show the impact of graduating with a health or social science degree, which also show no effect of grade retention. Overall there is no impact of retaining on the field of study. This is somewhat surprising as students for a large extent fail on the basis of their major courses, which should serve as preparatory courses for their field of study in tertiary education.

There is one crucial educational difference: promoted students are faster to finish education than grade retainers. The upper left panel of figure 9 shows the time in education since the moment that students could retain. The IV estimate for the time difference is 0.5 years or six months. As such, the difference in time is not 12 months. The upper right panel shows that retained students are faster to finish higher education when they eventually start, gaining 0.4 years. The lower left panel shows that this is because they are more likely to finish what they start. This is mainly due to promoted students switching from university education to professional college education, leading to delays. Lastly, depicted in the lower right panel of Figure 9, promoted students take on average about 0.1 more gap years than retained students. Thus, although retained students make up some time during higher education, they end up with a delay of about six months.

5.3 Earnings

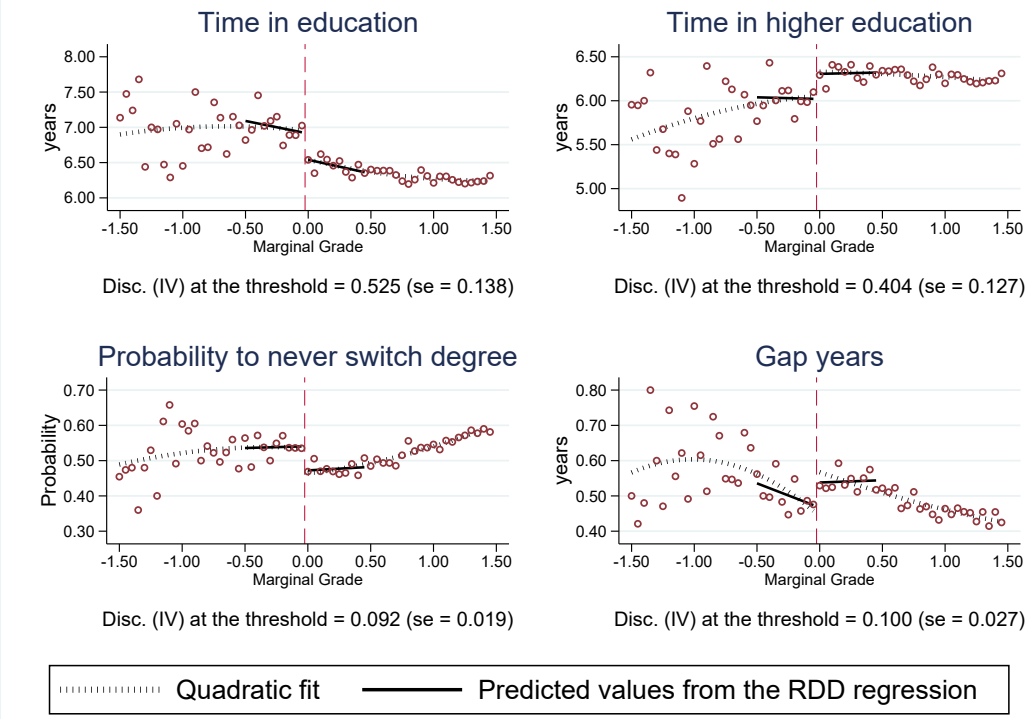
Figure 10 shows a statistically and economically significant earnings penalty for retained students at the age of 28. Table 4 presents IV results that confirm the significant difference in earnings. The first four columns of the upper panel show that the regression results are very stable across different grade point bandwidths. These columns show that students earn about 3000 euro less than promoted students at the age of 28, this is an 8.5% decrease.[21] Column (5) includes pre-determined characteristics into the regression. The results are similar to the regression that excludes the pre-determined characteristics but with the same bandwidth in column (2). This similarity of results again validates my approach. The last two columns use a non-parametric regression technique with a triangular kernel. These results are slightly larger than the parametric regression results.[22]

The earnings penalty is not limited to the age of 28. Figure 12 plots the different IV estimates for different ages using a bandwidth of $\pm$ 0.50. The penalty increases until the age of 27 and is slightly lower for the age of 28. The estimates of the earnings penalty for ages 29 and 30 are in line with the estimates for earlier ages. Yet, the estimates for these ages are based on only a part of the sample as not everybody has reached these ages in 2019 (the last year of available tax records). Still, the earnings penalties for ages 29 and 30 indicate that the earnings penalty is here to stay.

[21] Students that have a passing grade (a 5.5) as their marginal grade earn on average 35.224,82 euro

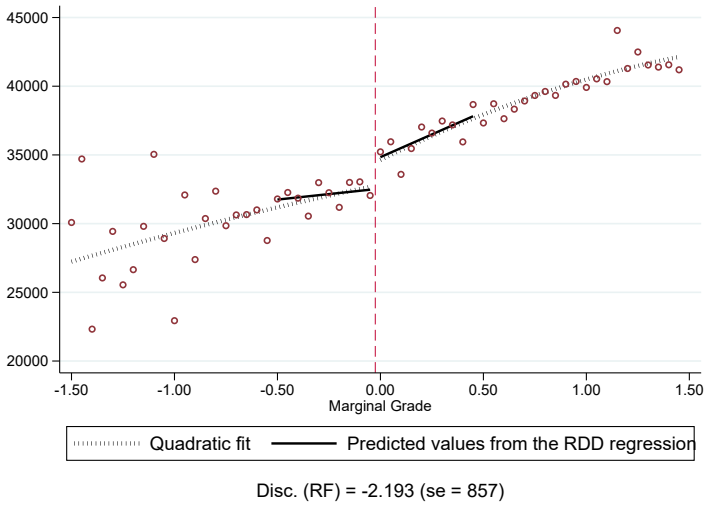
[22] Further robustness checks are provided in Appendix I.

Figure 9: Time in education, program switches and gap years



Notes. The upper left panel shows the total time spent in education since the exit exams in either 2006, 2007, or 2008, as measured by 2019. Students still in education get the maximum amount of time. The upper right panel shows the time spent in higher education alone. The lower left panel shows the probability of finishing the first higher education degree a student started. The lower right panel shows the gap years students take. A gap year is a year of non-enrollment between years of enrollment. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. Scoring below 0 implies certain grade retention, and scoring above 0 implies a retake and so possible passing. Each red circle is the average for a bin of 0.05 of a grade. The solid black lines are the predicted values from the RDD regression. The dashed line is a second-degree polynomial estimated separately, left and right of the cutoff. The fuzzy-RDD (IV) estimate printed below the graph is based on the grades that fall within a 0.5 bandwidth, below and above, the passing grade. The scatter plot uses 62,774 observations. The RDD estimate uses 15,965 observations.

Figure 10: Earnings at 28



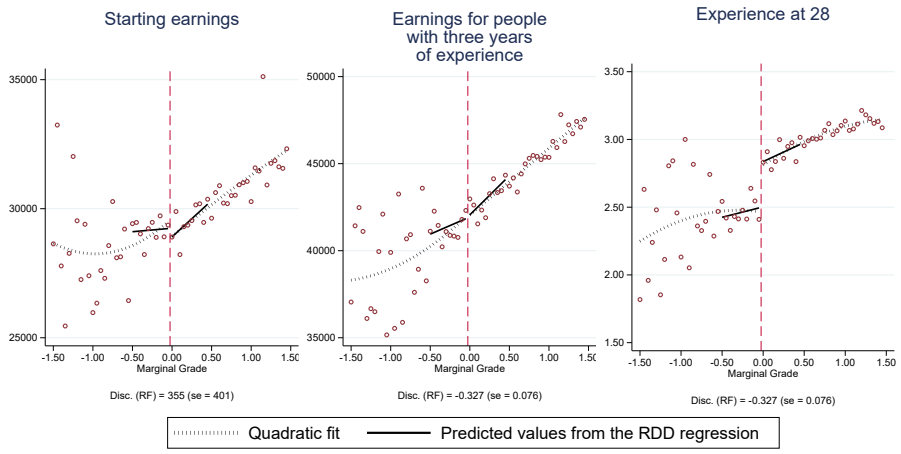
Notes. Earnings are either from employment or self-employment. Moreover, earnings are deflated and include zero earnings. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. Scoring below 0 implies failing, and scoring above 0 implies a retake and so possible passing. Each red circle is the average for a bin of 0.05 of a grade. The solid black lines are the predicted values from the RDD regression. The dashed line is a second-degree polynomial estimated separately, left and right of the cutoff. The reduced form RDD estimate printed below the graph uses a 0.50 grade point bandwidth, below and above, the passing grade. The scatter plot uses 62,774 observations. The RDD estimate uses 15,965 observations.

Table 4: Earnings at 28 - IV estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Retain	-2,894***	-3,097***	-3,314**	-3,110*	-2,910***	-3,636**	-3,349***
	(937)	(1,207)	(1,508)	(1,629)	(1,178)	(1,594)	(1,319)
Obs.	26,357	15,604	9,951	6,967	15,604	9,951	15,604
Band-width	±0.75	±0.50	Opt.	±0.25	±0.50	Opt	±0.50
Pre-det. Char.					✓		
Est. Tech.	Par.	Par.	Par.	Par.	Par.	N-Par.	N-Par.

Notes. All columns include major, year, and cutoff margin fixed effects. Opt. implies using the bandwidth chosen by the mean-squared error optimal bandwidth algorithm. Par. implies parametric or OLS regressions. Non-P. implies non-parametric regressions that use a triangular kernel. The non-parametric regression in column (6) is estimated using the data-driven methods of Calonico, Cattaneo and Titunik (2014). The standard errors are displayed between the brackets and are clustered on school basis. * p<0.1, **p<0.05, and *** p<0.01.

Figure 11: Earnings conditional on experience, and experience itself



Notes Starting Earnings are the earnings in the first year a person earns in excess of full-time minimum wage. Experience measures the number of years in which earnings are in excess of full-time minimum wage. The middle panel, earnings conditional three years of experience, does not include people without three years of experience. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. Scoring below 0 implies failing, and scoring above 0 implies a retake and so possible passing. Each red circle is the average for a bin of 0.05 of a grade. The solid black lines are the predicted values from the RDD regression. The dashed line is a second-degree polynomial estimated separately, left and right of the cutoff. The RDD estimate printed below the graph is based on the grades that fall within a 0.5 bandwidth, below and above, the passing grade. The scatter plot uses 61,728 observations. The RDD estimate uses 15,682 observations.

Figure F1 in the Appendix shows a similar age plot as Figure 12 but then for the difference in hours. The pattern in earnings seems initially due to a difference in hours, reflecting the later start of grade retainers. Until the age of 28, retained students work less than promoted students, as a larger fraction of the retained is still in education. At the age of 28, Figure F1 shows that there is no longer a significant difference in hours between retained and promoted students. The phaseout of the difference in hours is simultaneous to the phaseout of the difference in enrollment. The left panel of Figure F2 shows a difference in educational enrollment for age 27, but the right panel shows no difference in enrollment for age 28. This phaseout in enrollment explains why the earnings penalty at age 27 is larger than the penalty at age 28. Still, the absence of a difference in hours at age 28, an age at which there is still a large earnings penalty, indicates an enduring difference in wages between retained and promoted students. Such an enduring wage effect could be due to two reasons. First, the wage effect could be due to employer discrimination. Second, the effect could be due to grade retainers having less experience, and therefore fewer earnings in each given year.

The right panel of Figure 11 shows that students that start working have similar earnings, making discrimination or signaling effects of grade retention an unlikely explanation. The working start is defined as the year wherein a student earns more than full-time minimum wage for the first time. The implied labor-market starting age corresponds closely to the graduation age. Figure F1 shows that there is no difference in hours in this first working year. Therefore, it seems that promoted and retained students have similar start-up wages. These similar wages

imply that employers do not use grade retention as a signal. Employers can learn grade retention status by looking at the enrollment data that typically feature on Dutch resumés. The absence of a signaling value of grade retention is consistent with studies by Baert and Picchio (2021) and Di Stasio and van de Werfhorst (2016).

The difference in experience can explain the difference in earnings. The reduced form difference in experience at the age of 28 is -0.31 year, see the right panel of Figure 11. The accompanying instrumental variable estimate is -0.46 or 5.5 months. These differences are consistent with the differences in graduation times. The Appendix G shows that 5.5 months of experience is valuable. Figure G1 plots the experience earnings profiles for people with 6 months of experience, where experience is defined in the same way as in the rest of the paper. The figure shows an earnings-experience profile that is quite steep. Moreover, moving from 2.5 years of experience to 3 years of experience increases earnings by 2,857 euro, which is consistent with the findings above.[23]

The middle panel of Figure 11 plots the reduced form effects for people with three years of experience. It does not show any jump at the cutoff, indicating that earnings are similar when experience is similar. This is a strong indication that the retained and promoted students have similar earnings profiles, only the retained students are set back for a year. Based on Oosterbeek et al. (2021) it is likely that the effect of the delay on earnings persists until the age of 40.[24]

5.4 The treatment effects for different complier groups and infra-marginal students

The IV regressions identify effects for the compliers at the cutoff. However, it is of considerable interest to know whether the effects of grade retention are similar for different complier groups and for different cutoff points. First, this gives us an indication of the total effect of grade retention. Second, if the treatment effects are local to the compliers at the cutoff, we can mitigate any adverse effects by changing the cutoff. Alternatively, if the effects are non-local, it implies that the RDD effects apply to a wider group of students.

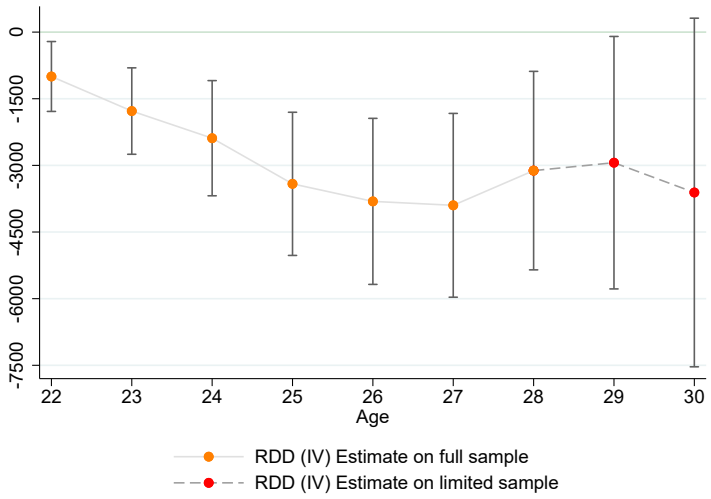
The Bertanha and Imbens (2020) test indicates an equality of outcome distributions between complier types, indicating external validity of the complier results to never and always takers. Bertanha and Imbens (2020) propose to assess this external validity using an augmented Hausman test. This test of the equality of outcome distributions has two parts. The first part compares the outcome distributions of the treated compliers (E[Y|$Grade_i \uparrow$, $Retain_i$=1]) with the always-takers (E[Y|$Grade_i \downarrow$, $Retain_i$ =1]).[25] In my application, these are those who are

[23] Figure G2 show profiles that are estimated using the canonical mincer regression framework. The figure also shows similar steep earnings-experience profiles.

[24] Oosterbeek et al. (2021) document a delay in labor-market entry due to school-starting age in the Netherlands, which also leads to lower levels of experience at any given age. In their paper, the lower level of experience causes the earnings of the affected to be lower until the age of 40.

[25] $Grade_i \uparrow$ denotes approaching the cutoff from below and $Grade_i \downarrow$ denotes approaching the cutoff from

Figure 12: RDD estimates of the earnings difference for the ages 22 to 30



Notes. Each dot is the estimate of a separate RDD earnings regression that uses a 0.50 grade point bandwidth, below and above, the passing grade. The vertical lines indicate the confidence intervals. Earnings are either from employment or self-employment. Moreover, earnings are deflated and include zero earnings.

retained due to scoring too low and those who failed the retake. The second leg compares the outcome distributions between untreated compliers (E[Y|$Grade_i \downarrow$, $Retain_i$ =0]) and never-takers (E[Y|$Grade_i \uparrow$, $Retain_i$=0]). In my case, these groups consist respectively out of those who pass on the retake and those who are promoted even though they scored too low.

The upper part of Table 5 shows that different complier groups have the same potential outcomes. The joint F-tests for equality of complier types are never close to statistical significance across the four columns. This indicates that we cannot reject that the compliers are the same as always-takers and never-takers. Bertanha and Imbens (2020) argue that this exogeneity result actually implies external validity to other cutoff points. According to them, the "independence between compliance types and potential outcomes implies independence between treatment participation and potential outcomes."

A more direct way to assess the external validity to other cutoff points is the method of Dong and Lewbel (2015). Dong and Lewbel (2015) propose to use the treatment effect derivative (TED) to extrapolate the treatment effects to these other points. If changing the cutoff does not affect the (difference) in potential outcomes, the TED equals the marginal threshold treatment effect (MTTE). Using the TED we can find new treatment effects, $\beta_1(c^{new})$, using a Taylor expansion. Given c is the current 5.5 cutoff, and c^{new} is the new cutoff and $\beta_1'(c)$ the treatment effect derivative (TED), we get the following:

above

$$\beta_1(c^{new}) \approx \beta_1(c) + (c^{new} - c)\beta_1'(c)$$

A TED equal to zero indicates the absence of treatment heterogeneity around the cutoff. This intuition is depicted in Figure D1 of the Appendix. If the treatment effects are homogeneous, the potential outcomes are parallel, such as in the right panel of Figure D1, and the TED will be close to zero. On the other hand, when the treatment effects are heterogeneous, such as the left panel of Figure D1, the TED at the current cutoff will be different from zero, and it is likely that treatment effects at other cutoff points are different. Dong and Lewbel (2015) show how to estimate the TED.

The lower part of Table 5 shows the results of Dong and Lewbel (2015) extrapolation exercise on the reduced form regressions. The estimated TED for all outcomes is small. As a result, a small decrease of the threshold does not lead to different treatment effects. The last row of Table 5 shows that we cannot extrapolate forever. When the extrapolation step is large, in this case one whole point, the standard errors become too large to be informative. In essence, this analysis says that the difference in potential outcomes is at least similar around the cutoff, and adverse effects for marginally affected students cannot be reduced by reducing the cutoff. Of course, the total effect would be reduced when we reduce the cutoff.[26]

[26]My preferred interpretation of lowering the cutoff on the marginal grade is that it is an approximate of slightly leaner rules or grading.

Table 5: Exogeneity, external validity and extrapolation

	University degree	Earnings at 28	Starting earnings	Experience at 28
	Bertanha and Imbens (2020) exogeneity test			
E[Y|$Grade_i \downarrow$, $Retain_i$ =1]	-0.011	2,462	3,373	-0.074
- E[Y|$Grade_i \uparrow$, $Retain_i$=1]	(0.105)	(4,603)	(2,962)	(0.578)
E[Y|$Grade_i \downarrow$, $Retain_i$ =0]	-0.047	250	-1253	0.238
- E[Y|$Grade_i \uparrow$, $Retain_i$=0]	(0.035)	(1,594)	(830)	(0.152)
Joint F-test	1.814	0.299	3.545	2.453
p-value	0.403	0.861	0.170	0.293
	Marginal threshold treatment effect (Dong and Lewbel, 2014)			
$\gamma_1(0)$	0.011	-2,142***	284	-0.341***
	(0.015)	(695)	(371)	(0.067)
TED	-0.044	-653	-922	-0.066
	(0.039)	(1,871)	(989)	(0.184)
$\gamma_1(-0.1)$	0.016	-2,067***	376	-0.334***
	(0.014)	(628)	(336)	(0.059)
$\gamma_1(-0.25)$	0.023	-1,978***	514	-0.324***
	(0.014)	628	334	0.057
$\gamma_1(-1)$	0.056	-1,489	1,206	-0.274*
	(0.036)	(1,659)	(875)	(0.158)

Notes. All columns include major, year, and cutoff margin fixed effects. The estimates use a 0.75 point interval. $Grade_i \uparrow$ implies approaching the cutoff from below, and $Grade_i \downarrow$ implies approaching the cutoff from below. $\gamma_1(X)$ is the estimate of the reduced form at X relative to the cutoff. The standard errors are displayed between the brackets and are clustered on school basis. * p<0.1, **p<0.05, and *** p<0.01.

6 Conclusion

Grade retention is a common educational practice that has direct labor market impacts as it delays labor-market entry and extends learning by one year. This is the first paper to evaluate these opposing labor market impacts of grade retention. Additionally, the combination of the data, identification strategy, and setting, let the paper overcome all the identification problems that are specific to grade retention.

The first key result is that (late) grade retention does not affect educational degrees of academic students, indicating no educational advantages of grade retention. The second key result is that grade retention comes with a large annual earnings penalty, indicating little overall benefit for retaining those who have to retain.

The earnings penalty of grade retention is likely due to the delay in academic progression, which leads to later labor-market entry and less experience at any given age. Other remedial interventions (summer school or after-school instruction time) do not delay students and could therefore be a more attractive instrument against learning deficiencies (Lavy et al., 2018; Pyne, Messner, & Dee, 2020).

An important limitation of this study is that it cannot rule out positive grade retention effects at lower grade levels or lower ability levels. In fact, this limitation should be an important avenue for future research. Additionally, it is important to keep in mind that grade retention can have an incentive effect on students.[27] Students might not like to retain, and therefore they work harder. Nonetheless, the design in this study is not affected by the incentive effect of grade retention, whether it exists or not.

References

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational evaluation and policy analysis*, *31*(4), 480–499.

Baert, S., & Picchio, M. (2021). A signal of (train) ability? grade repetition and hiring chances. *Journal of Economic Behavior & Organization*, *188*, 867–878.

Belot, M., & Vandenberghe, V. (2014). Evaluating the 'threat' effects of grade repetition: exploiting the 2001 reform by the French-speaking community of Belgium. *Education Economics*, *22*(1), 73–89.

Bertanha, M., & Imbens, G. W. (2020). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, *38*(3), 593–612.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal*, *14*(4), 909–946.

[27] Belot and Vandenberghe (2014) do not find evidence of an incentive effect of grade retention in Belgium, but Koppensteiner (2014) does find such an effect in Brazil.

Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, *115*(531), 1449–1455.

Clark, D., & Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy*, *122*(2), 282–318.

Cockx, B., Picchio, M., & Baert, S. (2019). Modeling the effects of grade retention in high school. *Journal of Applied Econometrics*, *34*(3), 403–424.

Cornelisz, I., Meeter, M., & van Klaveren, C. (2019). Educational equity and teacher discretion effects in high stake exams. *Economics of Education Review*, *73*, 101908.

Diamond, R., & Persson, P. (2016). *The long-term consequences of teacher discretion in grading of high-stakes tests* (Working Paper No. 22207). National Bureau of Economic Research.

Di Stasio, V., & van de Werfhorst, H. G. (2016). Why does education matter to employers in different institutional contexts? A vignette study in England and the Netherlands. *Social Forces*, *95*(1), 77-106.

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

Eide, E. R., & Showalter, M. H. (2001). The effect of grade retention on educational and labor market outcomes. *Economics of Education Review*, *20*(6), 563–576.

Eren, O., Lovenheim, M. F., & Mocan, N. H. (2018). *The Effect of Grade Retention on Adult Crime: Evidence from a Test-Based Promotion Policy* (NBER Working Papers No. 25384). National Bureau of Economic Research.

Ferreira Sequeda, M., Golsteyn, B., & Parra Cely, S. (2018). *The effect of grade retention on secondary school performance: Evidence from a natural experiment* (Tech. Rep.). Maastricht University, Research Centre for Education and the Labour Market (ROA).

Figlio, D., & Özek, U. (2020). An extra year to learn english? early grade retention and the human capital development of english learners. *Journal of Public Economics*, *186*, 104184.

Fruehwirth, J. C., Navarro, S., & Takahashi, Y. (2016). How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects. *Journal of Labor Economics*, *34*(4), 979–1021.

Gary-Bobo, R. J., Goussé, M., & Robin, J.-M. (2016). Grade retention and unobserved heterogeneity. *Quantitative Economics*, *7*(3), 781–820.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, *86*(1), 226–244.

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, *1*(3), 33–58.

Jepsen, C., Mueser, P., & Troske, K. (2016). Labor market returns to the ged using regression discontinuity analysis. *Journal of Political Economy*, *124*(3), 621–649.

Kolesár, M., & Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, *108*(8), 2277–2304.

Koppensteiner, M. F. (2014). Automatic grade promotion and student performance: Evidence from brazil. *Journal of Development Economics*, *107*(C), 277–290.

Landaud, F., & Maurin, E. (2020). *Aim high and persevere! competitive pressure and access gaps in top science graduate programs* (Working Papers). HAL.

Lavy, V., Kott, A., & Rachkovski, G. (2018). *Does Remedial Education at Late Childhood Pay Off After All? Long-Run Consequences for University Schooling, Labor Market Outcomes and Inter-Generational Mobility* (NBER Working Papers No. 25332). National Bureau of Economic Research, Inc.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Machin, S., McNally, S., & Ruiz-Valenzuela, J. (2020). Entry through the narrow door: The costs of just failing high stakes exams. *Journal of Public Economics*, *190*, 104224.

Manacorda, M. (2012). The cost of grade retention. *Review of Economics and Statistics*, *94*(2), 596–606.

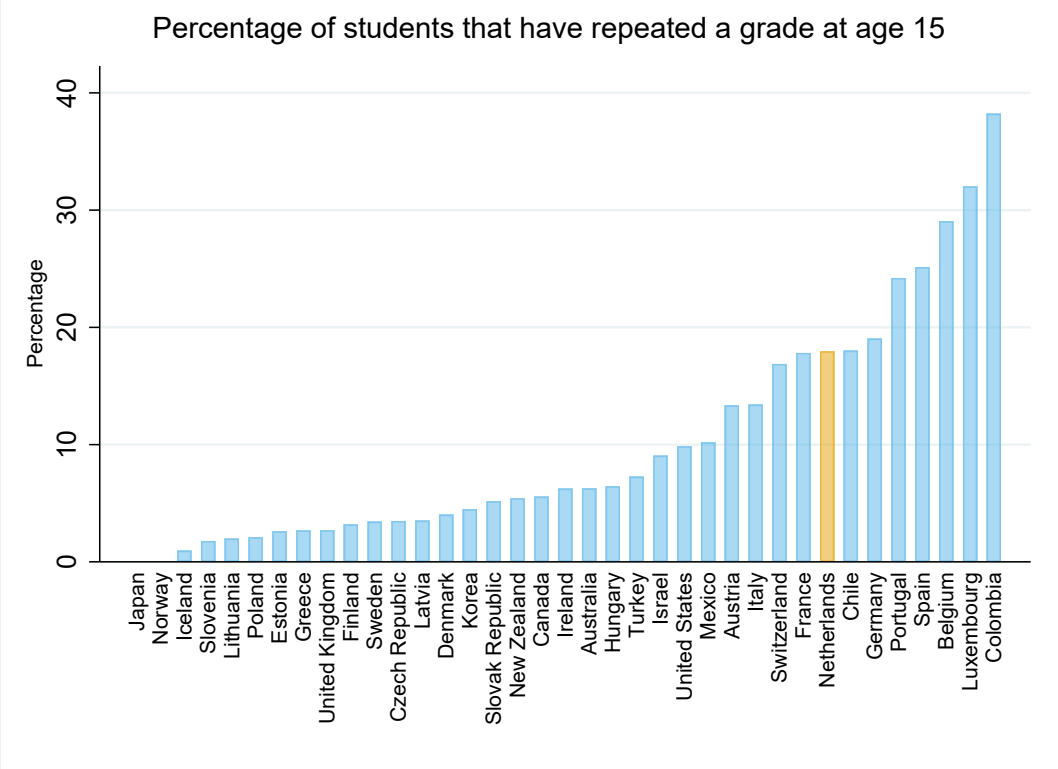
Oosterbeek, H., ter Meulen, S., & van der Klaauw, B. (2021). Long-term effects of school-starting-age rules. *Economics of Education Review*, *84*, 102144.

Pyne, J., Messner, E., & Dee, T. S. (2020). The dynamic effects of a summer learning program on behavioral engagement in school. cepa working paper no. 20-10. *Stanford Center for Education Policy Analysis*.

Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from florida. *Journal of Public Economics*, *152*, 154–169.

Appendix A

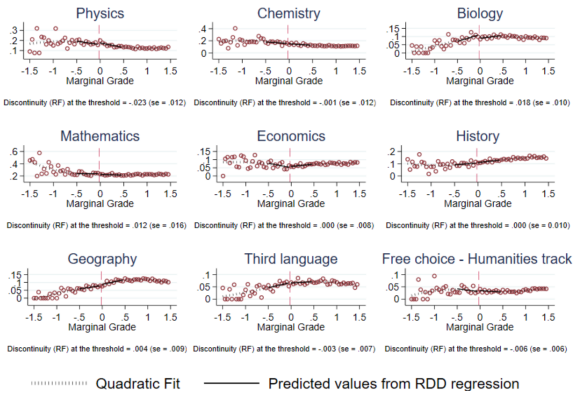
Figure A1: Percentage grade retainers per country



Notes. Each bar represents the percentage of students that declare themselves as a grade-retainer at the age of 15 per country. The orange bar represents the percentage of retained students in the Netherlands. The data is from Pisa 2018 on OECD countries.

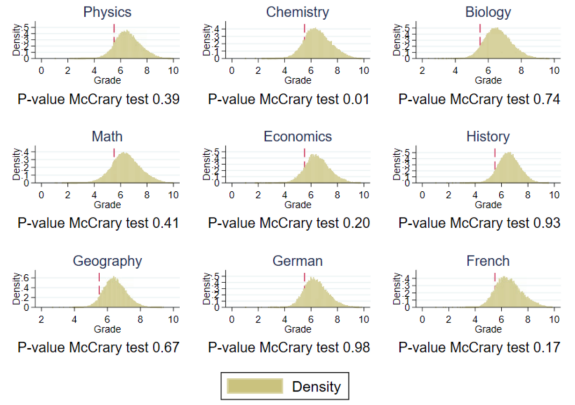
Appendix B

Figure B1: Components of the running variable



Notes. Each panel shows an RDD plot where an indicator for a specific course is the dependent variable. A course indicator is a dummy that is one if the marginal grade is of the type that is printed above the graph. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. The scatter plots use 62,774 observations.

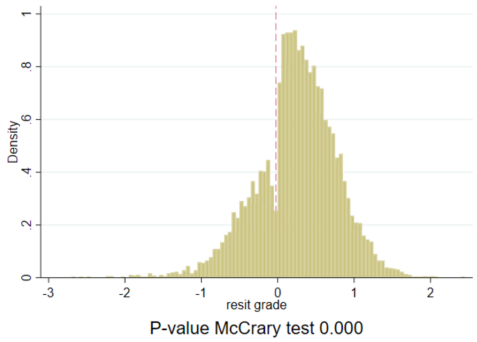
Figure B2: Distribution of the major course grades



Notes. Each panel shows a course grade distribution. Each bar is a bin of 0.05 of a grade. The McCrary test p-values are estimated with the method of Cattaneo et al. (2020) and are printed below each graph.

Appendix C

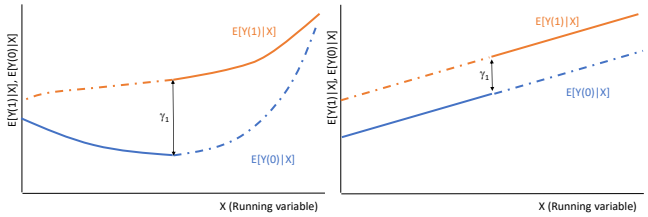
Figure C1: Distribution of the retake grades



Notes. Each bar has an interval of 0.05 grade. Grades are normalized to be 0 at a 5.5, the passing-threshold. Each bar represents the percentage of students scoring the particular grade at the retake. The McCrary test p-value is estimated with the method of Cattaneo et al. (2020) and is printed below each graph. The figure is based on students that take a retake exam from 2006 to 2008, a total of 29.150 students.

Appendix D

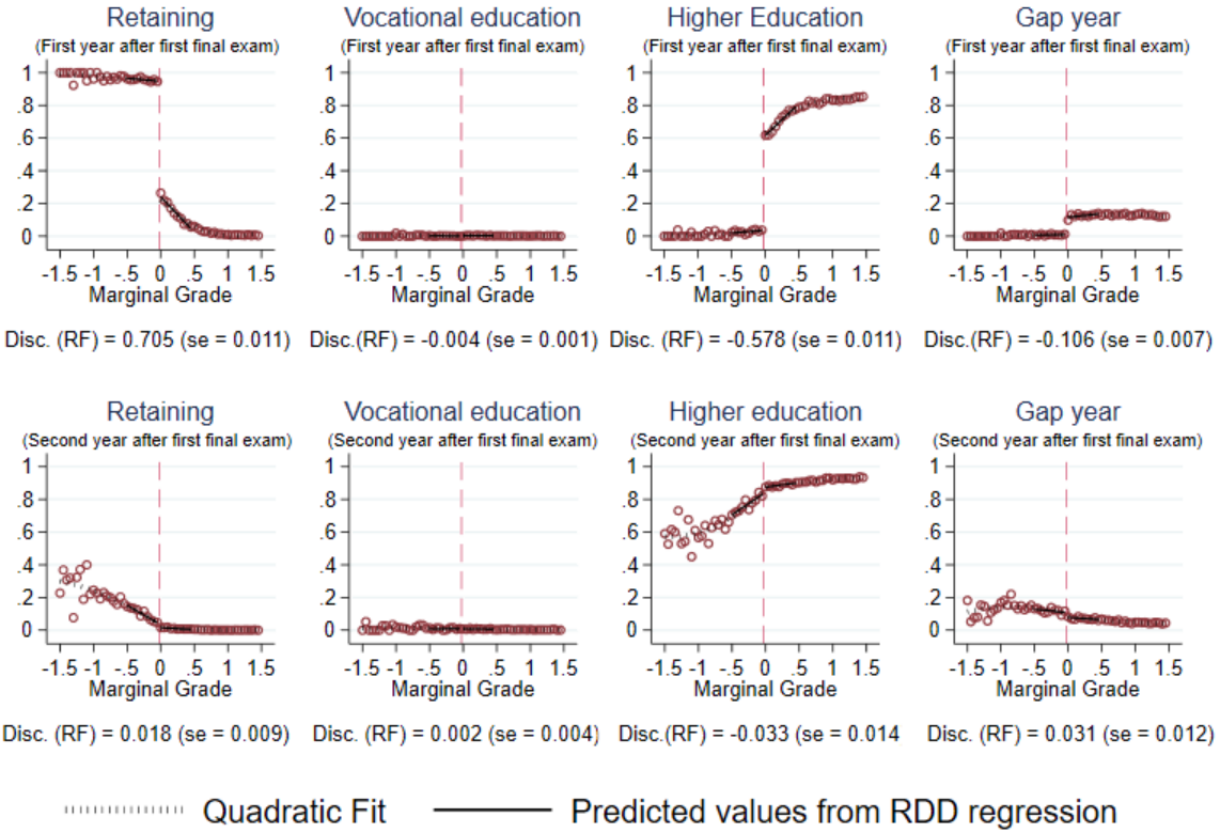
Figure D1: Illustration of heterogeneous and homogeneous potential outcomes



Notes. The orange lines are the potential outcome for the treated subjects. The blue lines are the potential outcomes of the non-treated subjects. The vertical axis is the outcomes of interest. The horizontal axis the running variable. The solid lines are the observed potential outcomes and the dotted lines are the counter factual (or non-observed) outcomes.

Appendix E

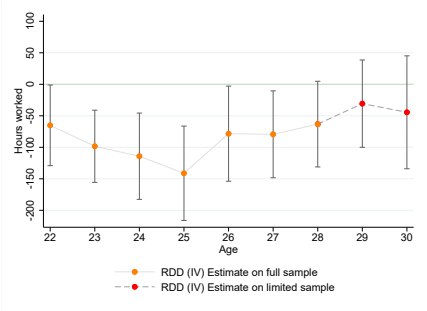
Figure E1: What do students do after their (first) central exam?



Notes. The upper part of the figure shows what the students do in the year after their first central exam. The lower part shows what they do in the second year. The first column shows the share of students that retain. The second, the share that enrolls into vocational education. The third column shows the students that enroll into higher education, and the last column shows the students that take a gap year. A gap year is defined as a year of non-enrollment between years of enrollment. The cohort of students considered did their first final year in 2006, 2007, or 2008. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. Scoring below 0 implies certain grade retention, and scoring above 0 implies a retake and so possible passing. Each red circle is the average for a bin of 0.05 of a grade. The solid black lines are the predicted values from the RDD regression. The dashed line is a second-degree polynomial estimated separately, left and right of the cutoff. The RDD estimate printed below the graph is based on the grades that fall within a 0.5 bandwidth, below and above, the passing grade. The scatter plots use 62,774 observations. The RDD estimates use 15,965 observations.

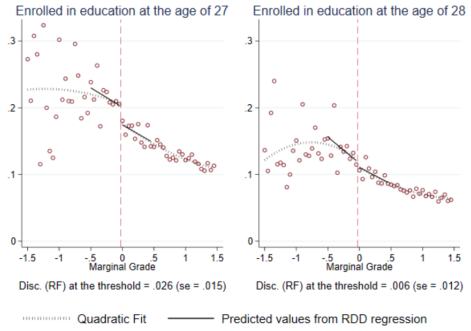
Appendix F

Figure F1: RDD estimates of the hours difference for the ages 22 to 28



Notes. Each dot is the estimate of a separate RDD earnings regression that uses a 0.50 grade point bandwidth, below and above, the passing grade. The vertical lines indicate the confidence intervals. Earnings are deflated and include zero earnings. The orange dots are estimated using all three cohorts, the red dots are estimated using the cohorts that have reached the particular age in 2021.

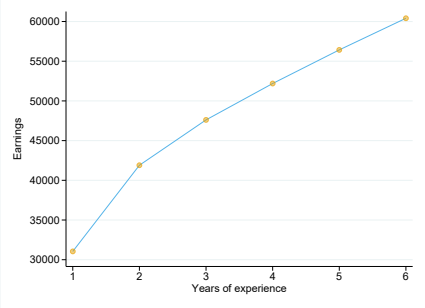
Figure F2: In education at 27 and 28



Notes. The figure show the percentage of students still enrolled in education at the age of 27 and age of 28. The marginal grade is the distance to passing (failing) the course that determines qualification for a retake. The scatter plot uses 62,774 observations. The RDD estimates use 15,965 observations.

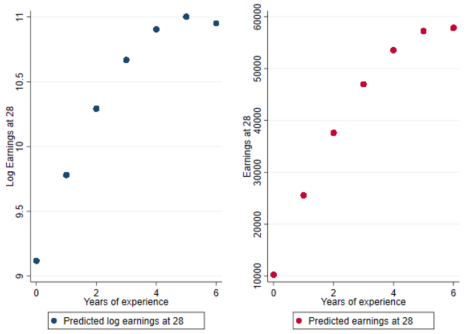
Appendix G

Figure G1: Earnings-experience curve for people with 6 years of experience



Notes. Experience is measured as a years of earnings that are in excess of full time minimum wage. Earnings are either from employment or self-employment. Moreover, earnings are deflated and include zero earnings.

Figure G2: Experience-earnings profiles using Mincer equations



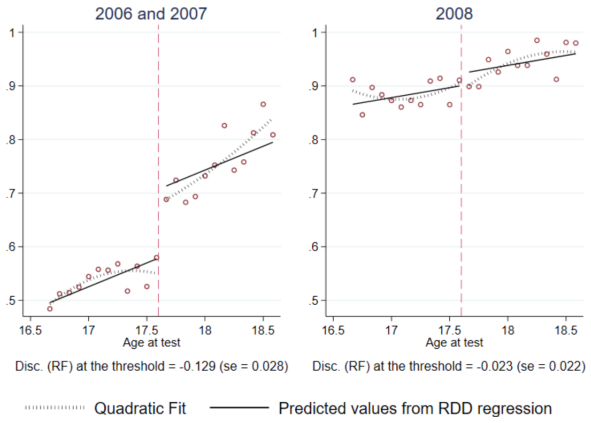
Notes. The figures plot predicted earnings at 28 based upon years of experience, years of experience squared and years of schooling. Experience is measured as a years of earnings that are in excess of full time minimum wage. Years of schooling is based on the time it nominally takes to achieve the degree of a students. Earnings are either from employment or self-employment. Moreover, earnings are deflated and include zero earnings.

Appendix H

Students can choose to re-do the last year of secondary school in a transition center.[28] This is a potential interesting treatment variation, as these centers can target instruction specifically to the needs of the retaining students. Therefore, we could expect that retaking in a transition center has more benefits than staying in the old school. Nonetheless, students entering transition centers have the option to only re-do the failed courses in these centers. Therefore, students in transition centers only attend school for on average 15 hours, whereas students retaining in normal schools attend for a minimum of 28 hours.[29] The lower number of instruction hours could imply less educational benefits of retaining in a transition center.

Before 2008, the government paid for enrollment into transition centers if students were older than 18, but not for younger students. In 2008, the Rutte-regeling made it possible to pay for enrollment into transition centers through a secondment agreement with the original secondary school, de facto making transition centers free of charge for students younger than 18.[30] This implies that in 2006 and 2007, students that were below the age of 18 had to pay for the transition centers, and in 2008 they did not. Figure H1 shows an enrollment cutoff at 18 in the years 2006 and 2007, whereas there is no such cutoff in 2008.

Figure H1: Enrollment into transition centers



Notes. The figure shows the enrollment into transition centers for retaining students in 2006 and 2007 together in the left panel en 2008 in right panel.

I use the (exogenous) increase in enrollment to investigate if attending a transition center leads to better outcomes than retaining in a normal secondary school with a difference-in-difference-in-difference strategy. The first difference is between those that have to retain and those that are promoted. The second difference is between students that are older or younger

[28] It is not uncommon to let students retain in a transition center. For instance, students in Chicago that are older than 15 go to such a center (Jacob & Lefgren, 2009).

[29] There is only data on hours for a small part of the students in transition centers.

[30] See: https://zoek.officielebekendmakingen.nl/stb-2007-348.html

than 18 when they take the central exam. The third difference is between the students that took the central exam (for the first time) in 2006 and 2007, or 2008. I test 2007 against 2006, as a placebo intervention test. As the expansion of the Rutte-regeling only came into effect in 2008, we do not want to see a treatment effects in 2007. As such, we get the following equation:

$$y_i = \rho_1 2008_i \times Retain_i \times Young_i + \rho_2 2007_i \times Retain_i \times Young_i + \rho_3 X_i + e_i \quad (1)$$

Where 2008_i implies that the student took the central exam (for the first time) in 2008, 2007_i implies that the student took the central exam in 2007. $Rertain_i$ implies that the students had to retain, whereas $Young_i$ implies that the students are younger than 18 in October in the year when they took the central exam. All interactions between 2007_i, 2008_i, $Retain_i$ and $Retain_i$, and their individual effects are includes in X_i.

Table H1 shows the results of regression 1. The first column shows a 15% increase in the enrollment rate of young retained students in 2008 due to the Rutte-regeling. The placebo intervention, the year 2007, does not show a significant effect on enrollment. The increase in 2008 in transition center enrollment does correspond to educational attainment. Column 2 shows that there is no effect of 2008 on the educational attainment of young retaining students, nor do we find an effect for the placebo intervention year 2007. Similarly, we do not find effects on earnings in column (3). This implies that retaining in a transition center does not have a different treatment effect as retaining in the school that young students originally attended.

Table H1: Difference-in-Difference-in-Difference regression results

	Transition center enrollment	University degree at 28	Earnings at 28
2008	0.147***	-0.009	1,798
	(0.008)	(0.030)	(2,307)
2007	-0.009	0.005	428
	(0.009)	(0.031)	(2,405)

Notes. Each column is a seperate regression that compares, 2006 against 2008 and 2007, retain against promoted, and young (age<18) against old (age≥18). Each regression includes the interactions and individual effects of the three differences. The regression coefficients in 2008 show the causal effect of extending reimbursement of transition centers to young retaining students. The 2007 coefficients are placebo interaction as if the program would have been implemented in 2007 instead 2008. The sample only includes students within a one year range of 18, a total of 87,475 students. The standard errors are displayed between the brackets, * p<0.1, **p<0.05, and *** p<0.01.

Appendix I

Table I1: Robustness tests

	University degree at 28	Earnings at 28	Starting Earnings	Experience at 28
Panel A: Different polynominal regressions				
1st degree poly.	0.011	-2,193***	355	-0.33***
	(0.018)	(857)	(443)	(0.08)
Obs.	*15,669*	*15,604*	*13,735*	*15,604*
2nd degree poly.	0.019	-2,475**	495	-0.34***
	(0.030)	(1,271)	(656)	(0.13)
Obs.	*15,669*	*15,604*	*13,735*	*15,604*
Panel B: Placebo interval regressions				
Placebo interval at -0.5	-0.029	149	-332	-0.05
	(0.077)	(3,659)	(2,078)	(0.37)
Obs.	*4,875*	*4,860*	*4,150*	*4,860*
Placebo interval at 0.5	0.009	54	-309	0.02
	(0.020)	(962)	(462)	(0.08)
Obs.	*33,596*	*33,417*	*30,000*	*33,417*
Panel C: Donut hole regressions				
Donut hole of 0.2	0.009	-1,599*	725*	-0.31***
	(0.018)	(880)	(433)	(0.08)
Obs.	*18,227*	*18,152*	*16,052*	*18,152*
Donut hole of 0.3	0.002	-2,669**	701	-0.47***
	(0.031)	(1,355)	(722)	(0.13)
Obs.	*17,922*	*17,848*	*15,876*	*17,848*

Notes. Notes. All regressions use a bandwidth of 0.5. 1st degree poly. implies a linear running variable. 2nd degree poly. implies a cubic running variable. Placebo interval at -0.5 implies that the cutoff is at -0.5 and the that the bandwidth varies between -1 and 0. Placebo interval at 0.5 implies that the cutoff is at 0.5 and the bandwidth varies between 0 and 1. The donut hole of 0.2 and 0.3 respectively do not include the 4 and 6 closest grade points from the regression. clustered standard errors are reported between the brackets, * p<0.1, **p<0.05, and *** p<0.01.