UNIVERSITÀ POLITECNICA DELLE MARCHE

Dipartimento di Scienze Economiche e Sociali

MODELING THE EFFECTS OF GRADE RETENTION IN HIGH SCHOOL

STIJN BAERT, BART COCKX AND MATTEO PICCHIO

February 2016

Abstract

A dynamic discrete choice model is set up to estimate the effects of grade retention in high school, both in the short- (end-of-year evaluation) and long-run (drop-out and delay). In contrast to regression discontinuity designs, this approach captures treatment heterogeneity and controls for grade-varying unobservable determinants. We deal with initial conditions and with partial observability of the track choices at the start of high school. Forced track downgrading is considered as an alternative remedial measure. In the long-run, grade retention and its alternative have adverse effects on schooling outcomes and, more so, for less able pupils.

Indirizzo: Stijn Baert. Sherppa, Ghent University; University of Antwerp; Université catholique de Louvain; IZA. E-mail: `Stijn.Baert@UGent.be`.

Bart Cockx. Sherppa, Ghent University; IRES, Université catholique de Louvain; IZA; CESifo. E-mail: `Bart.Cockx@UGent.be`.

Matteo Picchio. Corresponding author. Department of Economics and Social Sciences, Marche Polytechnic University, Piazzale Martelli 8, 60121 Ancona, Italy; CentER, Tilburg University, The Netherlands; Sherppa, Ghent University, Belgium; IZA, Bonn, Germany. E-mail: `m.picchio@univpm.it`. Tel.: +39 071 220 7176. Fax: +39 071 220 7102.

Modeling the Effects of Grade Retention in High School*

Stijn Baert, Bart Cockx and Matteo Picchio

1   Introduction

Grade repetition is practiced in many OECD countries. In 2009 in the OECD on average 13% of 15-year-olds are reported to have repeated at least one year either in primary or high school (OECD, 2012). This practice is very unevenly distributed. In France, Luxembourg, Spain, Portugal and Belgium more than 30% of the 15-year-olds are reported to have some delay in their school curriculum. In contrast, in the Scandinavian countries and in the United Kingdom more than 95% are *on time* at that age. Different societies seem, therefore, to have very different views on the effectiveness of grade retention as a remediation for unsatisfactory performance of pupils. For instance, in some countries, like France, a consensus has grown that grade repetition is bad and the government has taken actions to make schools accountable for overuse OECD (2012, p. 55). By contrast, in other countries, such as in the United States (US), there has been a revival of policies supporting retention in case a certain level of academic achievement is not attained by third grade (Schwerdt et al., 2015). Hence, the practice of grade repetition remains controversial and heavily debated.

Also in the scientific literature, arguments pro and contra grade retention are debated. Proponents argue that by repeating the same grade, low-achieving students have extra

time to catch up to the grade-level requirements, both in terms of knowledge and emotional maturity. By having more time to develop the skills needed in the subsequent grades, students would be less at risk of failure in the future and may even, relative to the counterfactual of promotion to the next grade, increase competencies and earnings in the long-run (Eide and Showalter, 2001). Moreover, the threat of retention might be an incentive device to work more diligently and harder (Manacorda, 2012). Opponents, by contrast, stress the personal and academic costs associated to grade retention. It might (i) hurt pupils' self-esteem (Browman, 2005; Byrd et al., 1997), (ii) generate psychological costs of separating students from their peers (Alexander et al., 1994), (iii) produce financial costs to the families and to society in terms of teaching resources (Eide and Goldhaber, 2005), and (iv) induce lower earnings because of the delayed entry into the labor market, but also because retention can be a negative signal to employers and, hence, lead to lower wages (Brodaty et al., 2013).

Many empirical studies have tried to deliver more insights into this debate by estimating the impact of grade repetition on test scores of academic achievement, but also on other outcomes, such as school drop-out, wage and, recently,[1] on juvenile crime . The estimation of the causal impact is, however, complicated by selection bias. Retained pupils are more likely to have a lower innate ability and weaker social background than promoted students. If these characteristics are not observed by the researcher, the estimates of the impact of grade retention on educational achievements tend to be biased downwards. The early literature indeed mostly found negative achievement effects of grade retention (Holmes, 1989), although less so in studies that matched treated and control students on measures of ability or academic achievement (Allen et al., 2009).

More recent studies, based on Regression Discontinuity Design (RDD),[2] Instrumental Variables (IV)[3] and factor analytic dynamic models (FADM) (Carneiro et al., 2003; Heckman and Navarro, 2007)[4] also take selection on unobservables into account. These studies, but not all, generally find more positive short-run effects on (test scores of) aca-

---

[1]See e.g. Depew and Eren (2015).

[2]Jacob and Lefgren (2004) and Jacob and Lefgren (2009) use RDD to evaluate respectively short- and long-run effects of grade repetition in Chicago Public Schools, Manacorda (2012) for high school students in Uruguay, Greene and Winters (2007) and Schwerdt et al. (2015) for third graders in Florida and Depew and Eren (2015) for fourth and eighth graders in Louisiana.

[3]Eide and Showalter (2001) evaluate grade repetition in high schools in several states in the US, Alet et al. (2013) and D'Haultfœuille (2010) study retention respectively in first-second grade and in fifth grade in France. Dong (2010) uses a control function approach (which is equivalent to IV) to estimate the effect of repeating Kindergarten in the US.

[4]Fruehwirth et al. (2014) study the effect of retention in the US between kindergarten and fourth grade. Gary-Bobo et al. (2014) consider sixth to eighth graders in France.

demic achievement, in particular if retention occurs early in primary school. However, in the long-run, effects on test scores and high school completion remain negative or are, at most, neutral in case of early retention in primary school.[5,6]

In the present study we examine the short- and long-run effects of grade retention in high school on educational achievement in Flanders, the Dutch speaking region in the North of Belgium. The Flemish case is particularly interesting for the following reasons. The PISA studies that measure since 2000 the academic achievements of 15-year-olds in a wide range of OECD countries show that the average performance in the assessed skills of Flanders has been persistently close to top. On the other hand, in Flanders the spread in the scores is also much higher than in most other countries and educational performance is highly segmented according to social background.[7] Retention is a very popular remedial measure in Flanders. In the school year 2014-15 on average 4.6% of all high school students were retained. In earlier years, which this study refers to, this figure was even higher.[8] We are interested in knowing which role retention has played in the aforementioned high, but unevenly distributed, performance of the Flemish school system.

A particular feature of the Flemish high school system is that it consists of very hierarchically ordered tracks that students enter at the start, generally at the age of twelve. This has been labeled the *cascade-system*, since many pupils start off in a high track and gradually downgrade, i.e. trickle down the cascade. Upgrading from a lower to a higher ordered track is not possible. Downgrading can be a free choice of the student, or imposed by the staff meeting of teachers at the end of the school year if a student did not pass the exams in the main subjects of the track she attended. The imposed downgrading is an alternative remedial measure that avoids grade repetition by reorienting the student to a less demanding track. In this case the student can still decide to stay in the high track,

---

[5]The study of Eide and Showalter (2001) finds positive effects on high school completion and on labor market earnings, but these were very imprecisely estimated and not statistically significantly different from zero.

[6]The aforementioned studies evaluate the ex-post effect of retention. These disregard, however, that the ex-ante threat of retention may encourage students to study harder in order to avoid grade repetition (Manacorda, 2012). Foureaux Koppensteiner (2014) finds that the introduction of an automatic promotion system in Brazilian high schools, removing this deterrence effect, significantly reduces academic achievement. However, conversely, Belot and Vandenberghe (2014) tentatively conclude that the reintroduction of grade repetition in grade 7 in high school of the Belgian French speaking Community did not lead to better medium-run outcomes among those most at risk of grade repetition.

[7]Source: Department of Education, Ghent University (www.pisa.ugent.be/nl/resultaten/vlaamse-publicaties).

[8]Source: Flemish Ministry of Education (www.ond.vlaanderen.be/onderwijsstatistieken/rapporten).

but then she is forced to repeat the grade. The aim of this paper is to study the effect of grade retention within this cascade system and also to compare it to track downgrading. Other countries also separate students into different tracks OECD (2012, p. 58-60), so that downgrading as remediation is not only of interest to the Flemish educational system.

Using data from retrospective surveys conducted on representative samples of youth belonging to the 1978 and 1980 birth cohorts, we estimate the average treatment effects of grade repetition on retained pupils in eighth, ninth and tenth grade (i.e. the second, third and fourth grade of high school in Flanders). To that purpose, we model and estimate the sequence of decisions that these students (or their parents) take throughout their high school career each time at the end of the schooling year: track choice (only at the start of high school), the decision to drop-out of school (only at ages when schooling is no longer compulsory), the decision to repeat the grade or not and the decision to downgrade or not. These students' decisions are conditional on the teachers' overall end-of-year evaluations: pass (A), forced downgrade (B) or fail (C). These teachers' evaluations are also modeled and estimated based on past schooling outcomes.

Our approach fits into the aforementioned FADM and is, hence, most comparable to the work of Fruehwirth et al. (2014) and Gary-Bobo et al. (2014). A common feature of this approach is that it exploits the panel structure of the data to identify not only selection on unobservables, but also *essential heterogeneity* in the treatment effect (Heckman et al., 2006). The key identifying assumption is that unobserved determinants of both treatment and outcomes are assumed to be captured by a *low dimensional set of common causes* (Fruehwirth et al., 2014, p. 14). Since, contrary to Fruehwirth et al. (2014), we do not have separate measurements of the multidimensional unobserved ability, we restrict the factor structure of these unobservables to be *uni*dimensional. An advantage of imposing this low dimensional factor structure is that it enables, in addition, to take time-varying unobserved determinants into account.[9] We allow for grade-varying unobserved heterogeneity. This is important in our analysis, because this can capture the impact of the changing environment (school, classroom and peer effects) over grades.[10]

Our paper also contributes methodologically to the FADM-approach. First, we face an initial conditions problem, because pupils may already have been retained in kindergarten or primary school.[11] In order to take into account that this early retention may also affect

---

[9]Gary-Bobo et al. (2014) do not allow for this.

[10]Because of the relatively small sample size and because each year only a relatively small fraction of the pupils repeat the grade, allowing for year-varying instead of grade-varying unobserved heterogeneity was not feasible.

[11]In contrast to many other countries, in Flanders pupils can be retained multiple times. We explicitly

the outcomes in high school, we follow the approach proposed by Wooldridge (2005) and condition the unobserved heterogeneity distribution on the number of years of schooling delay at the start of high school. Second, we do not only consider educational achievement as outcome (as the other authors using the FADM approach), but also the high school drop-out (as, e.g., Jacob and Lefgren, 2009), the average delay by grade (as, e.g., Schwerdt et al., 2015) and the attained track level. Considering not only educational achievement as outcome, but also other (long-term) outcomes is important, because the effect of grade retention has been shown to be more negative on the latter than on the former. Third, by explicitly modeling tracking and the imposed downgrading, we can contrast the relative efficacy of retention and (forced) downgrading. Finally, we propose a method to deal with a problem of partial observability that we face at the start of high school. In the analysis we distinguish five different tracks, but we can only observe the chosen track from eighth grade onwards. In grade 7 we can only observe a more global division into three tracks. To do so we follow a similar approach as Mroz et al. (2015) who account for partial observability of the onset time of a disease by integrating over all possible onset times. We integrate the likelihood function over the possible tracks that could have been chosen in grade 7 given the observed choice in grade 8.[12]

We find that grade retention has a neutral effect on the evaluation in the next grade. In contrast, the long-term effects are largely adverse. Pupils repeating (for the first time) grade 8 have a 14 percentage point lower chance to graduate from high school. Alternative timing of the retention, i.e. in grade 9 or 10, does not affect these treatment effects. Retention causes an increased likelihood of downgrading both in the short-run, by the next grade, and, although slightly less, in the longer run, by the last year of compulsory education. By this last year the likelihood of being in the highest track is 14 percentage points lower, while that of being in the lowest track is enhanced by the same amount. In line with the research based on RDD, we find that students who are at the margin of being retained increase their academic achievement in the short-run. However, as Fruehwirth et al. (2014), we find that retention of the lowest ability students decreases the academic achievement and drives the significantly positive effect on high school drop-out. When comparing retention to track downgrading as an alternative remedial measure, we observe a slightly more adverse effect on the academic achievement in the short-run and on high school drop-out, but the difference is not statistically significant. The main difference is

---

allow for this in the analysis.

[12]Contrary to Mroz et al. (2015), we exploit that the information available in the future observed outcomes and in the *cascade* nature of track choices in the Flemish high school system to weigh in this integration the possible track choices by their probability of occurrence.

that the delay in the last compulsory school year is much reduced in case of downgrading. Relatively to allowing students to unconditionally pass to the next grade (awarding an A), this remedial might stimulate academic achievement in the short-run, but this advantage disappears in the longer run. Overall, we conclude in line with the existing literature that retention is an inadequate remedial measure. Moreover, downgrading in a hierarchical tracking system does not perform sufficiently better. Policy recommendations are clear: stop retention and downgrading.

This study is organized as follows. In Section 2, we present the Flemish (Belgian) educational system with a focus on the functioning of high school. Section 3 describes the data and summarizes basic descriptive statistics of the variables used in the empirical analysis. Section 4 presents the econometric model. Section 5 reports our empirical findings, which are quantified on the basis of counterfactual simulations. Finally, Section 6 concludes.

2   The Flemish High School System

Flanders is the Dutch speaking region of Belgium, situated in the Northern part of the country. Belgium is a federal country with several competencies devolved to its three Regions (Flanders, Brussels and Wallonia) and three Communities (Dutch, French and German speaking). While the federal authorities are competent for all matters of National importance, territorial and person-related issues are left to Regions and Communities. Since 1988, the Flemish Community is in charge of all aspects of education policy in Flanders.

Compulsory education starts on 1 September of the year in which the child turns 6 years old and ends on 30 June of the year in which she reaches the age of 18.[13] The start of compulsory education coincides with the beginning of primary school. However, children might start one or more years earlier if in kindergarten they are suggested to do so.[14] Grade retention and grade skipping are also allowed in primary school. Hence, pupils start high school at different ages. The regular starting age is the year in which they turn 12. In our research sample outlined in the next section, only 1.1% started high

---

[13]From the age of 15 (conditional on passing the first two years of full-time high education) or 16 (unconditionally), only part-time education is mandatory.

[14]The choice to send children to primary education is formally made by their parents. However, in practice, parents follow teachers' judgment of whether their child is school ready (Baert and Cockx, 2013). In our sample, 1.4% of children started primary school in the year they turned 5 and 1.1% started it when 7 or 8.

school in the year they turned 11 and 3.7% started with delay.

In the beginning of high school students are grouped in hierarchical tracks according to their abilities and interests. This is a quite common practice in OECD countries to take the diversity of skills and educational preferences of pupils into account. In this study, as in Van de gaer et al. (2006) and Van Houtte et al. (2012), we refer to *tracking* as a system in which students are allocated to different pathways, i.e. *tracks*, in which they are taught entirely different curricula and may be denied to pursue a track in case of unsatisfactory performance. This differs from *setting* or *banding*, when pupils in the same curriculum are taught at different difficulty levels according to their ability (Gamoran et al., 1995). In Flemish high school, four main tracks can be distinguished: (i) the general track (GHS) provides a primarily theoretical general preparation for higher education; (ii) the technical track (THS) consist of a mix of theoretical and practical classes aiming at both direct labor market entry after completion or entry in primarily technical higher education; (iii) the vocational track (VHS) teaches practical skills that prepare for particular professions; from age 15 or 16, students in this track may move to part-time education, in which formal classroom training can be combined with on-the-job training; (iv) the arts track (AHS) combines general education with active arts practice. In this study, we neither consider AHS, nor part-time education in VHS, because too few pupils in our sample choose it. Students graduate from high school if they successfully pass the six grades of GHS or THS or the seven grades of VHS. All high school graduates can enter higher education without passing any central entry exam, except for the study of medicine.

Track mobility in high school is possible at the start of each academic year, but it is constrained in the following two ways. First, tracks are hierarchical with the following ordering from high to low: (i) GHS, (ii) THS and (iii) VHS. It is only possible to move down these tracks, not upwards.[15] Moreover, within GHS and THS, a further division of hierarchically ordered sub-tracks can be identified. We label these, respectively, GHS^+/GHS^- and THS^+/THS^-. Even if according to the rules it is not forbidden to move up tracks, in practice it is only possible to move down, because students would not have the pre-requisite for certain courses of the higher tracks.[16] This hierarchical ordering justifies to model the track choice as an ordered choice model (see Section 4.2.1). A second constraint is that track changes are not permitted between the before last and last grade, i.e. between grade 11 and 12 in GHS and THS and between and grade 12 and 13 in VHS.

---

[15]According to the rules, not forbidden to move upwards but, in practice, this is not feasible because students would not have the pre-requirements for certain courses of the higher tracks.

[16]More concretely, GHS^+ comprises the curricula including Latin and/or Ancient Greek and THS^+ comprises the curricula focused on industrial sciences and on commerce.

At the end of each academic year, pupils receive an evaluation: A, B or C. Pupils getting an A are promoted to the next grade and, if they wish, can downgrade the track. Pupils obtaining a C must repeat the grade and, if they wish, can downgrade the track. Pupils who have been awarded a B are forced to downgrade in case they want to be promoted to the next grade. They can only avoid track downgrading by repeating the same grade.

3   Data and Sample

We base our analysis on a survey conducted on two random samples of respondents, one born in 1978 and the other in 1980. Each sample consisted originally of about 3,000 individuals. These individuals were surveyed at the age of 23.[17] They were asked to provide information on some strictly exogenous characteristics (such as gender, birth date, parents' level of education and number of siblings), as well as years in which they started primary school, and year-by-year detailed retrospective information on their high school career: track choices, end-of-year evaluations (A, B or C) and timing of school drop-out or graduation. The data lack information on school and class characteristics and on the place of living within Flanders. The information on higher education and labor market participation is not used in this research.

The original sample contains $5,915$ pupils. In order to have a sample of pupils with a homogeneous educational, social and family background, we removed from the sample pupils whose grandmother on mother's side had a foreign nationality (583 pupils deleted), pupils who need special help, temporarily or permanently, and were therefore in special schools, and pupils who started high school when older than 15 (473 pupils deleted). As aforementioned, we also dropped students entering the arts track (103 students), those reporting a break of one or more years in high school attendance, those leaving school before the end of compulsory education (9 pupils), those ending in part-time education (183 students), those with inconsistent or missing information on the progression of the grade, evaluation and grade mobility (397 pupils) and those with missing values for some of the covariates used in the econometric model outlined in the following section (146 students). Since only 42 students were retained in first grade and only 46 students made track transitions involving more than two steps, we deleted their records from our sample.

---

[17]The survey was conducted by the so called *SONAR* research team. The team also surveyed the 1976 birth cohort, but because no detailed information on school tracks was available for this cohort, it was not retained in our analysis. The 1978 cohort was also interviewed at age 26 and the 1980 cohort at 29, but since these surveys did not add any information on the high school career, they were neglected.

After applying these selection criteria, we ended up with a sample of $3,933$ pupils who were observed in each year of their high school career.

Table 1: The Panel Structure of Our Sample

	Absolute frequency	Relative frequency (%)
Distribution of the number of multiple observations of pupils		
5	6	0.15
6	2,824	71.80
7	861	21.89
8	200	5.08
9	37	0.94
10	4	0.10
11	1	0.02
Total	3,933	100.00
Distribution of the number of grade observations in the pooled sample		
1	3,933	15.74
2	4,094	16.39
3	4,121	16.49
4	4,178	16.72
5	4,288	17.16
6	4,003	16.02
7	368	1.47
Total	24,985	100.00

Table 1 clarifies the panel structure of our sample. In the top panel we tabulate the number of years in high school and the corresponding number and fraction of individuals in the retained sample. Most students are observed during six years, the typical duration of high school in case of regular promotion in each grade. Because students can be retained multiple times, students are observed up to 11 years in high school. No student stays less than five years in high school. This is a consequence of compulsory schooling until age 18 and because very few pupils have started high school beyond the age of 13. The bottom panel of Table 1 reports the distribution of observations over each high school grade. Having multiple observations per grade will help in identifying the grade-varying unobserved heterogeneity (Section 4.5). The total number of year observations in our sample is $24,985$, i.e. 6.35 years per individual on average.

Table 2 displays summary statistics of schooling attainment and choices modeled in the empirical analysis. In the first columns overall statistics referring to the whole sample are reported. The subsequent columns refer to the subsamples of pupils who have never repeated a grade and those who have repeated a grade at least once in high school. First, we report some outcomes and decisions at the end of the schooling year averaged over the high school career. In our sample on average in each year 91.4% of the pupils get an A, the highest evaluation, while 4.6% and 3.7% obtain a B and a C, respectively. Pupils who have never repeated a grade realize far much better evaluations than those who did

repeat at least one grade. Around 4.4% of the pupils are retained on average at the end of the academic year.[18] Over the high school career this leads to 874 out of the 3,933 pupils (22.2%) being retained at least once (see Table 3 below). On average 91.7% of the students graduate with a high school diploma. However, 14.9% of the students who repeat a grade at least once drop-out from high school. 91.7% of the pupils never change track, while each year on average 6.4% start the academic year with a one-step track downgrade and 1.9% with a two-step downgrade. Repeaters are only marginally more likely than

Table 2: Summary Statistics of Outcome Variables: Schooling Attainment and Choices

	Whole sample		No repetition		Grade repetition	
	Mean	SD	Mean	SD	Mean	SD
Outcomes and decisions at the end of the year[§]						
Evaluation: A	0.914	0.280	0.966	0.181	0.763	0.425
Evaluation: B	0.049	0.215	0.032	0.176	0.096	0.295
Evaluation: C	0.037	0.189	0.002	0.045	0.141	0.348
Retention	0.044	0.206	0.000	0.000	0.174	0.379
High school graduation with diploma	0.917	0.276	0.939	0.239	0.851	0.356
No downgrade	0.917	0.276	0.925	0.263	0.894	0.308
One-step downgrade	0.064	0.245	0.060	0.237	0.076	0.265
Two-step downgrade	0.019	0.136	0.015	0.119	0.122	0.171
Cumulative delay (years)						
Cumulative delay at the beginning of grade 7	0.027	0.217	0.028	0.217	0.022	0.210
Cumulative delay at the beginning of grade 8	0.027	0.217	0.028	0.217	0.022	0.210
Cumulative delay in the last year of high school	0.309	0.619	0.028	0.217	1.289	0.568
Track at the beginning of grade 7						
GHS/THS	0.951	0.215	0.953	0.212	0.945	0.228
VHS	0.049	0.215	0.047	0.212	0.055	0.228
Track at the beginning of grade 8						
GHS$^+$	0.275	0.447	0.329	0.470	0.116	0.320
GHS$^-$	0.409	0.492	0.386	0.487	0.476	0.499
THS$^+$	0.101	0.301	0.084	0.277	0.151	0.358
THS$^-$	0.137	0.344	0.123	0.328	0.177	0.382
VHS	0.078	0.268	0.077	0.267	0.080	0.271
Track at the end of high school						
GHS$^+$	0.142	0.349	0.177	0.382	0.022	0.147
GHS$^-$	0.381	0.486	0.411	0.492	0.279	0.449
THS$^+$	0.109	0.312	0.097	0.296	0.150	0.357
THS$^-$	0.214	0.410	0.174	0.379	0.356	0.479
VHS	0.153	0.360	0.142	0.349	0.193	0.395
Number of pupils	3,933		3,059		874	
Number of pupils × number of years of schooling	24,985		18,622		6,363	

[§] The statistics of the presented outcomes and decisions are yearly averages over the high education career.

non-repeaters to experience a one-step downgrade (7.6% versus 6.0%), but the likelihood of a two-step downgrade is much higher: 12.2% against 1.5%.

Second, Table 2 displays the average cumulative delay at the beginning of grade 7, grade 8 and in the last year of high school (which is prior to graduation for those who

---

[18]This figure is in line with the figures reported in OECD (2004, p. 262) for the whole Belgium.

drop-out early). At the beginning of high education, the average number of years of schooling delay is 0.03. By the end of high school pupils have been on average retained during 0.31 years. For those who have repeated at least one grade, the delay is on average 1.29 years. Note that the delay only starts to accumulate beyond grade 8: no one repeats grade 7.

Third, Table 2 reports the relative frequency of track choices at the beginning of grade 7 and grade 8. At the beginning of grade 7, we have only partial information about the school track choice. We only know whether the student is in the vocational track (VHS) or not (i.e. in GHS or THS). This partial observability generates a complication in modeling track choices at the start of grade 7 and subsequent downgrades. We explain how we deal with this in Subsection 4.4. We have more detailed information on the tracks only starting from grade 8. From this grade onwards, we can group track choices into five hierarchical categories: GHS^+, GHS^-, THS^+, THS^- and VHS. At the beginning of grade 7, 4.9% of pupils choose VHS. As a result of some downgrading decisions, this frequency increases up to 7.8% when moving to grade 8; 27.5% are instead in GHS^+, 40.9% in GHS^- and the remaining 23.8% is split almost evenly between THS^+ and THS^-. By the end of high school the fraction in the higher tracks has substantially decreased in favor of the lower tracks. For instance, by the end of high school the fraction in GHS^+ is about half of what it was at the start of grade 8, while the fraction in VHS has about doubled. This is a consequence of the important degree of downgrading in Flemish high school. This downgrading is highly correlated with grade repetition. For instance, at the start of high school the fraction in VHS was roughly equally distributed over the two subsamples, while at the end of high school the fraction in VHS was clearly higher in the group that repeated at least one grade.

Table 3: Summary Statistics of Covariates at the Beginning of High School

	Whole sample		No repetition		Grade repetition	
	Mean	SD	Mean	SD	Mean	SD
Female	0.512	0.500	0.549	0.498	0.382	0.486
Calendar day of birth	183.671	104.225	182.532	104.011	187.660	104.931
Father's education after primary school (years)	6.342	3.318	6.406	3.330	6.116	3.266
Mother's education after primary school (years)	5.931	2.990	6.007	3.000	5.666	2.939
Cohort						
1978	0.500	0.500	0.500	0.500	0.500	0.500
1980	0.500	0.500	0.500	0.500	0.500	0.500
Presence of siblings						
0	0.136	0.343	0.136	0.343	0.136	0.343
1	0.471	0.499	0.472	0.499	0.468	0.499
2	0.258	0.437	0.259	0.438	0.254	0.436
3 or more	0.135	0.342	0.133	0.339	0.142	0.349
Number of pupils	3,933		3,059		874	

11

Table 3 reports descriptive statistics of the strictly exogenous covariates conditioned upon in the econometric analysis. About one half of the sample is female and the average day of birth is close to the middle of the calendar year. Out of the $3,933$ pupils in our sample, $1,967$ are born in 1978 and $1,966$ are born in 1980. Almost one half of the pupils have one sibling, 13.6% are only child and 39.3% have more than one sibling. Pupils' fathers are more educated than pupils' mothers, having on average 6.3 years of successful education beyond primary school against 5.9 years for mothers. The parents of the pupils who experienced grade repetition are lower educated on average than the pupils who were systematically promoted from one year to the next. This reflects the selection problem mentioned in the Introduction.

4 The Econometric Model

In this section, we write down the econometric model. Schooling choices of pupils and teachers are taken sequentially over time. We exploit this sequentiality together with the assumption that there are some common unobservable (grade-varying) variables, independent of the strictly exogenous observed explanatory variables, to identify the causal relationship of past choices (e.g. the end-of-year evaluation) on subsequent schooling outcomes (i.e. subsequent choices). In a first subsection we make these assumptions explicit for an unspecified joint distribution of choices conditional on aforementioned observable and unobservable variables, and on initial conditions. Based on these general assumptions, we derive the likelihood function. In the following subsection, we clarify our modeling assumptions regarding each of the schooling choices by pupils and teachers. In subsection 4.3 we explain how we restrict the (grade-varying) unobserved heterogeneity distribution. In Subsection 4.4 we discuss how we deal with the problem that we observe the track choice at the start of high school only partially. Finally, in Subsection 4.5 we briefly argue why based on our modeling assumptions the unknown parameters are identified.

4.1 General Model Assumptions and the Likelihood Function

Given the set-up of the Flemish high school system, there are different choices that pupils (or/and their parents) have to make. At the start of high school they must first choose the track in which they commence. Then, at the start of each academic year, they must choose, based on the end-of-year evaluations (A, B or C), whether to repeat the grade

or not (only in case of a B) and whether to downgrade the track or not. From age 18 onwards, schooling is no longer compulsory, so that pupils can, from that point onwards, also choose to drop-out of school. Formally, these choices by pupils and teachers (i.e. their evaluations) can be represented by five discrete outcome variables for each pupil i ($i = 1, \ldots, N$) in each academic year t ($t = 1, \ldots, T_i$):

- tr_i: track choice at the beginning of high school. Since tracks are hierarchically ordered, tr_i is an ordered response taking on the increasing values $\{\mathrm{VHS}, \mathrm{THS}-, \mathrm{THS}+, \mathrm{GHS}-, \mathrm{GHS}+\}$.

- ev_{it}: evaluation at the end of each academic year t. This is an ordered response taking on the following increasing values $\{\mathrm{C}, \mathrm{B}, \mathrm{A}\}$ if pupils are not in the last grade or if they are not in the lowest track VHS. Alternatively, if the student is in the last grade year or in the VHS track, the regulations impose (Section 2) that the ordered response can take on two values only: C or A .

- out_{it}: school drop-out at the end of year t from age 18 onwards (i.e. the end of compulsory schooling age). out_{it} is an indicator equal to 1 if student i drops out of school in year t, 0 otherwise.

- re_{it}: choice to repeat the grade in year $t + 1$ if the evaluation at the end of year t is B ($ev_{it} = \mathrm{B}$). re_{it} is a dichotomous variable equal to 1 if the student chooses to repeat the grade when she gets a B, 0 if she chooses instead to downgrade.

- dow_{it}: choice to downgrade the track in year $t + 1$. This choice is in most cases defined as an ordered response, taking on values $\{0, 1, 2\}$, where 0 means 'no downgrade', 1 stands for 'one-step downgrade' and 2 is 'two-step downgrade'. When pupils are in VHS, which is the lowest track, they cannot downgrade further, so there is no choice to be taken. If they are in $\mathrm{THS}-$, they cannot make a 'two-step downgrade' and their choice is dichotomous, either 'no downgrade' or 'one-step downgrade'.

Four out of five of these choice variables are observed at the end of each academic year t, resulting in multiple observations for each individual. Only the track choice at the beginning of high school, tr_i, is observed just once. Moreover, the track choice is only partially observed, a point to which we return in Subsection 4.4.

As mentioned Section 3, pupils start high school at different ages due to different past retention histories either in primary school or in kindergarten. If we allow for past performance, like past grade retention, to affect future outcome variables, we have an initial

conditions problem. The cumulative years of delay at the beginning of high school, in_i, cannot be assumed to be exogenous, since they can be correlated to the unobserved determinants. Heckman (1981) proposes to solve the initial conditions problem by explicitly modeling the years of delay at the beginning of high school as a function of observed and unobserved determinants. However, we follow the alternative solution proposed by Wooldridge (2005). This consists in conditioning on the delay at the start of high school and to integrate out the unobserved heterogeneity conditional on this delay. We opt for Wooldridge's (2005) solution for two reasons. The first one is a practical one. Since we already model five different outcomes, Wooldridge's (2005) correction allows us to avoid the inclusion of a sixth to process to be modeled. The second is based on the simulation findings in Akay (2012), who pointed out that Heckman's (1981) and Wooldridge's (2005) methods perform similarly when the number of multiple observations per unit is larger than $5 - 8$. Table 1 shows that our panel structure fits to this case.

Let $\mathbf{Y}_{it} \equiv [ev_{it} \; out_{it} \; re_{it} \; dow_{it}]$ be the row vector collecting the four time-varying outcome variables, $\mathbf{z}_i$ be the vector of observed strictly exogenous explanatory variables and $\mathbf{v}_i$ a random term capturing the (possibly grade-varying) unobserved determinants of our outcome variables. We can always write the joint density of $(tr_i, \mathbf{Y}_i)$ conditional on $(\mathbf{z}_i, \mathbf{v}_i, in_i)$ as:

$$f(tr_i, \mathbf{Y}_i | \mathbf{z}_i, \mathbf{v}_i, in_i) = f(tr_i | \mathbf{z}_i, \mathbf{v}_i, in_i) \cdot \prod_{t=1}^{T} f(\mathbf{Y}_{it} | \mathbf{z}_i, \mathbf{v}_i, \mathbf{Y}_{it-1}, \ldots, \mathbf{Y}_{i1}, tr_i, in_i)$$

$$\equiv f(tr_i | \mathbf{z}_i, \mathbf{v}_i, in_i) \cdot \prod_{t=1}^{T} f(\mathbf{Y}_{it} | \mathbf{z}_i, \mathbf{v}_i, \Im_{it-1}), \tag{1}$$

where in_i is the initial condition, i.e. the cumulative numbers of years of delay at the beginning of high school and $\Im_{it-1}$ denotes the information set containing all the realizations of the endogenous variables from $t - 1$ until the start of the processes, i.e. $\Im_{it-1} \equiv (\mathbf{Y}_{it-1}, \ldots, \mathbf{Y}_{i1}, tr_i, in_i)$.

Assumption 1 *(Sequentiality):*
Within each academic year t and for $t = 1, 2, \ldots, T$, the four time-varying outcome variables in $\mathbf{Y}_t$ are realized sequentially with the following chronological order: evaluation at the end of the academic year, ev_t; school drop-out choice, out_t; choice to repeat the grade, re_t (choice available only in case of a B); track downgrade choice, dow_t.

The aforementioned chronological order is natural. Nevertheless, one might question the assumption that the choice to repeat the grade in case of a B precedes the downgrading

decision. Recall that in case of a B, the decision to repeat the grade or not influences the choice set of downgrading, because a B student who wishes to be promoted to the next grade cannot stay in the same track. We believe that it is more natural to assume that the choice of grade repetition comes first, because the decision to be promoted to next grade just restricts the choice set of the downgrading decision. By contrast, if downgrading is first decided upon, the choice of downgrading is implicitly mixed up with the decision to repeat the grade, because if the student decides to stay in the same track, she has implicitly simultaneously decided that she will repeat the grade: no choice is left. Nevertheless, we tested which order in the timing provides a best fit of the data. To that purpose, we estimated the model under both assumptions and used a Vuong (1989) test to determine whether one of these two non-nested models could be statistically rejected against the other. We find that the alternative order of events could be rejected against the one stated in Assumption 1. The value of the asymptotically standard Normal statistic is 3.544 and rejects the alternative hypothesis at a p-value of 0.0002.[19]

Based on Assumption 1, we can rewrite the conditional density in Equation (1) as:

$$f(tr_i, \mathbf{Y}_i | \mathbf{z}_i, \mathbf{v}_i, in_i) = f(tr_i | \mathbf{z}_i, \mathbf{v}_i, in_i) \cdot \prod_{t=1}^{T} \Big[f(ev_{it} | \mathbf{z}_i, \mathbf{v}_i, \Im_{it-1})$$
$$\cdot \ f(out_{it} | \mathbf{z}_i, \mathbf{v}_i, \Im_{it-1}, ev_{it})^{s_{it}}$$
$$\cdot \ f(re_{it} | \mathbf{z}_i, \mathbf{v}_i, \Im_{it-1}, ev_{it} = B, out_{it} = 0)^{1-lg_{it}}$$
$$\cdot \ f(dow_{it} | \mathbf{z}_i, \mathbf{v}_i, \Im_{it-1}, ev_{it}, out_{it} = 0, re_{it})^{c_{it}} \Big], \qquad (2)$$

where s_{it} is an indicator equal to 1 if student i is 18 or older (age from which she can legally drop-out), lg_{it} is an indicator equal to 1 if the student is attending the last grade of high school and c_{it} is equal to 1 if the student is in the GHS/THS tracks and 0 if she is in the VHS track (VHS students do not have the option to downgrade as already at the bottom of the track hierarchy). We cannot derive the likelihood function on the basis of the density in Equation (2), because we do not observe $\mathbf{v}_i$. Instead, we integrate $\mathbf{v}_i$ out after assuming that it is orthogonal to $\mathbf{z}_i$, conditional on in_i:

Assumption 2 *(Orthogonality conditional on in_i):*
$\mathbf{v}_i \perp\!\!\!\perp \mathbf{z}_i | in_i$.

---

[19]The complete estimation results and calculation of the test statistic can be found in the Internet Appendix which can be downloaded from the following address: http://users.ugent.be/~bcockx/IA_BCP.pdf.

15

Based on Assumption 2 and provided that we parametrically specify the conditional probability density functions of the outcome variables, we can integrate out the unobservables $\mathbf{v}_i$ from the joint probability density function in Equation (2) given a specification of the conditional cumulative distribution function $G(\mathbf{v}_i|in_i; \boldsymbol{\delta}_i)$, where $\boldsymbol{\delta}_i$ is a finite dimensional vector of unknown parameters. Taking the logarithm of this marginal joint probability density function for each individual i and summing over all the individuals in the retained sample yields the log-likelihood function:

$$
\begin{aligned}
\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^{N} \ln \Bigg\{ & \int_{\mathbb{R}^5} f(tr_i|\mathbf{z}_i, \mathbf{v}_i, in_i; \boldsymbol{\theta}_{tr}) \cdot \prod_{t=1}^{T} \Big[f(ev_{it}|\mathbf{z}_i, \mathbf{v}_i, \Im_{it-1}; \boldsymbol{\theta}_{ev}) \\
& \cdot \; f(out_{it}|\mathbf{z}_i, \mathbf{v}_i, \Im_{it-1}, ev_{it}; \boldsymbol{\theta}_{out})^{s_{it}} \\
& \cdot \; f(re_{it}|\mathbf{z}_i, \mathbf{v}_i, \Im_{it-1}, ev_{it} = B, out_{it} = 0; \boldsymbol{\theta}_{re})^{1-lg_{it}} \\
& \cdot \; f(dow_{it}|\mathbf{z}_i, \mathbf{v}_i, \Im_{it-1}, ev_{it}, out_{it} = 0, re_{it}; \boldsymbol{\theta}_{dow})^{c_{it}} \Big] \Bigg\} \mathrm{d}G(\mathbf{v}_i|in_i; \boldsymbol{\delta}) \\
\equiv \sum_{i=1}^{N} \ln \Bigg[& \int_{\mathbb{R}^5} \mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{\delta}) \Bigg] \mathrm{d}G(\mathbf{v}_i|in_i; \boldsymbol{\delta}),
\end{aligned}
\tag{3}
$$

where $\mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{\delta})$ is the individual contribution to the likelihood and $\boldsymbol{\theta}$ is the vector of parameters fully characterizing the probability density functions conditional on $\mathbf{v}_i$.

To be able to maximize this log-likelihood we must further impose functional forms and specify how explanatory variables and past choices affect current high school choices. In the subsequent subsection we specify the conditional probability density functions of each high school choice. In addition, we explain how we allow for essential heterogeneity in the treatment effects (Heckman et al., 2006), i.e. that the treatment effect may depend on unobservables. In Subsection 4.3 we clarify our assumptions regarding the unobserved heterogeneity distribution.

4.2   The Empirical Specification of the Educational Choices

We assume that all educational choices can be specified as (ordered) logits. However, the logit form is not strictly required for identification (Fruehwirth et al., 2014). It is sufficient that the choices can be written as index models with an additively separable independent error term, but the logit form is chosen to facilitate estimation.

Assumption 3 *(Logit and ordered logit probability density functions):*

The probability density functions of the dichotomous and ordered choice variables are assumed to have a logit forms.

4.2.1 The Track Choice at the Start of High School

The track choice takes value on $\{\text{VHS}, \text{THS}-, \text{THS}+, \text{GHS}-, \text{GHS}+\}$, which we relabel for notational convenience and to underline their hierarchical ordering by $\{1, 2, 3, 4, 5\}$. The probability density function of the track choice is an ordered logit determined by a linear index in the strictly exogenous observed explanatory variables $\mathbf{z}_i$, the unobservable determinants of track choice $\mathbf{v}_{i,tr}$ and the initial delay in_i:

$$\Pr(tr_i = 1 | \mathbf{z}_i, \mathbf{v}_{i,tr}, in_i) = \Lambda(\alpha_{1,tr} - \mathbf{v}_{i,tr} - \mathbf{z}_i'\boldsymbol{\beta}_{tr} - in_i\gamma_{tr}),$$
$$\Pr(tr_i = 2 | \mathbf{z}_i, \mathbf{v}_{i,tr}, in_i) = \Lambda(\alpha_{2,tr} - \mathbf{v}_{i,tr} - \mathbf{z}_i'\boldsymbol{\beta}_{tr} - in_i\gamma_{tr}) - \Lambda(\alpha_{1,tr} - \mathbf{v}_{i,tr} - \mathbf{z}_i'\boldsymbol{\beta}_{tr} - in_i\gamma_{tr}),$$
$$\Pr(tr_i = 3 | \mathbf{z}_i, \mathbf{v}_{i,tr}, in_i) = \Lambda(\alpha_{3,tr} - \mathbf{v}_{i,tr} - \mathbf{z}_i'\boldsymbol{\beta}_{tr} - in_i\gamma_{tr}) - \Lambda(\alpha_{2,tr} - \mathbf{v}_{i,tr} - \mathbf{z}_i'\boldsymbol{\beta}_{tr} - in_i\gamma_{tr}),$$
$$\Pr(tr_i = 4 | \mathbf{z}_i, \mathbf{v}_{i,tr}, in_i) = \Lambda(\alpha_{4,tr} - \mathbf{v}_{i,tr} - \mathbf{z}_i'\boldsymbol{\beta}_{tr} - in_i\gamma_{tr}) - \Lambda(\alpha_{3,tr} - \mathbf{v}_{i,tr} - \mathbf{z}_i'\boldsymbol{\beta}_{tr} - in_i\gamma_{tr}),$$
$$\Pr(tr_i = 5 | \mathbf{z}_i, \mathbf{v}_{i,tr}, in_i) = 1 - \Lambda(\alpha_{4,tr} - \mathbf{v}_{i,tr} - \mathbf{z}_i'\boldsymbol{\beta}_{tr} - in_i\gamma_{tr}). \tag{4}$$

The coefficients $\alpha_{1,tr} < \alpha_{2,tr} < \alpha_{3,tr} < \alpha_{4,tr}$ are the ordered threshold parameters.

The nonparametric identification of the selection on unobservables in the track choice at the start of high school requires a different assumption than the one for the other choices. In contrast to the subsequent choices, for the track choice we neither have multiple observations, nor a differential random timing of the choice. Hence, we assume that there is an exclusion restriction.

Assumption 4 *(Exclusion restriction):*
Day of birth in the calendar year is assumed to affect high school choices only through its effect on in_i.

Alet et al. (2013) make a similar assumption. They assume that quarter of birth has a direct impact on both first grade test score and first/second grade repetition, but not on subsequent test scores. They justify this by arguing that the existing literature (Bedard and Dhuey, 2006; Sprietsma, 2010) finds that the relative age has a declining impact over time and that grade repetition is one of the most powerful channels through which relative age may have long-run effect on student achievement. When we include day of birth as a regressor in the 4 subsequent choices, it is statistically significant only the decision to drop-out from school. The p-value of the joint test of significance is 0.034. If we exclude

the day of birth as a regressor from the drop-out equation, the p-value of the joint test of significance increases to 0.75.

4.2.2 The End-of-Year Evaluation

At the end of each academic year, teachers jointly evaluate in a staff meeting the global academic performance of the pupils in the past year. As mentioned in Section 2, students obtain one of the following three scores: A, B or C. An A allows students to be promoted to the next grade. Students getting a C must repeat the grade. Students with a B are imposed to downgrade the track, unless they accept to repeat the grade, in which case they can freely choose to downgrade or not. Because of the natural ordering of these scores, the staff's evaluation choices are modeled as an ordered logit, conditional on both the strictly exogenous observed and unobserved explanatory variables and the past educational choices of pupils and teachers:

$$
\begin{aligned}
\Pr(ev_{it} = C | \mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1}) &= \Lambda\big[\alpha_{1,ev} - \phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1})\big], \\
\Pr(ev_{it} = B | \mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1}) &= \Lambda\big[\alpha_{2,ev} - \phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1})\big] \\
&\quad - \Lambda\big[\alpha_{1,ev} - \phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1})\big], \\
\Pr(ev_{it} = A | \mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1}) &= 1 - \Lambda\big[\alpha_{2,ev} - \phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1})\big], \quad (5)
\end{aligned}
$$

where $\mathbf{x}_i \subset \mathbf{z}_i$ due to the exclusion restriction mentioned in Assumption 4 and $\phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1})$ is a linear index in its arguments, capturing the impact of observed and unobserved determinants, and past educational choices of pupils and teachers.

We consider two different specifications of the linear index $\phi_{ev}(\cdot)$. The first specification ignores *essential* heterogeneity in the effect of the past grade repetition on the evaluation outcome, while the second one explicitly allows for it. In the first specification the linear index takes the following form:

$$
\begin{aligned}
\phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1}) &= \mathbf{x}_i'\boldsymbol{\beta}_{ev} + \mathbf{v}_{i,ev} + in_i\gamma_{ev} + \mathbf{Idow}_{it-1}'\boldsymbol{\pi}_{ev} + \mathbf{Itr}_{it}'\boldsymbol{\eta}_{ev} \\
&\quad + \mathbf{Igr}_{it}'\boldsymbol{\delta}_{ev} + re_{it-1}\kappa_{ev} + pre_{it-1}\psi_{ev} \\
&\equiv \mathbf{w}_{it}'\xi_{ev} + re_{it-1}\kappa_{ev} + pre_{it-1}\psi_{ev}, \quad (6)
\end{aligned}
$$

where

- $\mathbf{v}_{i,ev}$ is unobserved heterogeneity affecting the evaluation;

- $\mathbf{Idow}_{it-1} = [\mathbf{1}_1(dow_{it-1})\ \mathbf{1}_2(dow_{it-1})]'$ is a column vector of two indicators that

18

determine whether the student chooses to downgrade one or two tracks at the end of the previous academic year (the reference student does not change track): $\mathbf{1}_A(x)$ defines the indicator function that is equal to one if $x \in A$ and zero otherwise; $dow_{it} \in \{0, 1, 2\}$ indicates the number of tracks that individual i chooses to downgrade at the end of year t in high school;

- $\mathbf{Itr}_{it} = \left[\mathbf{1}_{\{1\}}\left(tr_i - \mathbf{1}_{\{\forall t:t>1\}}(t)\sum_{s=1}^{t-1} dow_{is}\right) \ldots \mathbf{1}_{\{5\}}\left(tr_i - \mathbf{1}_{\{\forall t:t>1\}}(t)\sum_{s=1}^{t-1} dow_{is}\right)\right]'$ is a column vector of five indicators of the track chosen at the beginning of the t-th year in high school;

- $\mathbf{Igr}_{it} = \left[\mathbf{1}_{\{1\}}\left(t - \mathbf{1}_{\{\forall t:t>1\}}(t)\sum_{s=1}^{t-1} re_{is}\right) \ldots \mathbf{1}_{\{7\}}\left(t - \mathbf{1}_{\{\forall t:t>1\}}(t)\ \sum_{s=1}^{t-1} re_{is}\right)\right]'$ is a column vector of seven indicators of the grade at the beginning of the t-th year in high school;

- re_{it-1} is an indicator variable equal to one if individual i was retained at the end of the previous academic year $t-1$ (repeating the grade therefore in the current year t) and zero otherwise;

- $pre_{it-1} = \mathbf{1}_{\mathbb{N}_0}\left(\mathbf{1}_{\{\forall t:t>2\}}(t)\sum_{s=1}^{t-2} re_{is}\right)$ is an indicator equal to one if the student has ever repeated a grade in high school in years prior to the $(t-1)^{th}$ year.

Hence, we allow that past high school choices ($\Im_{it-1}$) affect the evaluations in a flexible way. The coefficient κ_{ev} is the transitory effect of grade repetition on the subsequent academic performance, while ψ_{ev} is the permanent effect. $\boldsymbol{\delta}_{ev}$ and $\boldsymbol{\eta}_{ev}$ capture that students' ability to get good evaluations depends on the current grade and track, respectively. Finally, $\boldsymbol{\pi}_{ev}$ is the (transitory) effect of having downgraded a track on the academic achievement in the subsequent year.

In the second specification of the linear index $\phi_{ev}(\cdot)$ we allow the short- and long-run effects of grade repetition to be heterogeneous in observed and unobserved abilities. To maintain a tractable model, we simplify by interacting re_{it-1} and pre_{it-1} with the linear index defined in Equation 6. Since this linear index is also a function of unobservables, this allows for *essential* heterogeneity in the treatment effect of grade repetition. The specification of the linear index is then given by the following expression:

$$\phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1}) = \mathbf{w}'_{it}\xi_{ev}\left(1 + re_{it-1}\kappa_{ev}^0 + pre_{it-1}\psi_{ev}^0\right)$$
$$+ re_{it-1}\kappa_{ev}(1 + pre_{it-1}\psi_{ev}^0) + pre_{it-1}\psi_{ev}(1 + re_{it-1}\kappa_{ev}^0) \quad (7)$$

19

If κ_{ev}^0 and ψ_{ev}^0 are jointly equal to 0, then we go back to the first specification of the linear index in Equation (6).

In the last two grades or if a student is in the VHS track, the evaluation is dichotomous, either A or C. In these cases, the ordered logit model described in Equation (5) collapses to a logit model with:

$$\Pr(ev_{it} = A | \mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1}) = 1 - \Lambda\big[\alpha_{2,ev} - \phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_i, \Im_{it-1})\big], \quad (8)$$

and the probability of getting a C is its complement to one.

4.2.3 The School Drop-Out

In Belgium, compulsory education ends on 30 June of the year in which a student turns 18. From that date onwards, students are at risk of high school drop-out without diploma. School drop-out is an interesting long-run outcome of grade repetition that, as mentioned in the Introduction, other authors have considered as well. We model it as a binary choice in the following way for pupils at risk ($s_{it} = 1$):[20]

$$\Pr(out_{it} = 1 | \mathbf{x}_i, \mathbf{v}_{i,out}, in_i, tr_i, \Im_{it-1}, ev_{it}) = \Lambda\left[\alpha_{out} + \phi_{out}(\mathbf{x}_i, \mathbf{v}_{i,out}, in_i, tr_i, \Im_{it-1}, ev_{it})\right], \tag{9}$$

where, similar to Equation (7),

$$\phi_{out}(\mathbf{x}_i, \mathbf{v}_{i,out}, in_i, tr_i, \Im_{it-1}, ev_{it}) = (\mathbf{w}_{it}'\xi_{out} + \mathbf{Iev}_{it}'\boldsymbol{\omega}_{out})(1 + re_{it-1}\kappa_{out}^0 + pre_{it-1}\psi_{out}^0)$$
$$+ re_{it-1}\kappa_{out}(1 + pre_{it-1}\psi_{out}^0) + pre_{it-1}\psi_{out}(1 + re_{it-1}\kappa_{out}^0), \tag{10}$$

with $\mathbf{Iev}_{it} = \begin{bmatrix} \mathbf{1}_{\{A\}}(ev_{it}) & \mathbf{1}_{\{B\}}(ev_{it}) \end{bmatrix}$.[21] Compared to Equation (7), ϕ_{out} has the extra argument, $\mathbf{Iev}_{it}$, i.e. the end-of-year evaluation. By the sequential ordering assumed in Assumption 1, $\mathbf{Iev}_{it}$ is predetermined with respect to the drop-out choice, so that it can be conditioned upon.

---

[20]Very few students (71, 1.7% of the sample) drop-out of school before the end of the academic year. In order to simplify the model and the timing of events, in these cases we advance the drop-out date at the end of the previous academic year, disregarding information on retention and track downgrade of the uncompleted academic year.

[21]Because of the limited number of students at risk of a drop-out decision, estimation was only possible if we grouped students with a B and a C into one category, so that for the drop-out decision the indicator $\mathbf{1}_{\{B\}}(ev_{it})$ was excluded. For similar reasons a coarser grouping was also imposed on $\mathbf{Idow_{it-1}}$, $\mathbf{Itr_{it}}$ and $\mathbf{Igr_{it}}$. See the Internet Appendix (http://users.ugent.be/~bcockx/IA_BCP.pdf) for more details.

4.2.4 The Choice of Repeating the Grade in Case of a B Evaluation

Students getting a B can choose either to repeat the grade or to downgrade the track. The choice is binary and, conditional on getting a B and on not dropping-out, the probability of repeating the grade is specified as follows:

$$
\Pr(re_{it} = 1 | \mathbf{x}_i, \mathbf{v}_{i,re}, in_i, tr_i, \Im_{it-1}, ev_{it} = B, out_{it}=0) =
$$
$$
\Lambda \left[\alpha_{re} + \phi_{re}(\mathbf{x}_i, \mathbf{v}_{i,re}, in_i, tr_i, \Im_{it-1}) \right]. \tag{11}
$$

Because we do not have enough B observations to empirically identify heterogeneous effects of past grade repetition on the current decision to repeat the grade, the function $\phi_{re}(\mathbf{x}_i, \mathbf{v}_{i,re}, in_i, tr_i, \Im_{it-1})$ is parametrized as in Equation (6), excluding thereby the possibility of heterogeneous retention effects across individuals:

$$
\phi_{re}(\mathbf{x}_i, \mathbf{v}_{i,re}, in_i, tr_i, \Im_{it-1}) \quad = \quad \mathbf{w}'_{it}\xi_{re} + re_{it-1}\kappa_{re} + pre_{it-1}\psi_{re}. \tag{12}
$$

Note that the choice of repeating the grade must be considered neither for students in VHS nor for those in the last grade, because these students may never obtain a B (see Section 2).

4.2.5 The Track Downgrade

In Belgium, at the beginning of high school, students can choose among different tracks characterized by different curricula. This tracking system is aimed at grouping students with similar abilities and preferences. Choosing the right track is important as it will determine future work and education opportunities. In Belgium track choice matters particularly, because tracks are hierarchically ordered and students can only move down the hierarchy. The Belgian system of tracking is therefore often referred to as a 'cascade' system.

We model track transitions by defining a categorical ordered dependent variable for track downgrade. As already mentioned in Section 4.2.2, the variable of interest is denoted as $dow_{it} \in \{0, 1, 2\}$. The values reflect the three possible choices: no downgrade, one-step downgrade and two-step downgrade. Students in the VHS track are already at the bottom of the cascade and cannot downgrade further. Consequently, we model track downgrade only for GHS/THS students. The probability density function of a track down-

grade for GHS and THS$^+$ students is specified as:

$$\Pr(dow_{it} = 0|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) =$$
$$\Lambda\big[\alpha_{1,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it})\big],$$
$$\Pr(dow_{it} = 1|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) =$$
$$\Lambda\big[\alpha_{2,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it})\big]$$
$$- \Lambda\big[\alpha_{1,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it})\big],$$
$$\Pr(dow_{it} = 2|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) =$$
$$1 - \Lambda\big[\alpha_{2,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it})\big]. \qquad (13)$$

The function $\phi_{dow}(\cdot)$ is a linear index similar to the one specified in Equation (7):

$$\phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it}) = (\mathbf{w}_{it}'\xi_{dow} + \mathbf{Iev}_{it}'\boldsymbol{\omega}_{dow} + re_{it}\tau_{dow} + re_{it}\mathbf{Iev}_{it}'\boldsymbol{\zeta}_{dow})$$
$$\times (1 + re_{it-1}\kappa_{dow}^0 + pre_{it-1}\psi_{dow}^0)$$
$$+ re_{it-1}\kappa_{dow}(1 + pre_{it-1}\psi_{dow}^0)$$
$$+ pre_{it-1}\psi_{dow}(1 + re_{it-1}\kappa_{dow}^0). \qquad (14)$$

where $re_{it} = 1$ for students with a B evaluation who decided to repeat the grade and $re_{it} = 0$ otherwise. As a consequence of Assumption 1, re_{it} and $\mathbf{ev}_{it}$ are predetermined and, hence, can be conditioned upon. We also allow for interactions between the latter two variables.

For particular groups of students the choice set is reduced. First, students in THS$^-$ cannot make a two-step downgrade: $dow_{it} \in \{0, 1\}$. Hence, for these students the ordered logit reduces to a standard logit:

$$\Pr(dow_{it} = 0|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) =$$
$$\Lambda\big[\alpha_{1,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it})\big], \qquad (15)$$

and the probability of making a one-step downgrade is equal to its complement to one.

Second, students with a B choosing to promote to the next grade are forced to down-

grade, so that $dow_{it} \in \{1, 2\}$.[22] Also in this case the ordered logit simplifies to:

$$\Pr(dow_{it} = 2 | \mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) =$$
$$1 - \Lambda\big[\alpha_{2,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_i, \Im_{it-1}, ev_{it}, re_{it})\big], \qquad (16)$$

and the probability of making a one-step downgrade is equal to its complement to one.

4.3 The Unobserved Heterogeneity Distribution

To maximize the log-likelihood function in Equation (3), we must assign some parametric form to the joint conditional distribution of the unobserved heterogeneity, $G(\mathbf{v}_i | in_i; \boldsymbol{\delta})$. First, we follow Wooldridge (2005) and assume that the correlation between $\mathbf{v}_i$ and in_i affects the mean only and can, hence, be captured by an additive term in the linear indexes of the modeled high school choices. A consequence is that the associated coefficients of in_i cannot be given a structural interpretation.

Second, in our preferred econometric specification, we allow $\mathbf{v}_i$ to be *grade-varying* random pupil-specific unobserved determinants of the outcome variables, similarly to Fruehwirth et al. (2014). The grade-varying unobserved determinants may capture pupils' heterogeneity in reacting to grade specific characteristics of high school, like teachers, subject complexities and subject variation. In contrast to Fruehwirth et al. (2014) who specify the heterogeneity distribution as mixtures of Normal distributions, we avoid too strict parametric assumptions on the distribution of grade-varying unobserved heterogeneity. In line with Heckman and Singer (1984), we assume that the grade-varying unobserved determinants of the different processes have a discrete cumulative distribution function G with a finite and, a priori, unknown number M of points of support. Each support point $\mathbf{v}^m$ of the discrete distribution is a 5×7 matrix, since it is the collection of 5 column vectors, one for each endogenous process, with one support point for each of the 7 grades in high school: $\mathbf{v}^m \equiv [\mathbf{v}_{tr}^m \ \mathbf{v}_{ev}^m \ \mathbf{v}_{out}^m \ \mathbf{v}_{re}^m \ \mathbf{v}_{dow}^m]'$. The first element of $\mathbf{v}^m$, $\mathbf{v}_{tr}^m$, is made up of 7 equal scalars since the track choice at the beginning of high school is observed only in grade 7 and thereby it does not make sense to consider grade-varying heterogeneity for this outcome variable. Since in our sample nobody drops-out of high school before grade 9, grade-varying heterogeneity for school drop-out cannot be identified in the first two grades of high school. The matrix below clarifies the structure of the

---

[22]Students in THS$^-$ who are promoted to next grade are forced to downgrade and, hence, the downgrading choice is not modeled for these students.

m-th point of support of the grade-varying unobserved heterogeneity:

$$\mathbf{v}^m \equiv \begin{bmatrix} \mathbf{v}^m_{tr} \\ \mathbf{v}^m_{ev} \\ \mathbf{v}^m_{out} \\ \mathbf{v}^m_{re} \\ \mathbf{v}^m_{dow} \end{bmatrix} \equiv \begin{bmatrix} v^m_{tr} & - & - & - & - & - & - \\ v^m_{ev,7} & v^m_{ev,8} & v^m_{ev,9} & v^m_{ev,10} & v^m_{ev,11} & v^m_{ev,12} & v^m_{ev,13} \\ - & - & v^m_{out,9} & v^m_{out,10} & v^m_{out,11} & v^m_{out,12} & v^m_{out,13} \\ v^m_{re,7} & v^m_{re,8} & v^m_{re,9} & v^m_{re,10} & v^m_{re,11} & v^m_{re,12} & v^m_{re,13} \\ v^m_{dow,7} & v^m_{dow,8} & v^m_{dow,9} & v^m_{dow,10} & v^m_{dow,11} & v^m_{dow,12} & v^m_{dow,13} \end{bmatrix} .$$

If we leave the structure of the points of support to be unrestricted, we would have to estimate many parameters for each point of support of the discrete distribution. Estimating our model with such a high-dimensional discrete distribution would be computational demanding. It would also be problematic since many support points would be identified on very few observations. For example, the unobserved determinants of grade 13 would be based only on those pupils in the VHS track; the identification of the unobserved determinants of school drop-out in grade 9 and 10 would be based only on 1 and 11 observations, respectively. The most important reason for restricting the dimensions of the points of support comes, however, from Fruehwirth et al. (2014). They demonstrate that by doing it, both *essential* heterogeneity in the treatment effect and time-varying heterogeneity can be accommodated, if both the treatment (i.e. grade repetition) and the outcome of interest (test score in their case) are affected by *common unobserved causes*. Fruehwirth et al. (2014) argue that this is important because lower ability students are more likely to be retained and to learn at a slower pace than high ability students and because unobserved shocks, such as divorce, that affect both selection into grade repetition and the related treatment effect.

We therefore impose the following one-factor structure on the unobserved heterogeneity distribution:

Assumption 5 *(One factor structure):*
For every point of support $m = 1, \ldots, M$, we impose:

- a one-factor loading specification: $v^m_{k,g} = \delta_k \cdot v^m_{ev,g}$, $\forall k \in \{tr, out, re, dow\}$ for $g = 7, \ldots, 13$;

- the equality of the support points of grade 12 and 13, i.e. $v^m_{ev,12} = v^m_{ev,13}$.

Based on Monte Carlo simulations for treatment effects in duration models, Gaure et al. (2007) find that the number M of the points of support is best chosen by minimizing the

Akaike Information Criterion (AIC). We follow this recommendation. The probabilities associated to the points of support sum to one and are specified as logistic transforms:

$$p^m \equiv \Pr(\mathbf{v}_i = \mathbf{v}_i^m) = \frac{\exp\left(\lambda^m\right)}{\sum_{h=1}^{M} \exp\left(\lambda^h\right)} \quad \text{with} \quad m = 1, \ldots, M \quad \text{and} \quad \lambda_M = 0. \quad (17)$$

The sample log-likelihood function in Equation (3) can be rewritten as

$$\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^{N} \ln\left[\sum_{m=1}^{M} p^m \mathcal{L}_{im}(\boldsymbol{\theta}, \boldsymbol{\delta})\right], \quad (18)$$

where $\mathcal{L}_{im}(\boldsymbol{\theta}, \boldsymbol{\delta})$ is the individual contribution to the likelihood function if the individual is of type m.

4.4 Partial Observability of Track Choices at the Start of High School

As mentioned in Section 3, at the beginning of high school (in grade 7), we have only partial information about the school track choice. We only know whether students are in the vocational track (VHS) or not (GHS/THS). Only from grade 8 onward we have detailed information on courses of study, so that we can group students into the five tracks. However, the cascade system of the institutional set-up jointly with the track position and track mobility of each student in subsequent grades convey some information about the possible starting track. For example, students who are in GHS$^+$ in grade 8, surely were also in GHS$^+$ in grade 7, as track upgrading is not allowed. For the same reason, students in GHS$^-$ in grade 8 were not in THS and VHS tracks in grade 7.

We modify the likelihood function to take into account this partial observability of the track at the beginning of high school taking into account the information regarding future track positions as well as the institutional restrictions. We integrate over the possible tracks in grade 7, given future information about tracks and mobility. This is similar to the strategy used by Mroz et al. (2015) to solve the partial observability of the time in which persons with diabetes progress to the next disease stage. However, different from Mroz et al. (2015), the institutional restrictions allow us to exploit more information. We demonstrate below that this leads to weighing in this integration the possible track choices by their probability of occurrence, rather than using uniform weights.

To show in what direction we modify the likelihood function and keep the notation simple, we rewrite the density in Equation (1) by ignoring the conditioning on the ob-

served and unobserved covariates and the individual subscript i, yielding:

$$f(tr, \mathbf{Y}|in) = f(tr|in)f(\mathbf{Y}|tr, in). \tag{19}$$

We assume that the probability of being in each track at the beginning of high school is related to the information we have in the future on tracks, mobility choices and academic achievements. Denote by $f(tr|in, \mathbf{Y})$ this probability density function. If we integrate Equation (19) over the possible tracks, we obtain

$$f(\mathbf{Y}|in) = \int f(tr|in)f(\mathbf{Y}|tr, in)f(tr|in, \mathbf{Y})\mathrm{d}tr. \tag{20}$$

Once we parametrize $f(tr|in)$ and $f(\mathbf{Y}|tr, in)$ as in Subsection 4.2, a particular parametrization of $f(tr|in, \mathbf{Y})$ is implied:

$$
\begin{aligned}
f(tr|in, \mathbf{Y}) &= \frac{f(\mathbf{Y}|tr, in)f(tr|in)}{f(\mathbf{Y}|in)} \\
&= \frac{f(\mathbf{Y}|tr, in)f(tr|in)}{\int f(\mathbf{Y}, tr|in)\mathrm{d}tr} \\
&= \frac{f(\mathbf{Y}|tr, in)f(tr|in)}{\int f(\mathbf{Y}|tr, in)f(tr|in)\mathrm{d}tr}.
\end{aligned} \tag{21}
$$

where the first equality follows from Bayes' rule, the second one from the definition of a marginal density, and the third one from the chain rule. Both the numerator and the denominator of Equation (21) depend indeed on the probability density functions that we have already parametrized in Subsection 4.2. Substituting Equation (21) into Equation (20) yields

$$f(\mathbf{Y}|in) = \int \frac{f(tr|in)^2 f(\mathbf{Y}|tr, in)^2}{\int f(\mathbf{Y}|s, in)f(s|in)\mathrm{d}s}\mathrm{d}tr. \tag{22}$$

Since tracks take value on five categories, the integrals in Equation (22) are just sums over the five possible realisations. The sample log-likelihood function in Equations (3) and (18) are modified along the lines dictated by Equation (22).

4.5   Identification

The identification of the interrelated dynamics between grade retention, track mobility and schooling attainment is difficult. In this subsection we summarize our arguments why we believe that the aforementioned characteristics of our model ensure identification.

In our analysis the timing of the central treatment of interest, grade retention (i.e. a C in the end-of-year evaluation), is not fixed, but (partially) randomly determined. The counterfactual outcomes that we are interested in (the end-of-year evaluations, the drop-out of high school, the subsequent track choices and the schooling delay) are explicitly associated to these treatment times. Heckman and Navarro (2007) provide general identification results for these types of models which do not require exclusion restrictions.[23] Fruehwirth et al. (2014) tailor these specifically to the evaluation of grade repetition. The *no anticipations* assumption (Abbring and van den Berg, 2003) and the assumption of common (to the treatment and the outcome) unobserved factors are central in the identification of these *factor analytic dynamic models* (FADM). While we discussed the latter assumption in Section 4.3, we did not yet consider the former.

The *no anticipations* assumption means that, after conditioning on all prior information, grade repetition in grade $g + k$ ($\forall g \in \{7, 8, ..., 13\}$ and $\forall k : k > 0$) may not directly affect pupil's schooling outcomes in grade g. This assumption is actually implicitly implied by the *no sequentiality* Assumption 1, because it means that the end-of-year evaluation which may result in a retention (in case of a C) always precedes the schooling outcomes of the subsequent year, such as drop-out, downgrading or the evaluation at the end of the subsequent year. We believe that it is reasonable to assume that pupils cannot be sure about retention before the end-of-year evaluation and, hence, cannot act in anticipation of that. A similar reasoning applies when we consider different treatments, such as receiving a B instead of a C as the end-of-year evaluation.

Our model is more complicated than those considered by Heckman and Navarro (2007) and Fruehwirth et al. (2014). First, as just mentioned, we do not only consider a single treatment, i.e. retention (C) versus no retention (A), but also other treatments, such as retention (C) versus being forced to downgrade (B). Since these treatments are essentially determined by the same timing, they should not affect the identification arguments. Second, in the aforementioned studies it is assumed that the treatment (retention) only realizes once, while in the institutional framework that we consider, the same pupil may be retained multiple times. Abbring and van den Berg (2003) prove in a similar framework as ours, but in continuous instead of discrete time, that the presence of *multiple-spell* data, which corresponds to multiple treatments in our case, relaxes the requirements on the identifying assumptions. The intuition is similar to the reasoning that having a panel data set with multiple observations per unit requires fewer identifying assumptions than

---

[23]The methods of Heckman and Navarro (2007) are related to the *timing-of-events* method proposed by Abbring and van den Berg (2003). The former is expressed in discrete time whereas the latter in continuous time.

a single cross-section. We therefore believe that observing multiple treatments should facilitate identification rather than impeding it. Third, the track choice at the beginning of high school is a static choice, so that potential selectivity based on unobservables cannot be credibly identified on the basis of the aforementioned identification strategy. The one-factor structure imposed on the unobserved heterogeneity distribution in Assumption 5 is sufficient for identification, but this would then crucially depend on this functional form restriction. We therefore believe that the exclusion restriction imposed in Assumption 4 is a firmer basis for identification. Fourth, in contrast to the aforementioned authors we face an initial conditions problem, because the process of grade retention can start already in kindergarten or primary school, while we observe timing of retentions only from the start of high school. As mentioned in Section 4.1, we solve this initial conditions problem following the methods proposed by Wooldridge (2005). Finally, since the conditional density functions in Equation (22) are identified on the basis of the aforementioned arguments, the partial observability problem in Subsection 4.4 should not inhibit identification.

5 Empirical Results

The estimated parameters of the econometric model are reported in an Internet Appendix which can be downloaded from http://users.ugent.be/~bcockx/IA_BCP.pdf. We present the estimates of three models: without unobserved heterogeneity, with grade-constant and with grade-varying unobserved heterogeneity. As mentioned in Section 4.3, the number of points of support of the discrete one-factor heterogeneity distribution are chosen such that the AIC is minimized. The resulting number of support points M is 3, both for the specification controlling for grade-constant and for grade-varying unobserved heterogeneity. According to the AIC the latter model is to be preferred. Hence, in this model is chosen as benchmark in the discussion below.

Since the marginal effects are complicated and their direction and magnitude are not determined entirely by single estimated coefficients, we prefer not to discuss the estimated parameters in the main text. Instead, we simulated the model under different counterfactual scenarios of interest to infer the average treatment effects on the treated (ATT) of retentions, relative to unconstrained or constrained promotions, i.e. A or B, in different moments of the high school career on future educational outcomes (Subsection 5.2). However, before reporting these counterfactual simulations, we first use the simulations to check the performance of our model in terms of goodness-of-fit (Subsection 5.1).

The simulations were conducted as follows. We randomly drew 999 vectors from the

asymptotic Normal distribution of the model parameters. This ensures that the simulations also capture the uncertainty due to estimation. Subsequently, in each of the 999 simulations the drawn parameters were used to calculate the probabilities associated to each heterogeneity type. These probabilities were then used to randomly assign to each pupil in the sample a heterogeneity type. Thereafter, based on these random draws of parameters and heterogeneity types, the full sequence of high school decisions from track choice at the start of high school until drop-out or high school graduation was simulated for each pupil in the sample. Each choice was simulated sequentially, based on the chosen logit specifications reported in Subsection 4.2. To determine the outcome of the choice a random draw from the standard and, across the sample, independent uniform distribution was compared to the thresholds that the simulated ordered logits imply. These thresholds determine segments on the unit interval that correspond to particular choices. The assigned choice depends on the segment in which the random number falls. Once a choice is a assigned it is saved and conditioned upon in the subsequent choice. In each of the 999 simulations this procedure ends if all sampled individuals either have dropped-out or graduated from high school.

In the sequel, the *model prediction* of a particular outcome refers to the average of these 999 simulations. The empirical percentiles provide estimates of the thresholds of the 95% confidence intervals (CI). Average treatment effects on the treated (ATTs) are estimated by the average difference between two (counterfactual) outcomes, one in which the treatment (e.g. a C obtained in a particular grade) is imposed and one in which a counterfactual treatment (e.g. an A or a B) applies. The average is taken over the individuals who in the simulation are treated.

5.1 Goodness-of-fit

Table 4 provides an insight into the goodness-of-fit of our preferred model, i.e. the model with grade-varying unobserved heterogeneity. It allows us to answer the question whether the simulated proportions, based on the aforementioned parameter sets and simulation procedure, predict the observed proportions in the data well. More concretely, simulated and actual fractions are compared (*I-III*) for school outcomes in grades 8 to 10, i.e. obtaining an A in the end-of-year evaluation (broken down by whether or not one repeated the previous grade for grades 9 and 10) and repeating the grade (forced after a C or chosen after a B); (*IV*) high school graduation; and (*V*) years of schooling delay in the last compulsory schooling year, i.e. in the school year that ends in the calendar year that the student becomes 18. We selected these outcomes for a goodness-of-fit analysis, because

Table 4: Goodness-of-fit

	Model: grade-varying unobserved heterogeneity		
	Actual probability	Simulated probability	95% CI
I. Outcomes in grade 8			
A	0.900	0.895	[0.881, 0.909]
Repeating the grade	0.040 *	0.048	[0.039, 0.057]
II. Outcomes in grade 9			
A	0.923	0.916	[0.904, 0.929]
Repeating the grade	0.047	0.045	[0.036, 0.054]
A, if repeated grade 8	0.828	0.829	[0.770, 0.888]
A, if did not repeat grade 8	0.927	0.921	[0.908, 0.933]
III. Outcomes in grade 10			
A	0.908	0.902	[0.887, 0.915]
Repeating the grade	0.061	0.060	[0.050, 0.071]
A, if repeated grade 9	0.863	0.808	[0.734, 0.871]
A, if did not repeat grade 9	0.910	0.906	[0.892, 0.919]
IV. High school graduation			
Diploma	0.915	0.919	[0.906, 0.931]
Diploma, if repeated grade 8	0.720	0.751	[0.675, 0.822]
Diploma, if did not repeat grade 8	0.923	0.927	[0.915, 0.939]
Diploma, if repeated grade 9	0.798	0.796	[0.722, 0.861]
Diploma, if did not repeat grade 9	0.921	0.925	[0.913,0.937]
Diploma, if repeated grade 10	0.825	0.809	[0.744, 0.869]
Diploma, if not repeated grade 10	0.921	0.926	[0.914, 0.938]
V. Years schooling delay at start last compulsory year			
Diploma	0.229	0.239	[0.219, 0.260]
Delay, if repeated grade 8	1.261	1.250	[1.173, 1.335]
Delay, if did not repeat grade 8	0.186	0.188	[0.171, 0.207]
Delay, if repeated grade 9	1.202	1.250	[1.169, 1.331]
Delay, if did not repeat grade 9	0.181	0.192	[0.174, 0.210]
Delay, if repeated grade 10	1.200	1.202	[1.144, 1.264]
Delay, if did not repeat grade 10	0.166	0.177	[0.160, 0.194]

Notes: All predictions are based on 999 simulations that allow for the uncertainty of the estimated parameters. The model prediction (*simulated probability*) is calculated as the average outcome for these simulations. ***, **, * indicate a significant difference between prediction and actual outcome at the 1%, 5%, 10% significance levels, respectively.

these are the relevant ones in the counterfactual simulations. We report for each of these outcomes the predicted sample average, as well as the predicted sample average for the subgroups that repeated grades in one of the preceding years and the groups that did not. For instance, for the school outcomes in grade 9 we not only report the sample averages, but also those of the subgroups that were retained in the previous year, i.e. in grade 8.

Actual and simulated probabilities for the mentioned outcomes are very comparable. Only with respect to repeating the grade after the first year in grade 2, the actual probability (4.0%) is different from the simulated probability (4.8%) at a 10% level of significance. Therefore, we conclude that our model captures the dynamic choices in high school very well.

5.2 ATTs Based on Counterfactual Simulations

In this subsection, we answer our research questions. We do this by presenting ATTs with respect to grade repetition, both with a short-term focus (effect in the subsequent grade, i.e. two years after being retained) and with a long-term focus (effect on high school graduation and on delay at the start of the last compulsory schooling year). In a first step, we present our benchmark simulation results, i.e. the ATTs of being retained (i.e. obtaining a C) in grade 8 (the first observed retention in high school) relative to be promoted to the next grade (i.e. obtaining an A). In a second step, we contrast the results obtained with our benchmark model with grade-varying unobserved heterogeneity to those obtained if the heterogeneity is constant over grades or if unobserved heterogeneity is completely ignored. In a third step, we consider alternative dimensions of the effect of retention: (i) Does the timing of retention matters (grade 8 compared to grades 9 and 10)? (ii) Does it affect school track choices? (iii) Is the effect heterogeneous across the treated population? Finally, we contrast grade retention with the alternative remedial strategy of forced track downgrading, i.e. we contrast a C with a B.

5.2.1 Benchmark: Short- and Long-Term Effects of Retention in Grade 8

Table 5 presents the average effect of retention in grade 8. In the short-term, the impact on the evaluation in the next grade, i.e. in grade 9, is virtually zero. This means that grade-repetition does not improve the academic performance and is therefore ineffective. In the long-term this adverse outcome is reinforced, because those who were retained in grade 8 are 13.5 percentage points less likely to graduate from high school with a diploma. In addition, these pupils catch up only marginally relative to the counterfactual

Table 5: Benchmark Simulations: The Effects of Retention in Grade 8

Treatment: C versus A evaluation after grade 8	Model: grade-varying unobserved heterogeneity		
	ATT		95% CI
I. Evaluation in grade 9: A			
All treated	0.001		[-0.092, 0.099]
II. High school diploma			
All treated	-0.135	***	[-0.233, -0.043]
III. Delay at start last compulsory year			
All treated	0.904	**	[0.817, 0.991]

Notes: All statistics are based on 999 random simulations of the treated sample that allow for the uncertainty of the estimated parameters. The ATTs are calculated by subtracting the average outcome in case of the counterfactual of *no* grade retention from the average outcome in case of a retention. ***, **, * indicate whether the ATT is significantly different from 0 (1) in panels *I* and *II* (panel *III*) at the 1%, 5%, 10% significance levels, respectively.

of no retention: their delay at the start of the last compulsory schooling year is only slightly lower than one year. This means that, if the student would have been promoted, she would not face a particularly higher risk to be retained in a subsequent year than if she was initially retained.

5.2.2 The Impact of (Grade-Varying) Unobserved Heterogeneity

Comparing Table 5 and Table 6 allows us to judge the bias in the treatment effects when not controlling for (grade-varying) unobserved heterogeneity. On the one hand, this shows the standard result that not controlling for unobserved heterogeneity biases in the direction of more adverse ATTs. On the other hand, this bias seems to be slightly over-corrected when unobserved heterogeneity is assumed to be time constant rather than grade-varying.

Table 6: ATTs of Retention in Grade 8 for Alternative Controls for Unobserved Heterogeneity

Treatment: C versus A evaluation after grade 8	Model: without unobserved heterogeneity			Model: grade-constant unobserved heterogeneity		
	ATT		95% CI	ATT		95% CI
I. Evaluation in grade 9: A						
All treated	-0.066		[-0.160, 0.024]	0.026		[-0.074, 0.120]
II. High school diploma						
All treated	-0.168	***	[-0.248, -0.086]	-0.126	***	[-0.203, -0.049]
III. Delay at start last compulsory year						
All treated	1.023		[0.948, 1.102]	0.850	***	[0.744, 0.940]

Notes: All statistics are based on 999 random simulations of the treated sample that allow for the uncertainty of the estimated parameters. The ATTs are calculated by subtracting the average outcome in case of the counterfactual of *no* grade retention from the average outcome in case of a retention. ***, **, * indicate whether the ATT is significantly different from 0 (1) in panels *I* and *II* (panel *III*) at the 1%, 5%, 10% significance levels, respectively.

5.2.3 Alternative Dimensions of the Effect of Grade Retention

In this subsection, we inspect whether the treatment has a different effect when occurring later, whether it also affects track outcomes and whether it is heterogeneous by the ability of the pupils. First, we investigate whether the timing of grade retention matters. What happens if students are retained in a later grade: grade 9 or 10 instead of grade 8? Table 7 presents the results. These differ only marginally from those presented in Table 5. This suggests that grade retention induces a psychological shock with, irrespectively of the repeated grade, an *immediate* adverse impact: if, by contrast, the adverse impact would have gradually built-up, then the long-term effects of retention in later grades should have been smaller than in earlier grades.

Table 7: ATTs of Grade Retention: Alternative Timing of Retention

| | Model: grade-varying unobserved heterogeneity | | | | |
| | Treatment: C versus A evaluation after grade 9 | | | Treatment: C versus A evaluation after grade 10 | |
	ATT		95% CI	ATT	95% CI
I. Evaluation in next grade: A					
All treated	-0.010		[-0.121, 0.098]	–	–
II. High school diploma					
All treated	-0.123	***	[-0.219, -0.031]	-0.139 ***	[-0.228, -0.055]
III. Delay at start last compulsory year					
All treated	0.878	***	[0.784, 0.957]	0.868 ***	[0.798, 0.930]

Notes: All statistics are based on 999 random simulations of the treated sample that allow for the uncertainty of the estimated parameters. The ATTs are calculated by subtracting the average outcome in case of the counterfactual of *no* grade retention from the average outcome in case of a retention. Panel *I* is empty for treatments in grade 10 since not all individuals reach grade 11 (and therefore outcomes for the latter year cannot be calculated for all individuals). ***, **, * indicate whether the ATT is significantly different from 0 (1) in panels *I* and *II* (panel *III*) at the 1%, 5%, 10% significance levels, respectively.

Table 8 reports the effect of grade retention in grade 8 on track choices, both in the short-term (in grade 9) and in the long-term (in the last compulsory schooling year). As to facilitate the presentation, we group the five tracks into three. The findings show that retention is increasing the likelihood of a track downgrade, and this already in the subsequent grade. In case of retention, the likelihood of being in the general track (GHS) falls by as much as 18 percentage points, while the likelihood of being in the vocational track increases by as much as 13 percentage points. Both effects are highly significant. This might be related to the adverse effect of retention on pupils' self-esteem mentioned in the Introduction of this study. In the long-term, i.e. when we consider the impact of retention on the track choice in the last compulsory schooling year, we observe that the negative effect on being in the general track is about 5 percentage points less negative than in the short-term. This suggests that part of the pupils that were retained in the general

track, would downgrade anyway subsequently, even if they would not have been retained in grade 8. By contrast, the impact on those in the vocational track does not change over time.

Table 8: ATTs of Grade Retention in Grade 8 on Track Choice

Treatment: C versus A evaluation after grade 8	Model: grade-varying unobserved heterogeneity		
	ATT		95% CI
I. Track at start of grade 9: GHS			
All treated	-0.184	***	[-0.277, -0.106]
II. Track at start of grade 9: THS			
All treated	-0.051		[-0.054, 0.165]
III. Track at start of grade 9: VHS			
All treated	0.132	***	[0.071, 0.210]
IV. Track in last compulsory year: GHS			
All treated	-0.136	***	[-0.217, -0.067]
V. Track in last compulsory year: THS at least once			
All treated	-0.005		[-0.106, 0.102]
VI. Track in last compulsory year: VHS			
All treated	0.142	***	[0.066, 0.220]

Notes: All statistics are based on 999 random simulations of the treated sample that allow for the uncertainty of the estimated parameters. The ATTs are calculated by subtracting the average outcome in case of the counterfactual of *no* grade retention from the average outcome in case of a retention. ***, **, * indicate whether the ATT is significantly different from 0 at the 1%, 5%, 10% significance levels, respectively.

Similar to Fruehwirth et al. (2014) and Gary-Bobo et al. (2014) we allow the treatment effect to depend on observed and unobserved characteristics, i.e. we allow for *essential* heterogeneity. However, in contrast to the aforementioned authors, we do not consider the treatment heterogeneity for the full sample, but only for the subsample of treated (i.e. retained) pupils. We believe that it would be problematic with our data to estimate the effect of retention on the untreated subsample, because many of them are very unlikely to be retained: there is no common support.

To obtain some insights in the extent of treatment heterogeneity, we ordered the treated individuals in each simulation according to the quartiles of the linear index of the evaluation at the end of grade 8 as defined by Equation (7). In Table 9 we report the average treatment effects of a retention in grade 8 by these quartiles. In line with the research based on RDD, we find that students who are at the margin of being retained, i.e. those in the fourth quartile, increase their academic achievement in the subsequent grade, be it only significant at the 10% level. The ATTs on high school graduation and delay are also less adverse for this quartile. As Fruehwirth et al. (2014), we find that retention of the lowest ability students decreases the academic achievement and drives the significantly positive effect on high school drop-out and on years of delay in the last compulsory

schooling year.[24] Finally, panel *III* also suggests that the high ability pupils significantly reduce the schooling delay by about 20%, while the low ability students do not catch-up at all.

Table 9: Treatment Heterogeneity of Retention in Grade 8

Treatment: C versus A evaluation after grade 8	Model: grade-varying unobserved heterogeneity		
	ATT		95% CI
I. Evaluation in grade 9: A			
First quartile	-0.156		[-0.357, 0.037]
Second quartile	0.001		[-0.149, 0.142]
Third quartile	0.057		[-0.087, 0.171]
Fourth quartile	0.101	*	[-0.009, 0.206]
II. High school graduation			
First quartile	-0.259	***	[-0.451, -0.082]
Second quartile	-0.119		[-0.302, 0.034]
Third quartile	-0.064		[-0.224, 0.076]
Fourth quartile	-0.096		[-0.260, 0.050]
III. Delay at start last compulsory year			
First quartile	1.060		[0.876, 1.271]
Second quartile	0.904		[0.750, 1.074]
Third quartile	0.844	**	[0.694, 0.996]
Fourth quartile	0.806	**	[0.662, 0.957]

Notes: All statistics are based on 999 random simulations of the treated sample that allow for the uncertainty of the estimated parameters. The ATTs are calculated by subtracting the average outcome in case of the counterfactual of *no* grade retention from the average outcome in case of a retention. The first quartile is the one with the lowest value for the linear index of the evaluation in grade 8. ***, **, * indicate whether the ATT is significantly different from 0 (1) in panels *I* and *II* (panel *III*) at the 1%, 5%, 10% significance levels, respectively.

5.2.4 The Effect of Grade Repetition Relative to Forced Downgrading

In the Flemish setting there are three (A, B and C) instead of two (pass or fail) evaluation outcomes. This allows to consider the relative performance of grade repetition (C) to forced downgrading (B)[25] as alternative remediation strategy. In Table 10 we present, as in the previous tables, the ATTs of repeating grade 8. However, instead of contrasting grade retention to an unconstrained promotion to the next grade (i.e. an A), we now compare it to obtaining a forced downgrade (i.e. a B). In the short run, a forced downgrade is a better, though statistically insignificant, strategy than imposing grade repetition. The likelihood of getting an A in the subsequent grade is 4 percentage points higher in case the student

---

[24]Gary-Bobo et al. (2014) by contrast find that the lowest ability students may gain in terms of short-run academic achievement. However, in line with our findings on drop-out, their likelihood of accessing grade nine decreases significantly, and more so than higher ability students.

[25]Recall that downgrading can be avoided by repeating the grade.

is awarded a B instead of a C. In the short-run, awarding a B seems even more effective than an A. In the benchmark contrast reported in Table 5, students with a C and an A were equally likely to be unconditionally promoted in the subsequent grade. This is, however, not surprising. In the lower track the academic requirements for passing the grade are set to a lower level.

Table 10: Grade Repetition (C) Relative to Forced Downgrading (B) after Grade 8

Treatment: C versus B	Model: grade-varying unobserved heterogeneity	
	ATT	95% CI
I. Evaluation in grade 9: A		
All treated	-0.040	[-0.120, 0.039]
II. High school diploma		
All treated	-0.050	[-0.140, 0.035]
III. Delay at start last compulsory year		
All treated	0.727 ***	[0.602, 0.848]

Notes: All statistics are based on 999 random simulations of the treated sample that allow for the uncertainty of the estimated parameters. The ATTs are calculated by subtracting the average outcome in case of the counterfactual of *no* grade retention from the average outcome in case of a retention. ***, **, * indicate whether the ATT is significantly different from 0 at the 1%, 5%, 10% significance levels, respectively.

In the longer run, for students who were actually forced to repeat the grade, the advantage of awarding a B rather than an A disappears, but it remains a slightly, better strategy than awarding a C. Awarding a B instead of a C increases the likelihood of high school graduation by 5 percentage points, but not significantly, while the accumulated delay in the last compulsory schooling year is, significantly, about 0.7 years shorter. Compared to the baseline which contrasts a C to an A (Table 5), remediating by a forced downgrade (a B) does not improve relative to the alternative of allowing promotion to the next grade (an A).

6 Conclusion

We empirically analyzed the short- and long-term effects of grade retention in high school on academic performance and high school graduation. To this end we set up a dynamic discrete model that captures all major high school choices in Flanders, the Dutch speaking region in the North of Belgium. Based on a rich survey of a sample of pupils born in 1978 and 1980, we estimated the school choices of these high school students based on a factor analytic dynamic discrete choice model (Carneiro et al., 2003; Heckman and Navarro,

2007). In contrast to regression discontinuity designs, this approach allows to capture treatment heterogeneity and to control for grade-varying unobservable determinants. We contributed to the literature by proposing a method to deal with initial conditions and take partial observability of track choice at the start of high school into account. Moreover, we considered *forced* track downgrading as an alternative remedial measure.

Since the estimation results could not be directly interpreted, counterfactual simulations of the model were used to obtain estimates of the aforementioned treatment effects. Even if our results indicate that grade retention leads to neutral effects on academic achievement in the short-run, in the long-run grade retention has adverse effects, because it leads the higher drop-out rates, substantial schooling delay and downgrading within the hierarchical tracking system in Flemish high school. In line with the findings of Fruehwirth et al. (2014) and Gary-Bobo et al. (2014), we also found substantial heterogeneity in the effects of grade retention relative to allowing the student to be unconditionally promoted to the next grade. Lower ability students are clearly more adversely affected than those with higher ability. This is important, because it explains why other studies using RDD, mentioned in the Introduction, obtained more favorable results. RDDs identify the treatment effect of higher ability students who are on the margin of being retained. Finally, our study finds that the alternative remedial measure used in the Flemish schooling system, namely *forced* downgrading, can significantly improve academic performance only in the short-run. In the longer run this alternative reduces, relative to grade retention, the schooling delay and the high-school drop-out, although not significantly for the latter. However, these long-run gains disappear when comparing this remediation strategy to the alternative of having these students unconditionally pass to the next grade. Hence, the challenge remains to find more successful remediation strategies.

References

Abbring, J.H., and G.J. van den Berg (2003) 'The nonparametric identification of treatment effects in duration models.' *Econometrica* 71(5), 1491–1517

Akay, A. (2012) 'Finite-sample comparison of alternative methods for estimating dynamic panel data models.' *Journal of Applied Econometrics* 27(7), 1189–1204

Alet, E., L. Bonnal, and P. Favard (2013) 'Repetition: Medicine for a short-run remission.' *Annals of Economics and Statistics* 111, 227–250

Alexander, K.L., D.R. Entwisle, and S.L. Dauber (1994) *On the success of failure* (New York: Cambridge University Press)

Allen, C. S., Q. Chen, V. L. Willson, and J. N. Hughes (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., and B. Cockx (2013) 'Pure ethnic gaps in educational attainment and school to work transitions: When do they arise?' *Economics of Education Review* 4, 185–223

Bedard, K., and E. Dhuey (2006) 'The persistence of early childhood maturity: International evidence of long-run age effects.' *The Quarterly Journal of Economics* 121(4), 1437–1472

Belot, M., and V. Vandenberghe (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

Brodaty, T. O., R. J. Gary-Bobo, and A. Prieto (2013) 'Does speed signal ability? The impact of grade retention on wages.' mimeo, THEMA, CREST-ENSAE, and CNRS, France

Browman, L.J. (2005) 'Grade retention: Is it a help or hindrance to student academic success?' *Preventing School Failure* 49(3), 42–46

Byrd, R.S., M. Weitzman, and P. Auinger (1997) 'Increased behavior problems associated with delayed school entry and delayed school progress.' *Pediatrics* 100(4), 654–661

Carneiro, P., K. Hansen, and J. J. Heckman (2003) 'Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice.' *International Economic Review* 44(2), 361–422

Depew, B., and O. Eren (2015) 'Test based promotion policies, dropping out, and juvenile crime.' Working Paper 2015-07, Department of Economics, Louisiana State University

D'Haultfœuille, X. (2010) 'A new instrumental method for dealing with endogenous selection.' *Journal of Econometrics* 154(1), 1–15

Dong, Y. (2010) 'Kept back to get ahead? Kindergarten retention and academic performance.' *European Economic Review* 54, 219–236

Eide, E.R., and D.D. Goldhaber (2005) 'Grade retention: What are the costs and benefits?' *Journal of Education Finance* 31(2), 195–214

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

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

Fruehwirth, J.C., S. Navarro, and Y. Takahashi (2014) 'How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects.' *forthcoming in Journal of Labor Economics*

Gamoran, A., M. Nystrand, M. Berens, and P.C. LePore (1995) 'An organizational analysis of the effects of ability grouping.' *American Educational Research Journal* 32(4), 687–715

Gary-Bobo, R.J., M. Goussé, and J.-M. Robin (2014) 'Grade retention and unobserved heterogeneity.' CESifo Working Paper Series 2014-06, CESifo

Gaure, S., K. Røed, and T. Zhang (2007) 'Time and causality: A Monte Carlo assessment of the timing-of-events approach.' *Journal of Econometrics* 141(2), 1159–1195

Greene, J.P., and M.A. Winters (2007) 'Revisiting grade retention: An evaluation of Florida's test-based promotion policy.' *Education Finance and Policy* 2(4), 319–340

Heckman, J.J. (1981) 'The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process.' In *Structural analysis of discrete data with econometric applications,* ed. C. Manski and D. McFadden (Cambridge: MIT Press) chapter 4, pp. 179–195

Heckman, J.J., and B. Singer (1984) 'A method for minimizing the impact of distributional assumptions in econometric models for duration data.' *Econometrica* 52(2), 271–320

Heckman, J.J., and S. Navarro (2007) 'Dynamic discrete choice and dynamic treatment effects.' *Journal of Econometrics* 136(2), 341–396

Heckman, J.J., S. Urzua, and E. J. Vytlacil (2006) 'Understanding instrumental variables in models with essential heterogeneity.' *Review of Economics and Statistics* 88(3), 389–432

Holmes, C. T. (1989) 'Grade level retention effects: A meta-analysis of research studies.' In *Flunking Grades: Research and Policies on Retention,* ed. L. A. Shepard and M. L. Smith (New York: The Falmer Press) pp. 16–33

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

— (2009) 'The effect of grade retention on high school completion.' *American Economic Journal: Applied Economics* 1(3), 33–58

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

Mroz, T., G. Picone, F. Sloan, and A. Y. Yashkin (2015) 'Screening for a chronic disease: A multiple state duration model with endogenous treatment.' *forthcoming in International Economic Review*

OECD (2004) *Learning for tomorrow's world – First results from PISA 2003* (Paris: OECD Publications)

— (2012) *Equity and quality in education: Supporting disadvantaged students and schools* (Paris: OECD Publications)

Schwerdt, G., M. R. West, and M. A. Winters (2015) 'The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida.' NBER Working Paper 21509, NBER

Sprietsma, M. (2010) 'Effect of relative age in the first grade of primary school on long-term scholastic results: International comparative evidence using PISA 2003.' *Education Economics* 18(1), 1–32

Van de gaer, E., H. Pustjens, J. Van Damme, and A. De Munter (2006) 'Tracking and the effects of school-related attitudes on the language achievement of boys and girls.' *British Journal of Sociology of Education* 27(3), 293–309

Van Houtte, M., J. Demanet, and P.A.J. Stevens (2012) 'Self-esteem of academic and vocational students: Does within-school tracking sharpen the difference?' *Acta Sociologica* 55(1), 73–89

Vuong, Q.H. (1989) 'Likelihood ratio tests for model selection and non-nested hypotheses.' *Econometrica* 57(2), 307–333

Wooldridge, J.M. (2005) 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity.' *Journal of Applied Econometrics* 20(1), 39–54