Modeling the Effects of Grade Retention in High School

Bart Cockx* Matteo Picchio† Stijn Baert‡

Abstract

A dynamic discrete choice model is set up to estimate the effects of grade retention in high school, both in the short-run (end-of-year evaluation) and in the long-run (drop-out and delay). In contrast to other evaluation approaches, this model captures *essential* treatment heterogeneity and controls for grade-varying unobservable determinants. In addition, forced track downgrading is considered as an alternative remedial measure. Our results indicate that grade retention has a neutral effect on academic achievement in the short-run. In the long-run, grade retention, just like forced downgrading, has adverse effects on schooling outcomes and, more so, for less able pupils.

Keywords: Education, grade retention, track mobility, dynamic discrete choice models, heterogeneous treatment effects.

JEL classification codes: C33, C35, I21.

*Corresponding author. Department of Economics, Ghent University; IRES, Université catholique de Louvain; IZA; CESifo. Sint-Pietersplein 6, 9000 B–Ghent, Belgium. Tel.:+32.9.264.78.77. E-mail: Bart.Cockx@UGent.be.

†Department of Economics and Social Sciences, Marche Polytechnic University; Sherppa, Ghent University; IZA; GLO.

‡Department of Economics, Ghent University; Research Foundation – Flanders; University of Antwerp; Université catholique de Louvain; IZA; GLO; IMISCOE.

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 insight into this debate by estimating the impacts 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

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

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) academic 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]

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, which 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.[6]

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 (s)he attended. The imposed downgrading is an alternative remedial

[2] Jacob and Lefgren (2004) and Jacob and Lefgren (2009) use RDD to evaluate respectively short- and long-run effects of grade repetition for students 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 and 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. (2016) study the effect of retention in the US between kindergarten and fourth grade. Gary-Bobo et al. (2016) consider sixth to eighth graders in France.

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

measure that avoids grade repetition by reorienting the student to a less demanding track. 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). More concretely, OECD (2012) compared 34 OECD countries with respect to this differentiation. It turned out that in 19 of these countries students were clustered in different – and more or less hierarchical – tracks, based on their ability and interests. As a consequence, 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: track choice (only at the start of high school) and subsequently, each year, the decisions to downgrade or not; to repeat the grade or not; and to drop-out of school or not (only at ages when schooling is no longer compulsory). These students' decisions are each time taken conditional on the teachers' overall end-of-year evaluations, which are also explained in our model. These teachers' evaluations and aforementioned decisions are also allowed to depend on past schooling outcomes.

Our approach fits into the aforementioned FADM and is, hence, most comparable to the work of Fruehwirth et al. (2016) (henceforth FNT) and Gary-Bobo et al. (2016) (henceforth GGR). A key assumption in this approach is that unobserved determinants of both treatment and outcomes are assumed to be captured by a *low dimensional set of common causes* (FNT, p. 996). Since we do not have separate measurements of the multidimensional unobserved ability as FNT have in their data, we restrict, as GGR, the factor structure of these unobservables to be *uni*dimensional. The main advantages of the FADM approach are that (i) 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); (ii) in contrast to studies that analyze cross-sections, it can take the dynamic selection into account which prior retention decisions and other schooling outcomes induce in subsequent treated and untreated groups; (iii) it can allow for selection on *time-varying un*observed determinants, which matters to avoid bias induced by changes in the environment (school, classroom and peer effects) over grades or time that cannot be controlled for in the data;[7] and (iv) in contrast with RDD, it can identify non-local treatment effects of individuals at some distance from performance thresh-

[7] In contrast to FNT, who can take *year*-varying heterogeneity into account, our data can only identify *grade*-varying heterogeneity (see Section 4.3). GGR assume that the unobserved heterogeneity is constant.

olds. Taking these issues into account matters. Studies based on FADM, including ours, find, as studies based on RDD, that pupils at the margin of retention *benefit* from retention in the short-run. However, the effect of retention differs for inframarginal and untreated pupils. FNT and our study report that the effect becomes negative for lower ability children. By contrast, GGR find that the effect of grade repetition is decreasing in ability.

We also contribute methodologically to the FADM-approach. First, 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 delay by the end of secondary education (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. Second, by explicitly modeling tracking and the imposed downgrading, we can contrast the relative efficacy of retention and (forced) downgrading. Third, in contrast to the existing literature, we allow for multiple retention and for the fact that pupils may already have been retained prior to the start of the period of analysis, i.e. in kindergarten or primary school. To take into account that early retention may also affect the outcomes in high school (i.e. an initial conditions problem), we follow Wooldridge (2005) and condition the unobserved heterogeneity distribution on the number of years of schooling delay at the start of high school. Fourth, in contrast to FNT, who consider non-parametric identification of the FADM in the presence of *continuous* valued outcomes, we demonstrate under which conditions the model is identified in case of ordered *discrete* outcomes (see Section 4.3). 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 two tracks. We follow the approach of Mroz et al. (2016) by considering the marginal likelihood, i.e. by summing the likelihood function over the possible tracks that could have been chosen in grade 7.

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. When comparing retention to forced track downgrading as an alternative remedial measure, we observe that this alternative improves relative to retention only in that it does not lead to as much schooling delay by the end of high school.

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. Section 6 concludes. An Online Appendix, which can be downloaded from http://users.ugent.be/~bcockx/IA_BCP.pdf, contains the proof of Proposition 1, details on the chosen empirical specification, and the full estimation results.

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 September 1 of the year in which the child turns 6 years old and ends on 30 June of the year in which (s)he reaches the age of 18.[8] 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.[9] 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 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

[8] 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.

[9] 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.

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 tertiary 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 tertiary education; (iii) the vocational track (VHS) teaches practical skills that prepare for particular professions; from age 15 or 16, students in the THS and VHS tracks 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 censor inflows into AHS and into 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 tertiary 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.[10] 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^{-}.[11] 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 grade 12 and 13 in VHS.

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. However, if they wish, they 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.

[10] According to the rules, it is 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.

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

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.[12] 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 tertiary education 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 pupils whose grandmother on mother's side had a foreign nationality (584 pupils deleted), pupils who needed 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 mentioned above, we also dropped students entering the arts track (103 students), 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 end-of-year evaluation and grade mobility (396 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 seventh grade and only 46 students made track transitions involving more than two steps, we deleted their records from our sample. 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.

As the survey was conducted at age 23, there is a concern that the detailed schooling history (track choice, end of year evaluations, and timing of school drop-out or graduation) is affected by recall bias (Sudman et al., 1997; Beckett et al., 2001; von Fintel and Posel, 2016). In order to minimize this bias, the survey contained a detailed calendar with a line for each school year which the respondents were asked to complete line by line in a face-to-face interview. Interviewers were specifically instructed to pay attention to the consistency in these responses as they were regarded key information in the survey. The systematic

[12]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 age 29, but since these surveys did not add any information on the high school career, they were neglected.

listing of each school year limited the risk of associating an incorrect timing to past events ("telescoping"). Moreover, given that these school events are salient, happen regularly (each schooling year) and refer to a relatively recent phase in the life of the respondents, mistakes are expected to be limited, certainly compared to, e.g., the recall of the irregular timing of transitions between labor market states (Gillian, 2002). The fact that we had to eliminate only about 6.7% ($= 396/5,915$) of the sample because of inconsistent or missing information regarding this schooling history is coherent with the hypothesis of a relatively small recall bias. The 1978 birth cohort reports 4.6 percentage points more inconsistent histories than the 1980 one. However, we believe that this is no reason for concern, because these inconsistencies are in no way related to any of the other predetermined individual characteristics.[13] These birth cohorts were interviewed two years apart and by different interviewers. Presumably, the quality of the interviewers and/or the instructions that they received has improved over time.

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 relative and absolute frequencies. 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 helps in estimating the model including grade-varying unobserved heterogeneity (Section 4.3), i.e. random effects common to students in the same grade. 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

[13] To check this we constructed an indicator variable equal to one if the individual was one of the 396 deleted from the sample because of an inconsistent schooling history and zero otherwise. We then estimated both a linear probability model with standard errors robust to heteroskedasticity, and a probit model of this indicator on the set of predetermined covariates reported in Table 3. Apart from the one associated to the birth cohort, none of the regression coefficients were found to be individually statistically significant at the 5% level. Furthermore, the p-value of joint significance of the latter coefficients was respectively 0.399 and 0.530. These results can be obtained from the authors upon request.

Table 1: 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		
7	3,933	15.74
8	4,094	16.39
9	4,121	16.49
10	4,178	16.72
11	4,288	17.16
12	4,003	16.02
13	368	1.47
Total	24,985	100.00

repeated a grade realize far much better evaluations than those who did repeat at least one grade. Each year around 4.4% of the pupils repeat the grade, either because they choose to do so or because they are forced to as they obtained a C.[14] Over the high school career this leads to 874 out of the 3,933 pupils (22.2%) being retained at least once (see Table 2 below). On average 91.7% of the students graduate with a high school diploma. However, among those students who repeat a grade at least once, 14.9% 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 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 drop-out early). At the beginning of high school, 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. As mentioned above, 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

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

Table 2: Summary Statistics of Outcomes: 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
Grade repetition	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 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.

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 the 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 strongly correlated with grade repetition. For instance, at the start of high school the fraction in VHS is roughly equally distributed over the two subsamples, while at the end of high school the fraction in VHS is 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	

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 Econometric Model

4.1 Model Setup and Assumptions

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 ($t = 1, \ldots, T_i$;

with T_i the number of years i is observed in high school):

- $tr_{i1} \equiv Y_{i01}$: track choice at the beginning of high school. Since tracks $\{\text{VHS}, \text{THS}^-, \text{THS}^+, \text{GHS}^-, \text{GHS}^+\}$ are hierarchically ordered, tr_{i1} is an ordered response taking on the following corresponding increasing values: $\{1, 2, 3, 4, 5\}$.
- $ev_{it} \equiv Y_{i1t}$: evaluation at the end of each academic year t. This is an ordered response taking on the following increasing values: $\{\text{C}, \text{B}, \text{A}\}$ if pupils are not in the last grade and if they are not in the lowest track, i.e. not in 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. If C is the outcome, then the student is automatically retained in the subsequent academic year $t + 1$ ($re_{it} = 1$). If the evaluation is A, the student is automatically promoted to the next grade in year $t + 1$ ($re_{it} = 0$).
- $out_{it} \equiv Y_{i2t}$: decision to drop-out of school at the *end* of year t from age 18 onwards (i.e. the end of compulsory schooling age). out_{it} is equal to 1 if student i decides to drop out of school at the end of year t, 0 otherwise. If $out_{it} = 1$, no school outcomes are observed for this individual in year $t + 1$.
- $re_{it} \equiv Y_{i3t}$: indicator equal to 1 if student i decides at the end of year t to repeat the grade in year $t + 1$, 0 otherwise. A student can make such a choice only if the end of year evaluation is B. The notation with respect to re_{it} does not distinguish between the *voluntary* decision to repeat the grade after a B evaluation and the aforementioned automatic *involuntary* grade retention after a C evaluation. In both cases $re_{it} = 1$. This should be kept in mind when interpreting the model assumptions and the empirical findings below.
- $dow_{it} \equiv Y_{i4t}$: choice at the end of year t 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 THS^-, they cannot make a 'two-step downgrade' so that their choice is dichotomous, either 'no downgrade' or 'one-step downgrade'. When one decides not to repeat the grade after a B evaluation, 'no downgrade' is not possible.

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. Furthermore, since pupils can have to repeat one grade, this generates multiple observations within individuals and within the same grade g, with $g = 7, \ldots, 13$. Only

the track choice at the beginning of high school (grade 7), tr_{i1}, is realized just once. Formally, the grade g is related to t by $g = 6 + t$, except if the individual has been retained in the past (i.e. $re_{is} = 1$ for some $s < t$) in which case $g = 6 + t - \mathbf{1}_{\{\forall t: t>1\}}(t) \sum_{s=1}^{t-1} re_{is}$, where $\mathbf{1}_A(x)$ defines the indicator function, equal to one if $x \in A$ and zero otherwise.

In what follows, we discuss three assumptions that are crucial to the identification of this model, as will be discussed in Section 4.3.

Assumption 1 *(Sequential ordering of outcomes)*

Within each academic year t and for $t = 1, \ldots, T_i$, the four time-varying outcome variables are realized sequentially with the following chronological order: evaluation at the end of the academic year, ev_{it}; school drop-out choice, out_{it}; choice to repeat the grade, re_{it}; and track downgrade choice, dow_{it}.

Assumption 1 implies that the model is recursive. The distributions of each outcome c in any time period t, conditional on the predetermined variables and lagged outcomes, can be seen as structural in the sense that they are invariant to any changes in endogenous and exogenous variables (Heckman and Vytlacil, 2007, Section 4). They can, hence, be used to determine causal relationships of past decisions, such as retention in a particular grade, on future outcomes, such as evaluations in subsequent grades.

The assumption that the decision to drop out of school in year $t+1$ (out_{it}) is made after the end-of-year evaluation is natural. Also, the decision to repeat the grade (re_{it}) in case that $ev_{it} = B$ or to downgrade (dow_{it}) in year $t + 1$ is naturally made after the evaluation and the decision not to drop out of school at the end of year t. However, one might question whether in practice the choice to repeat the grade in case of a B precedes the downgrading decision.[15] It is more natural to assume that the choice of grade repetition comes first, because the decision after a B evaluation to promote to next grade just restricts the choice set of the downgrading decision: one has to downgrade in this case. 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, (s)he has implicitly simultaneously decided that (s)he will repeat the grade: no choice is left.[16]

[15] One could argue that these decisions are taken jointly and that each possible combination of these outcomes should be modeled within a multinomial choice model. We did not follow this approach, because it is not clear that such a model is identified with the available data. Even if the assumption of sequential ordering of these choices is not correct, this is expected to have a negligible impact on our findings, because this assumption is only imposed in case the end-of-year evaluation is a B. This concerns less than 5% of the sample each year.

[16] We checked which order in the timing provides a best fit of the data by estimating also the model under the alternative ordering assumption and using a Vuong (1989) test to determine whether one of these two non-nested models could be statistically rejected

Assumption 1 implies *no perfect anticipation* (Abbring and van den Berg, 2003): the sequential ordering of the outcomes means that outcomes that realize in the future, e.g. obtaining a C at the end of year $t+1$ (i.e. to fail in year $t+1$ and to be retained in year $t+2$), cannot have an influence on schooling outcomes in the current year, e.g. the evaluation, the decision to downgrade or to drop-out at the end of year t. Note that this assumption does not rule out that pupils predict in year t that they have *more chance* to fail in year $t+1$, but it excludes that they are in year t *sure* that this will happen in year $t+1$ and that they act on this information in year t. The no anticipation assumption would be violated if for example a student who fails knows for sure that she will (not) fail in the next year. This is very unlikely.

The educational outcomes are modeled as a sequence of ordered choices each of which is determined by a latent linear index lying within an interval of a set of thresholds. This will be formally stated in Assumption 2. Before doing so, we explain how we deal with endogenous outcomes that are realized before the start of high school and introduce some additional notation.

As mentioned in Section 3, pupils start high school at different ages due to different retention histories either in primary school or in kindergarten, or because they may have skipped a grade. As a consequence, age at the start of high school can take on three values: $in_i = 0$ if the pupil is 12 years old at the start of high school (i.e. high school is started "on time"); $in_i = 1$ if the pupil is 13 years old (i.e. high school is started with one year of delay); $in_i = -1$ if the pupil is 11 years old (i.e. high school is started early). Even if we assume that retention in high school is a new, different, process with no history, we cannot ignore the dependence on earlier retention through in_i, because in_i and retention in high school are related through common unobserved determinants of high school outcomes. We solve this initial conditions problem in a similar way as Wooldridge (2005) in dynamic nonlinear panel data models with unobserved heterogeneity. This consists in integrating out the unobserved heterogeneity conditional on in_i and the observed exogenous explanatory variables. To implement this solution we must make some parametric assumption on how the unobserved heterogeneity distribution depends on in_i and on the observed exogenous explanatory variables. We follow Wooldridge (2005) by assuming that this dependence can be completely captured by adding in_i as a conditioning variable in the conditional mean of latent variables associated to the ordered high school outcomes in Assumption 2 below and that, conditional on in_i, the unobserved heterogeneity is independent

against the other, even if this should be rather considered as a test of functional form than a test of which order is correct. 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 3.798 and rejects the alternative hypothesis at a p-value of 0.0001. The complete estimation results and calculation of the test statistic can be found in the Online Appendix E.

of the observed exogenous explanatory variables.[17] The independence of the unobserved heterogeneity conditional on in_i will be formally stated in Assumption 3.3 below. This is a strong assumption. However, in view of the small number of pupils not starting high school on time (3.7% of the students start high school late, at age 13, and 1.0% start high school early, at age 11), ignoring this initial condition problem or specifying it differently would hardly affect the results. We checked that the parameter estimates are indeed hardly affected if we ignore in the estimation the 186 individuals who do not start high school on time. Estimation results can be obtained from the authors upon request.

Further, let $\mathbf{Y}_{it} \equiv [ev_{it}\ out_{it}\ re_{it}\ dow_{it}] \equiv [Y_{i1t}\ Y_{i2t}\ Y_{i3t}\ Y_{i4t}]$ be the row vector collecting the four time-varying outcome variables,[18] $\mathbf{x}_i$ a vector of predetermined observed explanatory variables, z_i a predetermined observed variable only affecting tr_{i1}. This exclusion restriction is required for identification and is further discussed in Section 4.3.

Assumption 2 *(Sequence of ordered latent variable models represented by linear indices and a set of thresholds)*

The track choice tr_{i1} and other ordered school outcomes Y_{ict} can be represented by the following latent variables:

$$tr^*_{i1} \equiv Y^*_{i01} = \mu(\mathbf{x}_i, z_i, in_i; \theta_{01}) + u_{i01} \quad (1)$$

$$Y^*_{ict} = \mu(\mathbf{x}_i, in_i, \Im^*_{ic-1,t}; \theta_{ct}) + u_{ict}, \quad (2)$$

with $c \in \{1, \ldots, 4\}$, $t \in \{1, ..., T_i\}$ and $\mu(\ldots; \theta_{ct})$ is linear in the vector of parameters θ_{ct} associated to the conditioning variables. $\Im^*_{ic-1,t}$ denotes the history of the schooling outcomes prior to the current discrete outcome Y_{ict} including the contemporaneous outcomes that precede Y_{ict} in time period t. So, $\Im^*_{ic-1,t} \equiv (Y_{ic-1t},\ ..., Y_{i1t}, \Im_{it-1})$. In the latter expression, $\Im_{it-1}$ denotes all the realizations of the endogenous variables from $t-1$ until the start of the processes, i.e. $\Im_{it-1} \equiv (\mathbf{Y}_{it-1}, \ldots, \mathbf{Y}_{i1},\ tr_{i1}, in_i)$. u_{ict} are

[17] Wooldridge (2005) does not require this independence assumption, because he follows the approach of Chamberlain (1980) to separately identify the dependence of explanatory variables on unobserved heterogeneity and their causal impact on the outcome. This is not possible in our study because it requires time variation in the explanatory variables, which there is not.

[18] In principle we should distinguish for $t > 1$ in the notation between observed and counterfactual outcomes $\mathbf{Y}_{it}(re_{i1}, ..., re_{it-1}) \equiv \mathbf{Y}_{it}(re_{is})_{s=1}^{t-1}$, where $(re_{is})_{s=1}^{t-1}$ denotes the counterfactual history of retention, one of which is only observed. In what follows we do not explicitly make this distinction for notational convenience. The reader should be able to deduce from the context whenever we refer to a counterfactual outcome, in which case the outcomes are implicitly conditioned on the counterfactual history $(re_{is})_{s=1}^{t-1}$. Treatment effects, e.g., should be defined in terms of the counterfactual outcomes. A particular treatment effect of interest is that of some history of retention relative to no retention in any grade of high school, i.e. $Y_{ict}(re_{is})_{s=1}^{t-1} - Y_{ict}(0, ..., 0)$. We refer to FNT for a discussion of the simpler case in which an individual can be retained at most once.

unobserved individual specific components with $E(u_{ict}) = 0$. Choices c are generated by the index Y^*_{ict} falling in various intervals determined by thresholds $\alpha_{1,c} < ... < \alpha_{J_c,c}$ where $J_c + 1$ is the number of possible choices of the outcome Y_{ict}.

Finally, in Assumption 3 we characterize in three steps the assumptions that we make in regard to the unobserved determinants of the considered outcome variables. We first formally state these assumptions and then discuss them.

Assumption 3

3.1 *One factor specification of unobservables*

$\forall c \in \{0, ..., 4\}, g \in \{7, \dots 13\} : u_{ict} = v_{ic}(g) + \epsilon_{ict}$. More specifically, if $t = 1$: $u_{ic1} = \delta_{c7} v_{i1}(7) + \epsilon_{ic1}$ and $\forall t > 1, c \neq 0 : u_{ict} = \delta_{cg}(re_{is})_{s=1}^{t-1} v_{i1}(g) + \epsilon_{ict}$, where $\delta_{cg}(re_{is})_{s=1}^{t-1} = \delta_{cg}(re_{i1}, \dots, re_{it-1})$.[19] In this specification, $v_{i1}(g)$ are unobserved grade specific "factors" common to all outcomes c. Further, δ_{c7} and $\delta_{cg}(re_{is})_{s=1}^{t-1}$ are unknown "factor loadings" associated to each outcome. Clearly, for $t > 1$, these factor loadings can depend on the history of past retention $(re_{is})_{s=1}^{t-1}$. Factor loadings δ_{17} and $\delta_{1g}(0, ..., 0)$ are normalized to 1. Finally, ϵ_{ict} are i.i.d. error terms. It is assumed that $v_{ic}(g) \perp\!\!\!\perp \epsilon_{ict}$ and that $\forall ct \neq (ct)' : \epsilon_{ict} \perp\!\!\!\perp \epsilon_{i(ct)'}$.

3.2 *Random independent grade-specific unobserved persistent shocks* $v^*_{i1}(g)$

$v_{i1}(7) \equiv v^*_{i1}(7)$ and $\forall g \in \{8, \dots, 13\} : v_{i1}(g) = \sum_{j=7}^{g-1} \delta^*_j(g) v^*_{i1}(j) + v^*_{i1}(g)$. In this specification, $\delta^*_j(g) \in \mathbb{R}$ is an unknown parameter that depends on j and g. Further, $v^*_{i1}(g)$ are mutually independent random persistent shocks: $\forall g \neq g' \in \{7, \dots, 13\} : v^*_{i1}(g) \perp\!\!\!\perp v^*_{i1}(g')$.

3.3 *Independence between unobservables and observables*

$\forall c, t, g : (v_{i1}(g)\ \epsilon_{i1t}) \perp\!\!\!\perp \left(\mathbf{x}_i\ in_i\ \Im^*_{ic-1,t}\right)$.

In Assumption 3.1 we restrict the unobservables *within* grade g between choices c to depend perfectly on each other: all random unobserved determinants of outcome c in grade g, $v_{ic}(g)$, are proportional to the unobserved determinants of outcome 1, $v_{i1}(g)$. This means that we impose a one-factor specification on the

[19] This specification allows the loading factors δ_{cg} to depend non-parametrically on the complete history of retention $(re_{is})_{s=1}^{t-1}$, not distinguishing between voluntary retention (after a B) or involuntary retention (after a C). To explain the notation, consider $\delta_{48}(0\ 1)$ for the case that $t = 3$. $\delta_{48}(0\ 1)$ is the loading factor for choice $c = 4$ (i.e. for the decision to downgrade dow_{i3} in year $t = 3$) of the unobserved heterogeneity $v_{i1}(8)$ in grade 8 of an individual i who passed grade 7 without retention ($re_{i1} = 0$), but who was retained in the first year of grade 8 ($re_{i2} = 1$), so that (s)he repeats grade 8 in year $t = 3$. While we show in Proposition 1 that the model is identified for this general loading factor specification, we will impose in the empirical application parametric restrictions as described in Section 4.4.

unobserved determinants within a grade. The aim of imposing such a low-dimensional set of "common" causes is to accommodate both *essential heterogeneity* in the treatment effect (Heckman et al., 2006)[20] and grade-varying unobserved shocks. This matters because low ability students are more likely to be retained and to learn at a slower pace than high ability students, and because there are unobserved shocks that affect both selection into grade repetition and the related treatment effect. FNT show that in a similar factor specification and similar assumptions the model is non-parametrically identified. We follow their argument to prove identification of our model in Section 4.3.

In Assumption 3.2, we follow FNT by assuming that in each grade a new independent, but persistent shock $v_{i1}^*(g)$ arrives.[21] Consequently, in each grade the unobserved determinant $v_{i1}(g)$ of the retention and other schooling outcomes consists in the sum of the current ($v_{i1}^*(g)$) and weighted past shocks $\left(\sum_{j=7}^{g-1} \delta_j^*(g) v_{i1}^*(j)\right)$, with $\delta_j^*(g)$ as weights (not adding up to one). The persistent dependence on past shocks induces dependence of the grade-specific shocks $v_{i1}(g)$ across grades. The reader might notice that this assumption is equivalent to allowing unrestricted dependence of the $v_{i1}(g)$ across grades g. In the estimation we exploit this equivalence and directly estimate the joint distribution of the $v_{i1}(g)$ for $g \in \{7, \ldots, 13\}$ rather than the distributions of the $v_{i1}^*(g)$, because this is more convenient in the estimation procedure. This also implies that we do not estimate the $\delta_j^*(g)$ (Subsection 4.2). We define the independent persistent shocks $v_{i1}^*(g)$, because this facilitates the identification proof (see Proposition 1 below).

As mentioned earlier, Assumption 3.3 implies that the distribution of unobserved heterogeneity $v_{i1}(g)$ is independent of in_i, the "initial condition". Hence, we no longer need to condition the unobserved heterogeneity distribution on in_i when we integrate it out of the likelihood function (Section 4.2). Furthermore, as is shown in the proof of Proposition 1 in the Online Appendix, key in Assumption 3.3 is that the unobservables do not depend on track choice. This means that there cannot be any higher order dependence on track choice that is not captured by the conditional mean, and excludes that the factor loading δ_{11} on the unobserved determinant $v_{i1}(7)$ depends on tr_i, i.e. there cannot be *essential heterogeneity* in the effect of track choice.

[20] Formally, Assumption 3.1 makes the factor loading δ_{cg}, which is associated to the unobserved determinants $v_{ic}(g)$ of the schooling outcome c in grade g, dependent on the past history of grade repetition $(re_{is})_{s=1}^{t-1}$. This generates individual specific unobservable gains of the treatment, i.e. of a particular history of retention, relative to the counterfactual of no retention: $\left[\delta_{cg}(re_{is})_{s=1}^{t-1} - \delta_{ct}(0, ...0)\right] \cdot v_{i1g}$. FNT consider the special case in which pupils can be retained at most once in a grade. Then, the history of retention is fully captured by the moment that it has taken place. We generalize by allowing for multiple grade repetitions in our analysis.

[21] In fact FNT allow for new shocks in each time period t. Our data only allow non-parametric identification of new shocks arriving in each grade (Subsection 4.3).

4.2 Derivation of the Likelihood Function

Let $\mathbf{V}_i$ be a random 5×7 matrix capturing for each of the 5 choices $c \in \{0, \ldots, 4\}$ the (possibly grade-varying) unobserved determinants of our outcome variables: $\mathbf{V}_i' \equiv \left[\mathbf{v}_{i,tr}' \; \mathbf{v}_{i,ev}' \; \mathbf{v}_{i,out}' \;\; \mathbf{v}_{i,re}' \; \mathbf{v}_{i,dow}'\right] \equiv [\mathbf{v}_{ic}]_{c=0}^{4}$, where $\forall c \neq 0 : \mathbf{v}_{ic}' \equiv [v_{ic}(g)]_{g=7}^{13}$ and $\mathbf{v}_{i0} \equiv v_{i0}(7)$. Based on the chain rule and Assumptions 1 and 3.1 (but ignoring in the notation the dependence of δ_{cg} on the history of retention $(re_{is})_{s=1}^{t-1}$), we can write the joint distribution $D(tr_{i1}, \mathbf{Y}_i | \mathbf{x}_i, z_i, \mathbf{V}_i, in_i)$ as:

$$
\begin{aligned}
D(tr_{i1}, \mathbf{Y}_i | \mathbf{x}_i, z_i, \mathbf{V}_i, in_i) \;&=\; D_{01}(tr_{i1} | \mathbf{x}_i, z_i, \mathbf{V}_i, in_i) \cdot \prod_{t=1}^{T_i} D_t(\mathbf{Y}_{it} | \mathbf{x}_i, \mathbf{V}_i, \Im_{it-1}), \\
&=\; D_{01}(tr_{i1} | \mathbf{x}_i, z_i, \delta_{07} v_{i1}(7), in_i) \\
&\cdot\; \prod_{t=1}^{T_i} \Big[D_{1t}(ev_{it} | \mathbf{x}_i, \delta_{1g} v_{i1}(g), \Im_{it-1}) \\
&\cdot\; D_{2t}(out_{it} | \mathbf{x}_i, \delta_{2g} v_{i1}(g), \Im_{it-1}, ev_{it})^{s_{it}} \\
&\cdot\; D_{3t}(re_{it} | \mathbf{x}_i, \delta_{3g} v_{i1}(g), \Im_{it-1}, ev_{it} = B, out_{it} = 0)^{1 - lg_{it}} \\
&\cdot\; D_{4t}(dow_{it} | \mathbf{x}_i, \delta_{4g} v_{i1}(g), \Im_{it-1}, ev_{it}, out_{it} = 0, re_{it})^{vt_{it}} \Big], \qquad (3)
\end{aligned}
$$

where s_{it} is an indicator equal to 1 if student i is 18 or older (and, thereby, can legally drop-out), lg_{it} is an indicator equal to 1 if the student is attending the last grade of high school (and, thereby, must repeat the grade after a B evaluation when not dropping out) and vt_{it} is equal to 1 if the student is in the GHS or THS track and 0 if (s)he is in the VHS track (and, thereby, cannot downgrade). Recall that in Equation (3) the grade g is related to t: $g = 6 + t$, except if the individual has been retained in the past (i.e. $re_{is} = 1$ for some $s < t$), in which case $g = 6 + t - \mathbf{1}_{\{\forall t : t > 1\}}(t) \sum_{s=1}^{t-1} re_{is}$.

If we specify the parametric forms $F(.|.; \theta_{ct})$ for the conditional probability distribution functions of the outcome variables $D_{ct}(.|.)$ and for $\delta_{cg}(re_{is})_{s=1}^{t-1}$ (as we do in Section 4.4 and Online Appendix C), the joint distribution function in Equation (3) defines the likelihood contribution for individual i conditional on the unobserved heterogeneity $\mathbf{v}_{i1}$. Since by Assumption 3.3 this unobserved heterogeneity is independent of the observed conditioning variables, including the initial condition in_i, we can integrate out this unobserved heterogeneity provided that we specify its distribution function. If we denote this distribution by $G(\mathbf{v}_{i1}; \boldsymbol{\rho})$, where $\boldsymbol{\rho}$ is a vector of unknown parameters, then we obtain the unconditional log-likelihood function by taking the logarithm of this marginal joint probability distribution function for each individual i and summing

over all the individuals in the retained sample:[22]

$$\begin{aligned}\ell(\boldsymbol{\theta},\boldsymbol{\rho}) \;=\; & \sum_{i=1}^{N} \ln \Bigg\{ \int_{\mathbb{R}^{T_i}} F(tr_{i1}|\mathbf{x}_i, z_i, v_{i0}(7), in_i; \boldsymbol{\theta}_{tr,1}) \cdot \prod_{t=1}^{T} \Big[F(ev_{it}|\mathbf{x}_i, v_{i1}(g), \Im_{it-1}; \boldsymbol{\theta}_{ev,t}) \\ & \cdot \; F(out_{it}|\mathbf{x}_i, v_{i2}(g), \Im_{it-1}, ev_{it}; \boldsymbol{\theta}_{out,t})^{s_{it}} \\ & \cdot \; F(re_{it}|\mathbf{x}_i, v_{i3}(g), \Im_{it-1}, ev_{it}=B, out_{it}=0; \boldsymbol{\theta}_{re,t})^{1-lg_{it}} \\ & \cdot \; F(dow_{it}|\mathbf{x}_i, v_{i4}(g), \Im_{it-1}, ev_{it}, out_{it}=0, re_{it}; \boldsymbol{\theta}_{dow,t})^{c_{it}} \Big] \mathrm{d}G(\mathbf{v}_{i1};\boldsymbol{\rho}) \Bigg\} \\ \equiv \; & \sum_{i=1}^{N} \ln \left[\int_{\mathbb{R}^{T_i}} \mathcal{L}_i(\boldsymbol{\theta},\boldsymbol{\rho}) \mathrm{d}G(\mathbf{v}_{i1};\boldsymbol{\rho}) \right], \end{aligned} \tag{4}$$

where $\mathcal{L}_i(\boldsymbol{\theta},\boldsymbol{\rho})$ is the individual contribution to the likelihood and $\boldsymbol{\theta}$ is the vector of parameters fully characterizing the probability distribution functions conditional on $\mathbf{v}_{i1}$.

4.3 Identification

We use a very similar identification argument as the one FNT present for the single-dimensional ability example in their Section IV.B. This approach differs from the one in Carneiro et al. (2003) in that it exploits higher-order moments. As such, it requires less "measurements", i.e. outcomes which do not depend on the treatment status.[23] Key in the identification argument is that the schooling outcomes in the first period are "free of selection", i.e. one cannot have been retained prior to the first period, neither can the outcomes in the first period be affected by a subsequent retention.[24] Under this assumption and the assumption that there is only one unobserved factor that affects both selection into retention and the schooling outcome (Assumption 3.1), one can identify the distribution of this unobserved factor from the cross moments between this first schooling outcome and the first selection into retention. Once this unobserved component is identified, one can condition on it to control for the selection of low ability students into retention and identify the effect of retention in the previous period on the schooling outcome in the next period. In each subsequent time period the same argument can be followed, conditional on the previously unobserved components. Using the within next period's cross moments between at least two outcomes, one can then identify within each subsequent time period a new unobserved persistent component that is independently distributed from the

[22] For notational convenience, we do not substitute out $v_{ic}(g)$ by $\delta_{cg} v_{i1}(g)$ in Equation (4).

[23] In the framework of Carneiro et al. (2003) (using only second order moments) identification of a single-dimensional ability, as in our case, would require at least three measurements. In the FNT approach two measurements (or one measurement and one selection equation free of selection) are sufficient.

[24] This is the "no anticipation" assumption discussed in Section 4.1, which follows from Assumption 1 on sequential outcomes.

previous ones and, hence, time-varying unobserved heterogeneity components. FNT (p. 997) argue that this approach is similar to "matching", where the conditional independence assumption holds after conditioning not only on observable, but also unobservable determinants.

Because our model differs from that of FNT in a number of respects, we must adapt the identification argument. A first point is that the schooling outcomes in the first period are not "free of selection". First, there is an initial conditions problem because not all pupils start high school at the same age. As explained in Section 4.1, this is solved as in Wooldridge (2005) by conditioning all schooling outcomes on this age. Second, the first observed schooling outcome tr_{i1} induces selection on the second outcome ev_{i1}, because individuals are not randomly selected into tracks.[25] Proposition 1 below states two sufficient conditions (a continuous exclusion restriction or the presence of two continuous explanatory variables) based on which this selective dependence is identified. Because all schooling outcomes in our data are discrete and not continuous, the sufficient conditions require *continuous* variables. If the schooling outcomes would have been continuous, e.g. test scores, such as in FNT, then this requirement could have been relaxed. Because of the continuous nature of the schooling outcomes in their model, FNT can prove identification under weaker conditions: they require neither exclusion restrictions nor continuous explanatory variables.

A second difference with FNT is that we cannot allow for unrestricted time variation in the unobserved heterogeneity. By the discrete nature of our outcome variables, this would require a continuous exclusion restriction or at least two continuous explanatory variables determining two outcome equations *within* each time period. Our data just contains an exclusion restriction for tr_{i1} the track choice at the start of high school and, hence, not in each time period (see below). However, by imposing that unobserved random shocks $v_{ic}(g)$ are constant within each grade, we can identify the distributions of these shocks on the basis of the cross moments of the error terms between two time periods within the same grade. This is because the idiosyncratic error terms ϵ_{ict} are unrelated to these cross moments, as they are independently distributed over time (Assumption 2.3). Hence, identification of the unobserved shocks $v_{ic}(g)$ is based on the individuals retained in grade g.

A final set of differences is that we consider more schooling outcomes and allow for multiple retention and an initial conditions problem. Considering more outcomes actually facilitates identification, because it allows to form more cross moments from which the unobserved factors can be identified. We do not exploit these. This means that our model is over-identified. Allowing for multiple retention makes the identification

[25] In the case of FNT, the first selection (into retention) occurs only in the second period. This means that the outcome and selection into retention in the first period are both "free of selection".

argument more tedious, as it multiplies the number of possible treatments (any combination of different timing and occurrence). It does not however complicate the identification argument in any substantial way.

The following proposition states two sufficient conditions for non-parametric identification of the general econometric model that we specified in the previous subsection.[26]

Proposition 1 *(Non-parametric identification)*

If (i) Assumptions 1-3 are satisfied, (ii) $\mathbf{x}_i$ *has full rank, (iii) the distribution functions of* $u_{ict} \equiv \delta_{cg}(re_{is})_{s=1}^{t-1} v_{i1}(g) + \epsilon_{ict}$ *($\forall c, t, g$) are absolutely continuous and strictly increasing over their supports, where* $v_{i1}(7)$ *and* ϵ_{ic1} *have finite moments (at least up to the fourth order), and, (iv)* $\forall ct : \delta_{cg}(re_{is})_{s=1}^{t-1} \neq 0$,[27] *then one of the two following conditions is sufficient for non-parametric identification of the model that we defined in Section 4.1:*

1. *There is at least one explanatory variable* z_i *that varies continuously over the set of real numbers* $\mathbb{R}$ *and that affects* $tr_{i1} \equiv Y_{i01}$*, but not any other outcome* $Y_{ict} \neq Y_{i01}$ *conditional on* tr_{i1} *(*exclusion restriction*), or*
2. *there is no exclusion restriction (*$\{z_i\} = \emptyset$*), but there are at least two components of* $\mathbf{x}_i$ *that vary continuously over the set of real numbers* $\mathbb{R}$*.*

Proof. See Online Appendix A. ■

Our model does not satisfy the second sufficient condition of Proposition 1. As a consequence, we rely on the aforementioned first condition. More concretely, the results that we report below are based on a model which includes day of birth only in the track choice equation. Calendar day of birth is a multi-valued explanatory variable that strongly determines track choice at the start of high school tr_{i1} and it is assumed that this variable can be excluded as determinant from any of the subsequent schooling outcomes. Two concerns with respect to this exclusion restriction must be discussed here. First, the literature on relative age effects shows that relative age affects socio-economic outcomes beyond the first schooling outcomes (Bedard and Dhuey, 2006; Sprietsma, 2010; Alet et al., 2013). However, Bedard and Dhuey (2006) and Sprietsma (2010) find that most of the relative age effect comes from its impact on the initial selection of pupils into different grades, from intra-class ability grouping and from initial grade repetition, rather than from a long-run *direct* impact. Alet et al. (2013) exploit this to assume that quarter of birth only has a direct impact on

[26] In contrast to the proof of FNT for the single dimensional ability case, we do not require that the distributions of $v_{i1}^*(g)$ are asymmetric. We indeed exploit fourth order rather than third order cross moments of the error terms of the schooling outcomes.

[27] This means that for each t all the v_{ict} are perfectly correlated over outcomes c: independence is not permitted.

both the first grade test score and the first/second grade repetition, but not on subsequent test scores. We make a similar assumption justified on the grounds that within the hierarchical school system in Belgium the track choice at the start of high school coincides very much with an ability grouping and that conditional on this grouping the relative age only plays a minor role in subsequent high school outcomes. The fact that day of birth does not significantly affect any of these outcomes and hardly has any impact on our findings when we do include it as explanatory variable in these subsequent outcomes is comforting.[28] Second, calendar day of birth does not vary over the full support of $\mathbb{R}$. This implies that there can be at most identification over a segment of the support of $\mathbf{v}_{i1}$. Moreover, since this variable is not continuous, identification is only ensured at discrete points of this support. Nevertheless, given that the exclusion restriction is multi-valued and, hence, close to continuous and since the inclusion of more than two schooling outcomes per time period provides further identifying restrictions, we believe that this cannot be a major problem in practice.

4.4 Model Restrictions

In order to facilitate estimation and increase precision, we impose further restrictions on our model:

1. We assume that the parameters θ_{ct} of the linear index $\mu(.;\theta_{ct})$ are separable in c and g and are therefore constant over time t within grade g: $\theta_{ct} = [\theta_c \ \theta_g]$. In addition, in the empirical analysis we consider two models. The first is a linear index in the conditioning variables, while the second adds, to allow for treatment heterogeneity, an interaction term of this linear index with the indicators of past grade repetition.

2. The dependence on the past history of outcome variables is restricted in the following way: (i) the preceding outcomes within the same period are generally included; (ii) the downgrading decision only in the previous year is conditioned upon (first order Markov property); (iii) the grade repetition of the previous year re_{it-1} is conditioned upon, and the history of grade repetition prior to year $t-1$ is summarized by zero-one indicator pre_{it-1}, equal to one if the student has ever repeated a grade in high school in years prior to year $(t-1)$; (iv) the average effect of the grade is captured by an indicator

[28] While the associated coefficient is highly significant in the track choice equation, it is not statistically significant in any other equation, except for the drop-out equation. The p-value of the joint test of significance of day of birth in all equations except for in the drop-out equation is 0.753. It becomes 0.034 if the drop-out equation is included. The significance of day of birth in the drop-out equation is unrelated to a direct relative age effect. It is rather the consequence of the rule that compulsory education ends on June 30 of the year in which students turn 18. Indeed, if we add in the drop-out equation an indicator equal to 1 if the individual was born after June 30 and 0 otherwise, the significance of the day of birth in the drop-out equation disappears. These estimation results are not reported for the sake of brevity, but available from the authors upon request.

of the grade occupied in the beginning of year t, which is determined by time t and the number of past grade repetitions; (v) an indicator of track choice at the beginning of year t is included; this depends on the initial track choice and the history of downgrading.

3. We assume a discrete distribution for the one-factor heterogeneity distribution $G(\mathbf{v}_{i1}; \boldsymbol{\rho})$, with unknown support points and probability masses. The number of support points is chosen by minimizing the Akaike information criterion (AIC).

4. The loading factors are assumed to be constant across grades: $\delta_{cg}(.) = \delta_c(.)$, and their dependence on the past grade repetition restricted to $(re_{i,t-1}, pre_{i,t-1})$; moreover, the parameters of the interaction of $v_{i1}(g)$ with $(re_{i,t-1}, pre_{i,t-1})$ (determining essential heterogeneity) are restricted to be the same as those of the interactions of the linear index of the observed conditioning variables with $(re_{i,t-1}, pre_{i,t-1})$ mentioned in point 1.

5. The ϵ_{ict} are assumed to have a logistic distribution, so that the $F(.|.; \theta_.)$'s in Equation (4) are (ordered) logit models.

The Online Appendix C contains a detailed description on how the different functions of the model are parametrically characterized as a result of these restrictions.

4.5 Partial Observability of Track Choice at the Start of High School

Finally, in the empirical application we do not maximize the log-likelihood exactly as expressed in Equation (4), because we face a problem of partial observability. As mentioned in Section 3, we have only partial information about the school track choice tr_{i1} at the start of high school. 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, we do observe the track in which students end up in (the first year of) grade 8, denoted by tr_{i2}. This information together with the observed outcomes in the first high school year $\mathbf{Y}_{i1}$ (e.g. that no student repeats grade 7) and the institutional setting in which students can only downgrade, and if they do, at most two tracks, conveys information about the possible starting track tr_{i1}. For example, students who are in GHS$^+$ (VHS) in grade 8 (7), surely were also in GHS$^+$ (VHS) in grade 7 (8), as track upgrading is not allowed.[29] For the same reason, students in GHS$^-$ in grade 8 could not have been in THS or VHS in grade 7. In order to accommodate for this partial observability for all pupils who were not in the vocational track at the start of high school, we form the

[29] For these groups (32.4% of the total sample, as can be seen from Table 2) there is actually no problem of partial observability.

marginal likelihood by summing the likelihood conditional on tr_{i1} over the possible initial track choices given the restrictions imposed by the institutional setting and given the observed information about tr_{i2} in grade 8. This is similar to the strategy used by Mroz et al. (2016) who solved the partial observability of the time at which persons with diabetes progress to the next disease stage. More concretely, Mroz et al. (2016) summed over the unknown time moments given the available information about the time period in which the next disease stage could have arisen. In our setting, taking the prior information into account that restricts the set of possible initial track choices is somewhat more involved. The reader can find the technical details of this procedure in Online Appendix B.

5 Empirical Results

The estimated parameters of the econometric model are reported in Online Appendix D. We present the estimates of three models: without unobserved heterogeneity, with grade-constant and with grade-varying unobserved heterogeneity. As mentioned in Subsection 4.4, 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 the one controlling for grade-varying unobserved heterogeneity. According to the AIC, the latter model is to be preferred. Hence, 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, which are therefore reported in Online Appendix D. 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 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 Online Appendix C. 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 determined segments on the unit interval that correspond to particular choices. The assigned choice depended on the segment in which the random number fell. Once a choice was assigned it was saved and conditioned upon in the subsequent choice. In each of the 999 simulations this procedure ended if all sampled individuals either had 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).[30] Average treatment effects on the treated (ATTs) are estimated by the average difference between two (counterfactual) outcomes: one in which the treatment of retention is predicted by the model – i.e. a C evaluation is realized – and one in which a counterfactual treatment of no retention – i.e., in our benchmark approach, an A evaluation – is enforced. 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 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 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

[30] This corresponds to the procedure proposed by Krinsky and Robb (1986). Woutersen and Ham (2013) criticize this approach in that the CI obtained in this way may not cover the true parameter and may result in a too narrow CI in cases that the function of the estimated parameters is non-differentiable, has zero derivatives, or unbounded derivatives. Since the functions we consider are (a difference of) a product of logits, which are all differentiable over their support, we do not face this complication.

Table 4: Goodness-of-fit

	Model: grade-varying unobserved heterogeneity			
	Actual probability		Simulated probability	95% CI
I. Outcomes in grade 8				
A	0.900		0.898	[0.883, 0.912]
Repeating the grade	0.040	*	0.050	[0.040, 0.062]
II. Outcomes in grade 9				
A	0.923		0.921	[0.908, 0.934]
Repeating the grade	0.047		0.044	[0.035, 0.055]
A, if repeated grade 8	0.828		0.857	[0.798, 0.916]
A, if did not repeat grade 8	0.927		0.925	[0.911, 0.936]
III. Outcomes in grade 10				
A	0.908		0.905	[0.891, 0.919]
Repeating the grade	0.061		0.058	[0.048, 0.069]
A, if repeated grade 9	0.863	*	0.806	[0.737, 0.869]
A, if did not repeat grade 9	0.910		0.910	[0.896, 0.924]
IV. High school graduation				
Diploma	0.915		0.913	[0.886, 0.934]
Diploma, if repeated grade 8	0.720		0.757	[0.654, 0.837]
Diploma, if did not repeat grade 8	0.923		0.921	[0.897, 0.940]
Diploma, if repeated grade 9	0.798		0.763	[0.644, 0.854]
Diploma, if did not repeat grade 9	0.921		0.920	[0.895, 0.940]
Diploma, if repeated grade 10	0.825		0.787	[0.688, 0.860]
Diploma, if not repeated grade 10	0.921		0.920	[0.897, 0.941]
V. Years schooling delay at start last compulsory year				
Diploma	0.229		0.234	[0.212, 0.257]
Delay, if repeated grade 8	1.261		1.209	[1.137, 1.292]
Delay, if did not repeat grade 8	0.186		0.182	[0.163, 0.204]
Delay, if repeated grade 9	1.202		1.242	[1.161, 1.338]
Delay, if did not repeat grade 9	0.181		0.187	[0.168, 0.206]
Delay, if repeated grade 10	1.200		1.187	[1.126, 1.254]
Delay, if did not repeat grade 10	0.166		0.175	[0.155, 0.196]

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.

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 (i) repeating the grade after (the first year in) grade 8 and (ii) getting an A evaluation in grade 10 after repeating grade 9, the actual probability is different from the simulated probability 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) 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 matter (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 Short- and Long-Term Effects of Retention in Grade 8

Model (1) of Table 5 presents the average effect of retention in grade 8 when controlling for grade-varying unobserved heterogeneity. 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 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, (s)he would not face a particularly higher risk to be retained in a subsequent year than if (s)he was initially retained.

Comparing Models (2) and (3) in Table 5 to Model (1) allows us to judge the bias in the treatment

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

Treatment: C versus A evaluation after grade 8	(1) Model: grade-varying unobserved heterogeneity			(2) Model: grade-constant unobserved heterogeneity			(3) Model: without unobserved heterogeneity		
	ATT		95% CI	ATT		95% CI	ATT		95% CI
I. Evaluation in grade 9: A									
All treated	0.003		[-0.089, 0.100]	0.028		[-0.075, 0.129]	-0.068		[-0.155, 0.025]
II. High school diploma									
All treated	-0.135	***	[-0.230, -0.044]	-0.126	***	[-0.206, -0.051]	-0.168	***	[-0.247, -0.089]
III. Delay at start last compulsory year									
All treated	0.900	**	[0.808, 0.987]	0.844	***	[0.739, 0.938]	1.024		[0.951, 1.100]

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.

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.

5.2.2 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 in the pupils' ability. 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 6 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 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

Table 6: 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	-0.008		[-0.119, 0.101]	–		–
II. High school diploma						
All treated	-0.123	***	[-0.222, -0.033]	-0.140	***	[-0.231, -0.059]
III. Delay at start last compulsory year						
All treated	0.875	***	[0.784, 0.962]	0.867	***	[0.796, 0.934]

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.

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 hardly changes over time.

Table 7: 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	-0.177	***	[-0.274, -0.962]
II. Track at start of grade 9: THS			
All treated	0.043		[-0.066, 0.156]
III. Track at start of grade 9: VHS			
All treated	0.134	***	[0.072, 0.212]
IV. Track in last compulsory year: GHS			
All treated	-0.131	***	[-0.217, -0.057]
V. Track in last compulsory year: THS at least once			
All treated	-0.011		[-0.111, 0.098]
VI. Track in last compulsory year: VHS			
All treated	0.142	***	[0.063, 0.224]

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 FNT and GGR we allow the treatment effect to depend on observed and unobserved charac-

teristics, 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 insight 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 (A-8) in Online Appendix C.2. In Table 8 we report the average treatment effects of a retention in grade 8 by these quartiles. Given that we restrict our analysis on the subsample of retained students, those with very high values in terms of the linear index must be very close to the threshold above which they would have obtained a B rather than a C. We therefore label pupils within the fourth quartile of this index, as those "at the margin of retention". In line with the research based on RDD, we find that these marginal students increase their academic achievement in the subsequent grade, be it only significantly at the 10% level. The ATTs on high school graduation and delay are also less adverse for this quartile. As FNT, 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.[31] 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.

The finding that grade retention has the most adverse implications for pupils with the lowest ability is surprising, because one could expect that these pupils are the least ready for promotion. Yet, promoting them appears to be a better option than having them repeat the grade. To understand this apparent contradiction, one should not forget that our analysis only identifies the *relative* outcome of retention versus that of not being retained. In *absolute* terms promoting these low ability pupils might still result in poor outcomes. Our findings only signify that by retaining these students matters will not improve, presumably because the extra time that these students thereby obtain to catch up with the grade level requirements is not sufficient to counterbalance the psychological costs associated with retention. The fact that in Flanders retained students lack sufficient professional guidance might explain this, but whether such professional support would be sufficient to reverse the negative consequences of grade retention remains to be proven.

[31] GGR 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.

Table 8: 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	-0.153		[-0.356, 0.042]
Second quartile	0.003		[-0.152, 0.151]
Third quartile	0.060		[-0.074, 0.180]
Fourth quartile	0.105	*	[-0.010, 0.206]
II. High school graduation			
First quartile	-0.265	***	[-0.461, -0.083]
Second quartile	-0.119		[-0.303, 0.043]
Third quartile	-0.062		[-0.229, 0.077]
Fourth quartile	-0.093		[-0.270, 0.061]
III. Delay at start last compulsory year			
First quartile	1.055		[0.869, 1.277]
Second quartile	0.902		[0.742, 1.073]
Third quartile	0.840	**	[0.688, 1.000]
Fourth quartile	0.802	**	[0.647, 0.962]

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.

5.2.3 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)[32] as alternative remediation strategy. In Table 9 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). Neither the ATT on the evaluation in the subsequent grade, nor the effect on high-school graduation of this alternative remediation strategy is significantly better than grade retention. Awarding a B rather than a C does lead to significantly less delay by the end of high school. However, the delay remains more important than in the counterfactual of an A (see Table 5). More generally, 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).

The fact that overall this alternative remediation strategy is not significantly better than grade retention does not mean that it may not perform better (or worse) at some levels of student's ability. Therefore, in

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

Table 9: Grade Repetition (C) Relative to Forced Downgrading (B) after Grade 8

Treatment: C versus B evaluation after grade 8	Model: grade-varying unobserved heterogeneity		
	ATT		95% CI
I. Evaluation in grade 9: A			
All treated	-0.038		[-0.119, 0.044]
II. High school diploma			
All treated	-0.049		[-0.143, 0.037]
III. Delay at start last compulsory year			
All treated	0.719	***	[0.591, 0.840]

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 forced downgrading 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.

Table A-13 of Online Appendix F, we report the ATTs after grade 8 of a C versus a B outcome by the four quartiles discussed above. This additional analysis shows that for the three highest quartiles the ATTs of a C versus a B roughly mirrors the aforementioned overall ATTs mentioned in Table 10 suggesting only minor heterogeneity in this treatment effect. By contrast, for the lowest ability group both the short- (evaluation in grade 9) and long-run effect (high school graduation) of a C versus a B resembles the corresponding ones found for the contrast between C and A, as reported in Table 9. This suggests that for the lowest ability group remediation by forced downgrading (B) would lead to a better performance than grade retention (C). Nevertheless, it would not improve upon just passing the grade (A) and, hence, it would neither be a remediation strategy to recommend. Finally, forced downgrading later in the educational career, i.e. after grade 9 or grade 10 (Table A-14), leads to qualitatively similar outcomes as in grade 8 (Table 9).

6 Conclusions

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 considering multiple ordered *discrete* schooling outcomes instead of a single *continuous* outcome, al-

lowing for multiple grade retention, proposing a method to deal with initial conditions and taking the partial observability of the 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 to higher drop-out rates, substantial schooling delay and downgrading within the hierarchical tracking system in Flemish high school. In line with the findings of FNT and GGR, 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 studies using RDD 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, improves relative to retention only in that it does not lead to as much schooling delay by the end of high school. However, relative to unconditionally passing to the next grade, there is no improvement. Hence, the challenge remains to find more successful remediation strategies.

Acknowledgements

Matteo Picchio acknowledges financial support by the Research Foundation - Flanders (FWO) when he was at Ghent University from October 2011 until October 2012. We are grateful to the Steunpunt SSL of the Flemish government for making the SONAR survey data available. We also wish to thank the participants in the Labour Health seminar in Tilburg (2013), in the seminar in Seville (2013), in the DSSE-ISFOL workshop on Human Capital and Education in Rome (2013), in the CESifo conference in the Economics of Education in Munich (2013), in the EALE Conference in Turin (2013), in the AIEL conference in Rome (2013), in the SIE Conference in Bologna (2013), in the seminar of the Department of Economics and Social Sciences in Ancona (2013), in the seminar of Competence Centre on Microeconemetric evaluation - EC JRC in Ispra (2016), in the seminar of the Department of Economics in Sheffield (2017), in the LEER workshop on Education Economics in Leuven (2017), in the seminar of the Department of Economics in Bristol (2017). We are particularly grateful for the very constructive comments of three anonymous reviewers which allowed us to significantly improve the quality of the manuscript.

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

Alet, E., L. Bonnal, and P. Favard (2013) 'Repetition: Medicine for a short-run remission.' *Annals of Economics and Statistics* 111/112, 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

Beckett, M., J. Da Vanzo, N. Sastry, C. Panis, and C. Peterson (2001) 'The quality of retrospective data: An examination of long-term recall in a developing country.' *Journal of Human Resources* 36(3), 593–625

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

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

Chamberlain, G. (1980) 'Analysis of covariance with qulaitiative data.' *Review of Economic Studies* 47, 225–238

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(2), 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

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

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 (2016) 'Grade retention and unobserved heterogeneity.' *Quantitative Economics* 7(3), 781–820

Gillian, P. (2002) 'Biases in the reporting of labour market dynamics.' Working Paper 02/10, Institute for Fiscal Studies (IFS)

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., and E. J. Vytlacil (2007) 'Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation.' In *Handbook of Econometrics,* ed. J. J. Heckman and E. Leamer, vol. 6B (Amsterdam: Elsevier) chapter 70, pp. 4779–4874

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

Krinsky, I., and A.L. Robb (1986) 'On approximating the statistical properties of elasticities.' *Review of Economics and Statistics* 68(4), 715–719

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 (2016) 'Screening for a chronic disease: A multiple stage duration model with partial observability.' *International Economic Review* 57(3), 915–934

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

Sudman, S., N. M. Bradburn, and N. Schwarz (1997) *About Answers: The Application of Cognitive Processes to Survey Methodology* (CA: San Francisco: Jossey-Bass Publishers)

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

von Fintel, D., and D. Posel (2016) 'Errors in Recalling Childhood Socio-economic Status: The Role of Anchoring and Household Formation in South Africa.' *Social Indicators Research* 126, 119–140

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

Woutersen, T., and J.C. Ham (2013) 'Calculating confidence intervals for continuous and discontinuous functions of parameters.' CeMMAP working paper 23/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, London