A Dynamic Model of Effort Choice in High School

Olivier De Groote*

June 11, 2020

Abstract

I estimate a dynamic model of educational decisions when researchers do not have access to measures of study effort. Students choose the academic level of their program and the probability to perform well. This differs from a pure discrete choice model that assumes performance follows an exogenous law of motion. I investigate high school tracking policies and obtain the following results: (1) encouraging underperforming students to switch to less academic programs reduces grade retention and dropout, (2) the decrease in the number of college graduates is small, and (3) a pure discrete choice model would ignore changes in unobserved study effort and find a large decrease in the number of college graduates.

Keywords: high school curriculum, early tracking, dynamic discrete choice, CCP estimation, effort

JEL: C61, I26, I28

*Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France. E-mail: olivier.de-groote@tse-fr.eu. This paper benefited from helpful comments at various stages from Peter Arcidiacono, Estelle Cantillon, Jan De Loecker, Koen Declercq, Thierry Magnac, Arnaud Maurel, François Poinas, Jo Van Biesebroeck, and Frank Verboven and several audiences at KU Leuven and Duke University. It also benefited from discussions during and after presentations at ULiège, RUG Groningen, UAB Barcelona, Toulouse School of Economics, University of Cambridge, LMU Munich, McGill University, UNC Chapel Hill, Oslo University, ENSAE-CREST, IWAEE 2017, EALE 2017, ECORES and LEER summer schools 2017, NESG 2018, CEPR Applied IO 2018, ES 2018 European Winter Meeting, SOLE 2019, and the Barcelona GSE summer forum 2019. I also thank Jan Van Damme, Bieke De Fraine, and the Flemish Ministry of Education and Training, for providing the LOSO dataset. Funding for this project was generously provided by the Research Foundation Flanders (FWO). Olivier De Groote acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

1 Introduction

Students follow different curricula during secondary education, depending on their preferences and ability. Many countries separate students in academic or vocational tracks. Academic curricula do not focus on skills that are directly useful in the labor market but provide preparation for higher education. To achieve the European 2020 target of 40% college-educated people, many countries aim to induce more students to choose academic curricula. In the US, there is a similar trend towards more academic course taking, especially in STEM (Science, Technology, Engineering, Math)-fields.[1]

This trend raises two related concerns. First, it is unclear how many students experience a causal effect of a more academic curriculum on success in higher education. Second, students who would experience large difficulties in completing an academic trajectory would waste time and effort they could otherwise spend on training skills that are of direct use in the labor market. Mismatch and failure can demotivate students and lead to unfavorable outcomes such as grade retention and dropout. These outcomes generate large costs for both students and society. In this paper, I investigate how to design policies that help in matching students to a study program. This is a general concern in the design of educational systems, but it is especially important in countries that allocate students into entirely different tracks at an early age.[2]

I use a dataset that combines rich information on student characteristics and ability, study program attendance, and performance in secondary education with data on higher education careers. I use micro-data of Flanders, the largest region of Belgium. Study programs consist of tracks and elective courses within each track. Tracking occurs gradually. Students choose a program when they enter high school at

[1] The 2011 NAEP report compares high school students graduating in 2005 to students graduating in 1990. They find that they take more academic credits (16 on average instead of 13.7). The percentage of students that followed a rigorous curriculum also increased from 5% to 13% (Nord et al., 2011).

[2] Germany and Austria differentiate from the age of 10. Belgium and the Netherlands differentiate from age 12. Most of these early tracking countries also face much higher rates of grade retention (OECD, 2013).

age 12 but this choice can be updated almost every year. Students that underperform can switch out of an academically rigorous program or have to repeat the grade. I study the impact of study programs that differ in their academic level on higher education enrollment and graduation and the extent to which the gradual tracking policy helps to improve these outcomes. Next, I look at an alternative policy that prevents students from repeating grades.

I develop a dynamic model of educational decisions in which high school students make yearly decisions about their level of academic effort. They do this by making a discrete choice between the study programs in their choice set and by deciding on the probability of good performance at the end of the year. Allowing students to influence both the study program and the distribution of performance is novel and it is particularly important for the counterfactual simulations in this paper as we expect students to change their (unobserved) study effort in response to tracking policies. The decision of a study program is based on the effort cost of studying today and the impact on future utility. The effort cost depends on (1) a fixed cost, independent of the expected performance, and (2) a variable cost, increasing in the probability to obtain a good performance outcome at the end of the year. Good performance is costly but required to continue in the program and eventually graduate. I allow for uncertainty in the form of performance and taste shocks, which could lead students to make decisions that are ex-post suboptimal. From a social perspective, mismatches also arise, in particular, because students do not take into account the costs of grade repetition and drop out on government spending.

In a pure discrete choice model, the Conditional Choice Probabilities (CCPs) in the data can be rationalized as choices made by optimizing agents in a given policy context. This way, utility can be identified and new CCPs can be calculated in counterfactual policy simulations (Magnac and Thesmar, 2002; Rust, 1987). Many dynamic models additionally include (other) state variables that transition over time but are assumed to be outside the control of the agents in the model, i.e. they are modeled as exogenous laws of motion. In the classic Rust (1987) bus engine model, CCPs capture the effect of mileage on replacing a bus engine. Conditional on this decision, transitions of mileage are outside the agent's control. In contrast to the CCPs, they are interpreted as primitives and kept fixed in counterfactual simulations. Therefore, a counterfactual increase in replacement cost would predict postponing some engine replacements, but it cannot predict a less intensive use of buses with older engines. In educational models, state transitions are often the result of a stochastic performance outcome, such as course grades or credits (Arcidiacono (2004), Arcidiacono et al. (2016), Declercq and Verboven (2018), Eckstein and Wolpin (1999), and Joensen and Mattana (2017)). By not determining it within the model, it excludes any effect a policy can have on the study effort students exert.

I adapt the pure discrete choice model by adding an unobserved choice variable: "effective study effort". This is a scalar that allows students to set the distribution of state transitions. In this context, it means they choose the probability to be successful at the end of the year. Making this a choice variable does not change the fit of the model to the data (as we can perfectly match this probability in the data using a flexible function of state variables), but it changes the impact of counterfactual simulations if study effort is expected to change. This innovation does not require additional data. Instead, I make use of an Euler equation that arises naturally in dynamic models but is not exploited in a pure discrete choice model: a first-order condition that equates the (unobserved) marginal cost of improving the distribution of performance today to the marginal benefit it is expected to generate in the future. The benefit of a better performance outcome (i.e. a better state) is easy to derive from a dynamic model. At its optimal level, this has to be consistent with the observed distribution of performance. In a pure discrete choice model, this trade-off does not arise as agents are assumed to be unable to change the performance distribution, regardless of its cost. The identifying power of this first-order condition allows me to transform flow utility and state transitions (primitives in the pure dynamic discrete choice models and identified in Magnac and Thesmar (2002)) to a new set of primitives: the fixed cost of a study program and the marginal costs of increasing effective study effort. Because primitives of the pure discrete choice model serve as inputs for the identification of the primitives in the current model, it is straightforward to apply standard approaches to identify and estimate a model with persistent unobserved heterogeneity and avoid solving it during estimation by applying CCP estimation and finite dependence (Arcidiacono and Miller, 2011; Hotz and Miller, 1993; Hu and Shum, 2012).

To look at the impact on higher education, I simultaneously estimate the structural effect of high school study program and grade retention on enrollment in and graduation from higher education. I control for gender, parental education, and two continuous measures of cognitive ability at the start of high school. To allow for persistent heterogeneity that is not captured by these controls, I allow for a discrete factor structure in the form of two unobserved types. The primitives during high school are functions of these observables and types and therefore generate heterogeneous predictions of high school decisions and outcomes. Moreover, they also have a direct impact on enrollment in and graduation from higher education. This allows me to control for both observed and unobserved characteristics of students in the estimation of the effects of high school programs and study delay on higher education outcomes. Unobserved heterogeneity and the causal effects of high school outcomes can be identified without strong functional form assumptions by exploiting rich panel data (Freyberger, 2018; Hu and Shum, 2012). The Flemish context is particularly useful for this purpose as most students make choices and obtain (at least one) performance outcome every year, during at least six years. Furthermore, some variables do not have a direct impact on higher education, which helps the identification of causal effects under weaker assumptions on the nature of unobserved heterogeneity (Heckman and Navarro, 2007). The results are also robust to applying an alternative identification strategy that follows Carneiro et al. (2003) and uses a large set of measures of unobserved traits to identify the types.

I find a positive impact of a higher academic level of study programs on obtaining a higher education degree, and a negative impact of past grade retention. Counterfactual simulations are therefore needed to predict the impact of policies that are expected to change both in opposite directions. Allowing underperforming students to switch to a program of lower academic level (downgrade) as an alternative for repeating a grade has important benefits in the long run. Without this possibility, the number of students with grade retention and high school drop out would increase by a third and the number of college graduates would decrease by 4%. A second counterfactual shows that this policy can be improved. Prohibiting students to repeat a grade (if they have the option to downgrade), would decrease the number of retained students by a third and dropout by 11%. Enrollment in higher education would go down but the number of students that obtain a higher education degree would not decrease significantly. A welfare analysis shows that this policy can create a loss for students, but it is largely offset by the taxpayers' gains through decreased educational spending and increased tax returns. The large impact of initial conditions on effort costs suggests these gains should be invested in early childhood education.

I compare this to the predictions of a pure discrete choice model and find an underestimation of the positive effects in both simulations. This can be explained by students reacting to the policy by improving their study effort. This makes them less likely to repeat grades or dropout, and more likely to graduate from an academic program. This in turn has an impact on higher education. In the policy where underperforming students are no longer allowed to repeat a grade, a pure discrete choice model would predict an important reduction in the number of college graduates (2.5%) while a model that allows for changes in study effort finds a much smaller and statistically insignificant effect of less than 1%.

This paper contributes to three strands of literature. First, I contribute to the estimation of dynamic discrete choice models in general, and educational decision making in particular. Since the seminal contribution of Keane and Wolpin (1997), dynamic discrete choice models have often been used to evaluate the impact of counterfactual policies on educational decisions. This includes the decision to stay in high school, go to college, or choose a major.[3] Allowing students to be forward-looking is important because they are expected to react in advance to an impact on their utility in the future. In the pure discrete choice model, the only channel through which students can respond is through observable choices like a study program or the years of schooling. This excludes any direct response on performance outcomes through the study effort they exert during the year. This means that the model is not capable of generating the results that were found in theory (Costrell, 1994), field experiments (Dubois et al., 2012), and natural experiments (Garibaldi et al., 2012). Therefore, some recent papers have included observable measures of study effort in the model (Fu and Mehta, 2018; Todd and Wolpin, 2018), but these are often unavailable and can only proxy for the actual effort students exert. I propose an alternative strategy that can be used when only program and performance data are available. Similar to Hu and Xin (2019), I allow for an unobserved choice variable that influences state transitions. The difference between our approaches is that I do not impose an exclusion restriction in the state transition rule.[4] Instead, I make use of a first-order condition coming from the dynamics in the model. A similar idea is applied in some job search models when the probability to find a job is modeled as a choice variable (Cockx et al., 2018a; Paserman, 2008; van den Berg and van der Klaauw, 2019). I embed this idea within the framework of dynamic discrete choice models in the spirit of Rust (1987), Hotz and Miller (1993) and Arcidiacono and Miller (2011): I formulate

[3] See e.g. Eckstein and Wolpin (1999), Arcidiacono (2005), Joensen and Mattana (2017) or Declercq and Verboven (2018). The same type of models have also been used to look at the impact of wage returns on educational decisions (Arcidiacono, 2004; Beffy et al., 2012).

[4] Hu and Xin (2019) demonstrate this by using time as an exclusion restriction in a finite horizon model. This rules out using it when the horizon is infinite or when other primitives depend on time. The latter is particularly important in educational choice models such as this application. This is because the grade a student is in, combined with the years of study delay are perfectly collinear with the calendar year. Both are expected to influence the flow utility of a program, thereby violating the exclusion restriction.

identification and estimation strategies that are generally applicable when researchers want to allow agents to have a direct impact on the distribution of state variables through an unobserved choice. I discuss the importance of this approach in the context of educational decisions and quantify the impact by showing the differences it generates in relevant policy simulations.

A second strand of literature investigates the returns to educational investments. Altonji et al. (2012) review the literature on the effects of high school curriculum on educational attainment and wages, initiated by Altonji (1995). Several papers find positive effects for intensive math courses (Aughinbaugh, 2012; Goodman, 2019; Joensen and Nielsen, 2009; Rose and Betts, 2004). Papers that compare academic and vocational curricula stress the importance of comparative advantages and heterogeneous effects (Kreisman and Stange, 2017; Meer, 2007). While investing in certain high school programs matters, selection into them is not random. This explains why investing in early childhood education is effective because it induces students to opt for better programs in later life through dynamic complementarities (Cunha and Heckman, 2009; Cunha et al., 2010; Heckman and Mosso, 2014). A separate literature looks at the causal impact for an individual student of being retained in school (Cockx et al., 2018b; Fruehwirth et al., 2016; Jacob and Lefgren, 2009; Manacorda, 2012). I contribute to this literature by jointly analyzing high school program choice and grade retention within a structural model.[5] This approach has several advantages: (1) it allows me to simulate an implementable policy to reduce grade retention, which goes beyond an interpretation of treatment effects by taking into account the impact of the threat of grade retention on effort decisions, (2) counterfactual policies illustrate dynamic complementarities by comparing remedial strategies in high school to the effects of initial conditions, (3) I identify new primitives that quantify the cost of grade retention and differences in effort costs between students with different initial

[5] I distinguish between high school programs of different academic level (tracks) and look at differences within tracks, based on math-intensity and the inclusion of classical languages in the curriculum.

conditions.

Finally, I contribute to the literature on educational tracking policies. Several papers look at the impact of tracking students at an early age (Cummins, 2017; Duflo et al., 2011; Fu and Mehta, 2018; Hanushek and Woessmann, 2006; Pekkarinen et al., 2009; Roller and Steinberg, 2020) or the long-run impact of the academic track for specific groups of students (Dustmann et al., 2017; Guyon et al., 2012). Cockx et al. (2018b) look at average treatment effects for outcomes within high school for students that are forced to repeat grades or switch tracks. Recent evidence also suggests that switching track can diminish the negative consequences of early track choice (De Groote and Declercq, 2020; Dustmann et al., 2017). I contribute to this literature by investigating the impact of flexibility in tracking policies during secondary education.

The rest of the paper is structured as follows. Section 2 highlights the methodological contribution of the paper by discussing the intuition in a simple two-period binary choice model and by showing how the model is identified without additional data. Section 3 describes the institutional context, the data, and policy issues, and section 4 applies the model to the data. I discuss the estimation results in section 5 and I simulate tracking policies in section 6. Finally, I conclude in section 7.

2 The cost of effort

This section describes the methodological contribution of the paper. To provide intuition, I discuss this in a two-period, binary choice model of educational decisions in which unobservables only enter through iid taste and performance shocks. Students choose to drop out of school and they can do that in period 1 (high school) or period 2 (college). Performance enters the model as a dummy equal to one when a student obtains a high school degree, which is required to access higher education. I will show how this model is identified in the main text and refer to the appendix for a proof that holds in a more general case that allows for multiple alternatives, time periods, and performance outcomes, as well as a time-invariant unobserved state variable. Finally, I discuss how to estimate the model without solving it, using the CCP estimator (Hotz and Miller, 1993).

2.1 A standard model set-up

Consider a student i right before entering the final year of high school in period $t = 1$. He can decide to stay in school ($j = 1$) by finishing high school. He can also choose to drop out of school ($j = 0$). If i stays in school, he incurs a flow utility equal to

$$u(x_i) + \varepsilon_{ij1}.$$

x_i is a vector of time-invariant state variables that are known to the econometrician and the student, such as parental background and observed measures of ability. ε_{ijt} is an extreme value type 1 distributed taste shock, unobserved by the econometrician but observed at time t by the student. In the spirit of Keane and Wolpin (1997), I call $-(u(x_i) + \varepsilon_{ij1})$ the "effort cost" of going to school.

After $t = 1$, i can obtain a high school degree. This creates a state variable in the form of a dummy that is only available in $t = 2$: $g_i = 1$. I allow for uncertainty and assume students obtain a degree if

$$f(x_i) + \eta_i > 0$$

with $f(x_i)$ an index and η_i a logit shock, realized in $t = 2$. The probability to obtain a degree is then

$$\Pr(g_i|x_i) = \frac{\exp(f(x_i))}{1 + \exp(f(x_i))}. \tag{1}$$

Students who drop out ($j = 0$) in $t = 1$ never return to school and receive a lifetime utility of dropping out. In $t = 2$, students that obtained a high school degree have to option to stay in school ($j = 1$) by going to college. They can also still opt for drop out. The lifetime utility after leaving high school is specified as follows:

$$\Psi_j(x_i) + \varepsilon_{ij2}.$$

I do not distinguish between different sources of this utility and directly estimate $\Psi_0(.)$ and $\Psi_1(.)$ as policy-invariant functions of x_i.[6] To keep this model simple, I make the following (strong) assumptions that are important to relax in the application, but avoid some complexities in explaining the methodological contribution: I do not allow for grade retention by assuming that students who failed have to drop out, and I assume $\Psi_0(x_i, g_i) \equiv \Psi_0(x_i)$. This implies that a high school degree only has value when it is used to get into college. As only differences in utility are identified, we need to treat $\Psi_0(x_i)$ as known (Magnac and Thesmar, 2002). I will, therefore, set $\Psi_0(x_i) = 0$ and interpret the other $\Psi_j(x_i)$ as differences in expected lifetime utility.

2.2 Solution when performance is not determined within the model

If we assume the state transition, i.e. the probability to obtain a degree, is exogenous to the student, we can directly solve the model by backward induction. In $t = 2$, i has a choice only when he obtained a high school degree. Since $t = 2$ is the final period, this is equivalent to a static model. He chooses to go to college if $\Psi_1(x_i) + \varepsilon_{i12} > \varepsilon_{i02}$ and drop out otherwise. Let d_{it} be a dummy equal to 1 if i chooses to go to school at time t. The probability to go to school in $t = 2$ for students with a high school degree is given by:

$$\Pr(d_{i2} = 1 | x_i, g_i = 1) = \frac{\exp(\Psi_1(x_i))}{1 + \exp(\Psi_1(x_i))} \ . \tag{2}$$

[6] These values could follow from differences in expected discounted wages, but also because of an intrinsic value of (not) going to college.

In period 1, the problem is dynamic. Students do not know if they will be successful in school. They only know the distribution of g_i and make a decision under uncertainty. There is also uncertainty about the next period taste shocks. The lifetime utility of drop out is given by $\Psi_0(x_i)+\varepsilon_{i01} = \varepsilon_{i01}$. The lifetime utility of choosing the high school option is represented by the conditional value function (added with taste shock ε_{i11}) :

$$v(x_i) = u(x_i) + \beta\gamma + \beta \left(\Pr(g_i|x_i) \ln\left(1 + \exp \Psi_1(x_i, g_i = 1)\right)\right) \tag{3}$$

with γ the Euler constant, $\beta \in (0, 1)$ the one-period discount factor and $\ln\left(1 + \exp \Psi_1(x_i, g_i = 1)\right)$ the logsum expression, net of γ.[7] The probability to go to school in period 1 is then given by

$$\Pr(d_{i1} = 1|x_i) = \frac{\exp v(x_i)}{1 + \exp v(x_i)} \ . \tag{4}$$

After fixing β, we can estimate $f(x_i)$, $u(x_i)$, and $\Psi_1(x_i)$ using a parametric form with the logit probabilities in (1), (2), and (4) in a likelihood function. They can also be recovered nonparametrically from log odds in the data.[8]

2.3 The wrong primitives

Assume now we want to use this model for a counterfactual that changes the impact of a high school degree. For example, we would like to know the change in college enrollment when college utility Ψ_1 increases. This would have a direct impact on college enrollment among high school graduates through (2). In a dynamic model, this would also have an impact before the college enrollment decision because a high school degree is required to enter college. We can see this in (4) with conditional value

[7] The logsum expression is the result of integrating over the type 1 extreme value taste shocks.

[8] E.g. $\Pr(d_{i2} = 1|x_i, g_i = 1)$ at each realization of x_i can be used to recover the value of higher education: $\Psi_1(x_i) = \ln\left(\frac{\Pr(d_{i2}=1|x_i,g_i=1)}{1-\Pr(d_{i2}=1|x_i,g_i=1)}\right)$.

functions specified in (3), which is increasing in Ψ_1. With a constant $\Pr(g_i|x_i)$, more students will have a high school degree and will be able to choose to go to college, further increasing college enrollment.

However, this is likely an underestimation of the effect. Staying in high school to go to college only makes sense for students who obtain a high school degree. This gives them a reason to exert more study effort to make sure they succeed. Therefore, keeping $\Pr(g_i|x_i)$ fixed in (4) is problematic. In particular, $f(x_i)$, the estimated index that predicts high school completion, is likely not policy invariant. Students with the same characteristics x_i will be more likely to obtain a high school degree after the change in college utility. Furthermore, an increase in study effort is expected to decrease the flow utility of going to school. Therefore $u(x_i)$ is also not policy-invariant.

2.4 Existing approaches in the literature

Performance takes the form of obtaining a high school degree. Later in this paper, it will also refer to evaluations at the end of each year in high school. Other examples of performance that have been used in dynamic discrete choice models are course grades (Arcidiacono, 2004; Arcidiacono et al., 2016; Eckstein and Wolpin, 1999), course credits (Declercq and Verboven, 2018; Joensen and Mattana, 2017), college admission probabilities (Arcidiacono, 2005) or length of study (Beffy et al., 2012).

The papers avoid the problem by assuming students are not able to influence study effort.[9] This is a strong assumption that has been rejected in both field and natural experiments (Dubois et al., 2012; Garibaldi et al., 2012). Nevertheless, in many cases, it can be argued that the impact of a particular counterfactual through this channel will be limited. As explained in the introduction of this paper, this is

[9]As in Keane and Wolpin (1997), one could instead ignore a year of schooling that is not successfully completed, i.e. $d_{i1} = 1$ only if we also observe $g_i = 1$. This way, we do not need to take stance on the process of g_i. A counterfactual change in d_{i1} now also captures the students that were already in school but needed this extra incentive to get a degree. The problem with this approach is that we do not observe how the probability to obtain a degree changes because of the policy and we cannot run counterfactuals that change the implications of that.

not the case for the counterfactuals put forward in this paper.

An alternative solution is to specify (3) and (1) such that $u(.)$ and $f(.)$ depend on a vector of observed measures of effort e, with $\nabla f(e) > 0$ and $\nabla u(e) < 0$ as in Todd and Wolpin (2018) and Fu and Mehta (2018). If e is also optimally chosen in the model, it could capture the expected change in study effort and the invariant mapping between f and x and u and x is no longer problematic. The problem with this approach is that we need to observe measures of effort and identify their impact on performance and utility, separately from student background. In most datasets, such measurements are also unavailable or are unlikely to capture its full extent.

2.5 Alternative approach: new primitives

Instead of assuming the flow utility (=- effort cost) in high school is constant, we now let it be a linear function of an index that characterizes the distribution of performance, called *effective* study effort. Let the probability to obtain a high school degree be given by $\phi(y_i)$ with $y_i \in (0, +\infty)$, a continuous choice variable, and $\frac{\partial \phi(y_i)}{\partial y_i} > 0$. We specify a linear utility function in y_i:

$$u(x_i, y_i) = -C^0(x_i) - c(x_i)y_i. \tag{5}$$

with $c(x_i) \equiv -\frac{\partial u(x_i,y_i)}{\partial y_i}$ the marginal effort cost of changing y_i and $C^0(x_i) \equiv -u(x_i, 0)$ a fixed effort cost. I assume the probability to obtain a high school degree is given by

$$\phi(y_i) = \frac{y_i}{1 + y_i}. \tag{6}$$

With the current functional form assumptions, y_i can be interpreted as the odds of obtaining a high school degree ($y_i = \frac{\phi(y_i)}{1-\phi(y_i)}$). Therefore, we can interpret $c(x_i)$ as the marginal cost to increase the odds by one unit and $C^0(x_i)$ as the component of the effort cost that does not change with the probability to obtain a degree. The fixed cost, therefore, captures a distaste to go to school because of differences in preferences or social norms. The cost can also be negative because students might enjoy going to school or parents can reward (or force) them to go. $c(x_i) > 0$ is the marginal cost of increasing the odds to obtain a degree. A higher level of y_i corresponds to a higher cost but also a higher chance to obtain a degree.

The equations above impose two assumptions that appear unintuitive if we consider y_i to be a measure of effort: (1) utility is linear in y_i (see (5)) and (2) performance depends only on y_i, not on student characteristics (x_i) (see (6)). To discuss these assumptions, I start by clarifying how y_i can be interpreted as a measure of *effective* study effort.

For ease of interpretation, we could write effective study effort as $y_i = \mathcal{Y}_i(e_i)$ with e_i a vector of all measures of effort that matter for performance (such as hours of study or minutes of attention paid in class) and $\mathcal{Y}_i(.)$ a production function. This function could depend on individual characteristics (such as ability or parental inputs). In this model, we assume the researcher does not know $\mathcal{Y}_i(.)$ or e_i, but specifies a transformation of performance outcomes y_i and allows this to be a choice variable of the student. This provides a flexible way to model effort, especially without access to effort data.

First, we can interpret marginal cost estimates as capturing differences in the cost of studying because of differences in student characteristics, even without identifying the production function. A student of low ability is expected to have a worse technology $\mathcal{Y}_i(.)$ and would, therefore, need a higher number of hours of study in e_i for the same increase in y_i than a high ability student. In the model, this will imply a higher marginal cost $c(x_i)$.[10] In models with more time periods, marginal costs can flexibly depend on past choices and outcomes to also capture how this influences study efficiency today. In contrast to the estimates of a production function, the interpretation of marginal cost estimates is in utility units, not in their effect on performance. This is important for defining the primitives of the model as we will allow performance to

[10]Further in this section we extend the state space with unobserved heterogeneity which allows for similar statements about the impact of unobserved ability.

change in counterfactuals.

Second, a counterfactual simulation is expected to change e_i but choice probabilities and welfare depend on $y_i = \mathcal{Y}_i(e_i)$. Therefore, we do not require data on e_i and we do not need to make assumptions on $\mathcal{Y}_i(.)$. We can even be more agnostic about the source of the change in y_i as it could also go through induced changes in study efficiency.

The assumption that utility is linear in y_i (5) is key for the identification of $C^0(x_i)$ and $c(x_i)$. This assures that fixed and marginal costs are primitives of the model and can be used in counterfactual simulations. Note that this holds for the specific choice of the researcher about y_i, i.e. about how the performance outcome depends on a choice variable that linearly enters the utility function.[11] In a pure discrete choice model, a researcher does not have this flexibility as utility and state transitions are assumed to remain the same in counterfactuals. The linear structure is required here to define the transformations of utility that are primitives. In this context, y_i is chosen to be the odds of obtaining a degree. This means that the linearity assumption should be interpreted as an assumption that the marginal cost of improving the odds does not change in counterfactual simulations. For counterfactuals that involve small changes in the utility of students, this choice might not be very important as the estimation approach will estimate the primitives in such a way that they predict the data in the status quo. For larger changes in behavior, it is important to choose y_i according to the application. An alternative assumption would be to let y_i be the probability to obtain a degree. However, a policy-invariant marginal cost would imply the unrealistic assumption that the cost to increase the probability to graduate from 50 to 51% is equally costly than to increase it from 98 to 99%. It is much more reasonable to assume a cost function that is increasing and convex, i.e. the utility function should be decreasing and concave. This is the case if y_i is defined as the

[11] This is a crucial difference with a pure discrete (or a discrete/continuous) model that would include some observable measure e_i in the model as e_i is data and cannot be specified by the researcher.

odds instead.[12]

The fact that y_i is defined by the researcher also explains why x_i does not enter the probability to obtain a high school degree as we can always define a new $y_i^{'}$ such that $\phi(y_i') \equiv \phi(x_i, y_i)$. In the application, I will show that in the case of discrete or multiple performance outcomes, it can be useful to make the probability depend on some elements of x_i to impose a structure that uses a single measure of effective effort to keep the model tractable with a diverse set of discrete performance outcomes.

2.6 Solution with endogenous performance distribution

The conditional value function of going to high school is now given by:

$$v(x_i, y_i) = u(x_i, y_i) + \beta\gamma + \beta\left(\phi(y_i)\ln\left(1 + \exp\Psi_1(x_i)\right)\right) \tag{7}$$

Instead of assuming the performance distribution is exogenous, we can let y_i be a choice variable by maximizing v with respect to y_i. This gives the following first-order condition (FOC):

$$\frac{dv(x_i, y_i)}{dy_i} = \frac{\partial u(x_i, y_i)}{\partial y_i} + \beta\left(\frac{\partial\phi(y_i)}{\partial y_i}\ln\left(1 + \exp\Psi_1(x_i)\right)\right) = 0 \text{ if } y_i = y_i^* \tag{8}$$

with y_i^* the optimal choice of y_i, $\frac{\partial u(x_i,y_i)}{\partial y_i} = -c(x_i)$ and $\frac{\partial\phi(y_i)}{\partial y_i} = (1 + y_i)^{-2}$.[13] This FOC is an Euler equation that equalizes today's marginal cost of effective study effort to its marginal benefit in the next period. In this (simple) case, we obtain a closed-form solution:

$$y_i^* = \sqrt{\frac{\beta\ln\left(1 + \exp\Psi_1(x_i)\right)}{c(x_i)}} - 1. \tag{9}$$

We see that the optimal level of y_i increases in the discounted surplus of being able

[12]Let $\bar{g}_i = \phi(y_i)$, then $\frac{du(x_i,y_i)}{d\bar{g}_i} = \frac{\partial u_i}{\partial y_i}\frac{d\phi^{-1}(\bar{g}_i)}{d\bar{g}_i} < 0$ with $\frac{\partial u_i}{\partial y_i} = -c(x_i) < 0$ and $\frac{d\phi^{-1}(\bar{g}_i)}{d\bar{g}_i} = \frac{1}{(1-\bar{g}_i)^2} > 0$, $\frac{d^2u(x_i,y_i)}{d\bar{g}_i^2} = -c(x_i)\frac{2}{(1-\bar{g}_i)^3} < 0$.

[13]The second-order condition is always satisfied: $\frac{d^2v(x_i,y_i)}{dy_i^2} = -2\beta(1+y_i)^{-3}\ln\left(1 + \exp\Psi_1(x_i)\right) < 0$.

to enter college ($\beta \ln(1 + \exp \Psi_1(x_i))$) and decreases in its marginal cost $c(x_i)$. Note that requiring an interior solution for y_i puts an upper bound on marginal costs. If they are larger than the discounted benefit of a high school degree, the student would have no incentive to exert effort.[14]

We can now revisit the counterfactual of section 2.3. An increase in Ψ_1 will result in a higher value of y_i (see (9)). According to (6), this leads to a better probability to obtain a degree $\phi(y_i)$. It also leads to lower flow utility because of the increase in the variable component of the effort cost $c(x_i)y_i$ (see (5)). Not only Ψ_1 but also flow utility and state transitions in the conditional value function (7) are now changing because the model is capturing a change in study effort.

2.7 Identification

I first discuss the identification of this simplified model which shows how a first-order condition can provide identifying power for the new primitives of the model. I then explain the intuition of the general proof of identification that can be found in the Appendix section A.

2.7.1 The simplified model

$\Psi_1(x_i)$ is still identified from period 2 choices as before. The difference now is that the performance index $f(x_i)$ and the flow utility $u(x_i)$ are endogenous (i.e. determined within the model) and depend on two new primitives: fixed costs $C^0(x_i)$ and marginal costs $c(x_i)$. Magnac and Thesmar (2002) show that for given state transitions, we can identify $u(x_i)$.[15] State transitions are nonparametrically identified from the data,

[14] This follows from an identification approach that uses a first-order condition and is satisfied when students expect a non-zero probability of obtaining a high school degree. In applications, it is standard to (implicitly) assume this on state transitions (see e.g. Arcidiacono (2005); Beffy et al. (2012); Declercq and Verboven (2018); Eckstein and Wolpin (1999); Joensen and Mattana (2017)). Hu and Xin (2019) show that exclusion restrictions can be used to identify a discrete unobserved choice.

[15] Note that dynamic discrete choice models are nonparametrically unidentified. Magnac and Thesmar (2002) show that we need to specify the distribution of ε_{ijt}, the discount factor β and

which implies the identification of the index $f(x_i)$. As they are now endogenous, we identify them at their optimal level in the data: $u^*(x_i)$ and $f^*(x_i)$. We cannot treat them as primitives anymore, but they remain useful from an identification point of view because we can use them to recover alternative primitives in a second step.

To see this, first note that the optimal level of effort in the data, y_i^*, is a transformation of the identified index $f^*(x_i)$. This is because (9) shows that students with the same value of x_i, choose the same level of y_i : $y_i^* \equiv y^*(x_i)$ and $\phi(y^*(x_i)) \equiv \Pr(g_i|x_i)$. Therefore, from (1) and (6) at the optimized value of y_i^* it follows that:

$$y_i^* = y^*(x_i) = \exp(f^*(x_i)). \tag{10}$$

With the identified objects $\Psi_1(x_i)$, $u(x_i, y^*(x_i)) \equiv u^*(x_i)$, and $y_i^* \equiv y^*(x_i)$, we can now proceed to the identification of the new primitives of the model. The FOC allows us to identify marginal costs from the marginal benefits at y_i^*. Rearrange (8) and substitute (10) such that marginal costs (a primitive) can be written as a function of the identified objects (assuming β is known):

$$c(x_i) = \beta \left(\frac{\ln(1 + \exp \Psi_1(x_i))}{(1 + y^*(x_i))^2} \right). \tag{11}$$

To identify the fixed costs, we substitute $y_i^* = y^*(x_i)$ into the utility function of high school (5) and rearrange:

$$C^0(x_i) = -u^*(x_i) - c(x_i)y^*(x_i). \tag{12}$$

Intuitively, we are exploiting data on performance outcomes in a more structural sense than in a pure discrete choice model. If two students have the same future value $\Psi_1(x_i)$ but different state transitions (so different $f^*(x_i)$), it now has to be rationalized by differences in marginal costs. While students who make different choices in period

normalize the utility of one option to be able to identify $u(x_i)$.

1 despite having the same future values and state transitions provide the variation to identify the fixed cost component of utility.

2.7.2 The general case

Dynamic discrete choice models are identified only after fixing β, normalizing the utility of one option, and specifying the taste shock distribution (Magnac and Thesmar, 2002). In the Appendix section A, I show that the degree of under-identification is the same as in the standard set-up of a pure dynamic discrete choice model as in Magnac and Thesmar (2002). It is therefore straightforward to generalize the model to multiple periods, alternatives, discrete performance outcomes, and time-varying state variables that affect all primitives of the model. Moreover, if we can identify a time-invariant unobserved state variable in the pure discrete choice model, we can also allow for it in the current model.

To summarize the extent of the proof, add a time subscript to the state variable x_{it} (which can contain t to allow for non-stationarity) and let ν_i be an unobserved time-invariant state variable or "type". Results from the literature on the identification of dynamic models can be used to identify $u_j^*(x_{it}, \nu_i)$ and $f_j^*(x_{it}, \nu_i)$ for each option j in the choice set (up to standard normalizations). The researcher now also needs to specify a transformation of $f_j^*(x_{it}, \nu_i)$ for which it is reasonable to assume that its effect on utility is linear and for which a FOC is expected to be satisfied in the data. After this choice, we can use the FOC to map the endogenous objects $u_j^*(x_{it}, \nu_i)$ and $f_j^*(x_{it}, \nu_i)$ into the primitives of the model: the intercept of the utility function ($-C_j^0(x_{it}, \nu_i)$ in the current model) and the marginal utility of the chosen transformation of $f_j^*(x_{it}, \nu_i)$ ($-c_j(x_{it}, \nu_i)$ in the current model).

The reason we can allow for both unobserved ability (in the form of types) and effort choice follows directly from the fact that we do not attempt to separately identify ability from a measure of effort we would observe in real life (such as hours of study). The choice variable here is *effective* study effort. By construction, this merges the impact that ability, hours of study, or any other variable known to the agent would have on performance. Performance outcomes in the data can still deviate from a prediction based on effective study effort, but only through an unexpected shock. This is why we could directly identify effective study effort from the observed data by integrating over the shocks when unobserved ability did not enter the model (see (10)). To allow for unobserved ability, we first need to identify how performance depends on it before we can do this, but this is the same requirement as in a pure discrete choice model.

2.8 The CCP method

The Hotz and Miller (1993) inversion theorem shows that the expected value of behaving optimally in the future can be identified directly from the distribution of taste shocks (ε), the conditional choice probabilities (CCPs) in the data $\Pr(d_{it}|x_i)$ and flow utility $u^*(x_i)$. This is especially convenient in the multi-period case as it does not require us to solve the model by backward induction during estimation, but it can also be illustrated in the simpler model. With extreme value type 1 taste shocks and a terminating action in the choice set, we can replace the expected value of behaving optimally in the future by a simple expression of CCPs.

To see this, note that the probability to drop out in period 2 is given by

$$\Pr(d_{i2} = 0|x_i) = \frac{1}{1 + \exp \Psi_1(x_i)},$$

taking logs and rearranging terms:

$$\ln(1 + \exp \Psi_1(x_i)) = -\ln \Pr(d_{i2} = 0|x_i).$$

This is the expression of the expected value of behaving optimally in the future (net of the Euler constant) that also entered the conditional value functions. Substituting this into the marginal costs (11):

$$c(x_i) = \beta \left(\frac{-\ln \Pr(d_{i2} = 0 | x_i)}{(1 + y^*(x_i))^2} \right)$$

and fixed costs still follow from (12).

3 Institutional background and data

This section describes the institutional context in Flanders (Belgium) and introduces the data. I make use of the LOSO dataset in which I follow a sample of 5, 158 students that started secondary education in 1990.[16] Students were actively followed during high school and therefore the data contains many individual characteristics, choices, performance outcomes, and test scores. Afterward, the students were asked to respond to surveys about educational and labor market outcomes which provide information about their higher education career. Details about the data and the context are omitted from this text but discussed in Appendix B.

3.1 Study programs

After finishing six grades in elementary school, students enroll in high school in the 7th grade, usually in the calendar year they become 12 years old. Students can choose between all schools in Flanders since school choice is not geographically restricted and free school choice is law-enforced. After obtaining a high school degree, they can enroll in higher education.

In full-time education, they choose between different high school programs, grouped into tracks that differ in their academic level. The academic track has the most aca-

[16]The LOSO data were collected by Jan Van Damme (KU Leuven) and financed by the Flemish Ministry of Education and Training, on the initiative of the Flemish Minister of Education. Note that throughout the paper I discuss the data for this sample of 5,158 students, which covers 80% of the original sample. In the data appendix I discuss in detail why some observations were dropped. This is mainly because some variables were missing, but also because students made choices that were not consistent with the tracking systems as explained in this paper.

demically rigorous curriculum. It provides general education and prepares for higher education. The middle track prepares students for different outcomes. Therefore, I follow Cockx et al. (2018b) and distinguish between a track preparing mainly for higher education programs (middle-theoretical), and a track that prepares more for the labor market (middle-practical). Students can also choose the vocational track. This track prepares them for specific occupations that do not require a higher education degree. Within each track, there is a choice between several programs. I aggregate them up to eight study programs. I split the academic track into four programs: classical languages, intensive math, intensive math + classical languages, and other. The middle-theoretical track is split between intensive math and other. This aggregation still allows for a substantial number of students in each program and corresponds to important differences in enrollment and success rates in higher education (Declercq and Verboven, 2015).

A student graduates from high school after a successful year in the 12th grade in the academic or one of the middle tracks, or the 13th grade in the vocational track. Leaving in the vocational track after grade 12 is also not considered dropout as students still obtain a certificate that is valued on the labor market (they do need to finish the 13th grade to have access to higher education). Compulsory education laws require students to pursue education until June 30th of the year they reach the age of 18. From the age of 15, they can also decide to leave full-time education and start a part-time program in which work and school can be combined.

Although each track prepares for different options after high school, enrollment in almost any higher education option is free of selection. Students from any track can enroll in almost any program of higher education (Declercq and Verboven, 2018). Therefore, selection into higher education only takes the form of self-selection. Similar to Declercq and Verboven (2015, 2018), I distinguish between three levels of higher education (professional college, academic college, and university) and allow for STEM and non-STEM majors. For universities, I also distinguish between five different campuses in Flanders.

3.2 Mobility

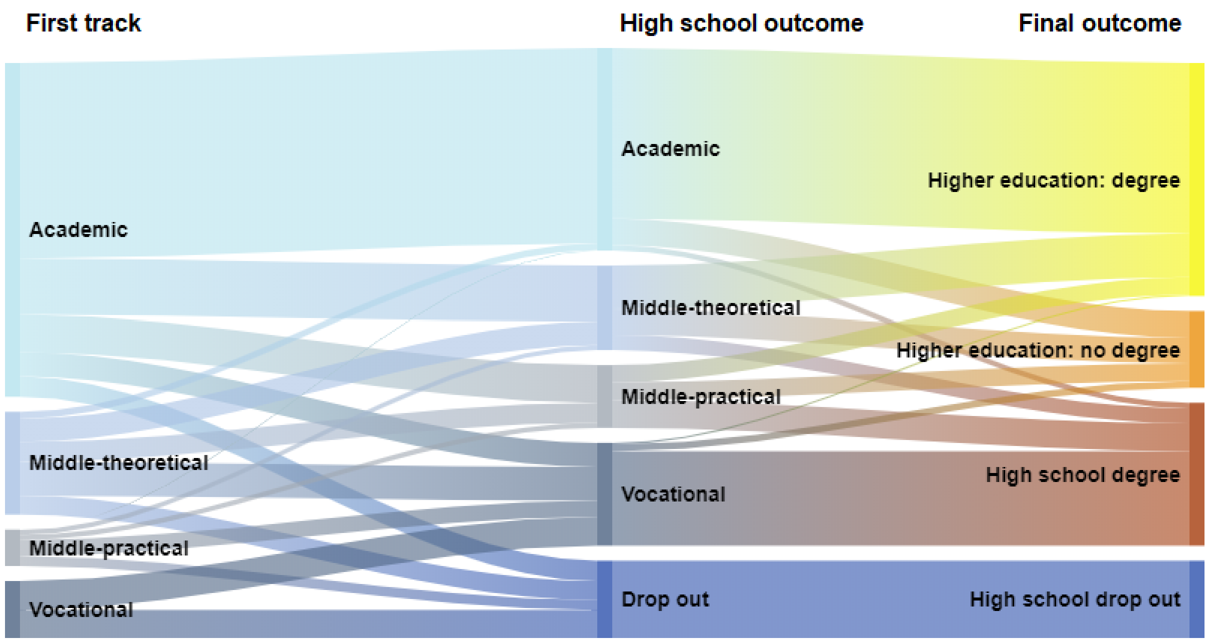
At the start of secondary education, all programs are available. The choice set in the future depends on the past program and performance during the year. Upward mobility, i.e. moving from a track of lower academic level to a more rigorous one, is practically impossible, except for switches between middle tracks and the academic track in the first two grades. Similarly, students can never enroll in programs with classical languages if they did not choose it from the start. Math-intensive programs are available from grade 9 on and the same restrictions apply. Finally, there can be no more switching between full-time programs from grade 11 on.[17]

The restrictions imply that a wrong choice at age 12 can have large consequences for educational attainment. As there can be uncertainty about performance and future preferences, many students prefer to keep their options open by choosing the academic track in the beginning and gradually move towards their final program. Figure 1 summarizes both the movements within high school (left side) and between high school and the final educational outcome (right side).[18] Most students start in the academic track but many transition to another one. There are similar movements from the other tracks but almost always in a downward fashion. Nevertheless, students that move down do not necessarily give up on obtaining a higher education degree. While it is very common for students that graduated from the academic track to obtain a higher education degree, most students in the middle-theoretical track and some in the middle-practical track also obtain one. For students that graduated from the

[17] These rules are not always formal and students have the legal right to ignore them. Nevertheless, this is a realistic description of the perceived rules by students as schools often advertise them as being binding. Cockx et al. (2018b) apply a similar set of rules. In Appendix B I discuss the data cleaning and more details about the rules. This shows that only a small number of observations have to be dropped because they are inconsistent with this description.

[18] A more detailed overview of transitions to higher education can be found in the Appendix Table A10 and Table A11. As in Declercq and Verboven (2018), I define a degree as three successful years of higher education in a time span of six years.

Figure 1: Transitions in the educational system



Note: Left: program chosen in grade 7, middle: last choice before leaving secondary education, right: final educational outcome. See Appendix Table A14 for data on these transitions. Students in the vocational track only obtain a full high school degree that gives access to higher education after an additional 13th grade. They obtain another type of degree after grade 12 and are therefore not considered dropouts if they leave before grade 13. Figure created using Google Charts.

vocational track, this is very uncommon. The students who started in the vocational track also often drop out of high school, which excludes the possibility to go to higher education.

3.3 Performance and the tracking policy

The transitions in high school are not always a smooth and voluntary process. Each study program comes with its performance standards. Teachers are expected to keep the quality standards within the program at a certain level. This is done by handing out a certificate to students, based on their performance during the year. An A-certificate is given to students that did not fail a single course. They can move to the next grade and continue in the program. If they failed on some courses, teachers need to decide on the certificate. This can still be an A-certificate, e.g. for students that only failed a small number of courses, but it can also be a B- or a C-certificate. A C-certificate means that the student failed on too many important courses and must repeat the grade to continue in full-time secondary education. A B-certificate indicates that the student failed on some important courses within the program. He can proceed to the next grade, but not in every program. Alternatively, a student with a B-certificate can decide to repeat the grade without being excluded from a program. In most cases, a B-certificate excludes the track a student is currently in and therefore encourages them to downgrade to another track. However, a B-certificate can also exclude elective courses only (see appendix Table A12). Most of the time students obtain an A-certificate. 7.1% of the certificates are B-certificates and 6.6% are C-certificates. A C-certificate always leads to grade retention if students do not want to leave full-time education, but also 1 out of 4 students with a B-certificate chooses to repeat grades.

Although the number of B- and C-certificates each year is low, many students obtain at least one of them during their high school career. 35% of students obtain a B-certificate and 30% a C-certificate. This results in a large degree of grade retention. Table A13 summarizes the number of students that obtain a B- or C-certificate or accumulate study delay. It then compares their educational outcomes with that of the average student. 32% of students leave high school with at least one year of study delay. These students are 22 %points less likely to enroll in and 24 %points less likely to graduate from higher education than students who were not retained. Part of this is also explained by the higher dropout rates in high school.

3.4 Implications for optimal policies and the required model

Figure 1 shows large differences between higher education outcomes when we compare different tracks.[19] These effects can be driven by observed and unobserved initial conditions that could have a direct impact on higher education outcomes. Appendix Table A9 shows the differences between student characteristics and their final program in high school. The dataset contains measures of cognitive ability (language and math), gender, and socioeconomic status (SES). The latter is defined as a dummy equal to one if at least one of the parents has completed higher education. Academic programs attract mostly female, high ability students with a high SES. In the model, I will control for these characteristics and also allow for an unobserved type to capture other characteristics (such as noncognitive ability).

Even if the effects are not all driven by initial conditions, it is still not clear if encouraging students to choose academically rigorous programs will have an impact on higher education outcomes as for many of them it might be too costly to exert enough effort to succeed, leading to study delay or even drop out. Therefore, the model should be able to identify how each policy affects two choices the students make: their study program and their study effort.

I will evaluate two policies that change the choice set of students that obtain a B-certificate. In the first counterfactual, I investigate the impact of allowing students to avoid grade retention by switching to another program. I do this by simulating what would happen if a B-certificate was equivalent to a C-certificate and forced students to repeat a grade. A second counterfactual looks at a new policy that is currently being implemented in Flanders. It simulates the impact of not allowing students with a B-certificate to repeat the grade.[20] This policy follows from a concern that grade

[19] Appendix Table A10 shows that there are also differences between study programs of the same track.

[20] This measure is part of a bigger reform in secondary education and is being applied on cohorts that enter high school from September 2019 on. In the actual implementation of the policy, it will still be possible in some cases to repeat the grade but only if students get their teachers' explicit permission (source: answer by the Flemish minister in parliament at 4 October 2018 on question 2410 in period 2017-2018).

retention is too high and costly for society. However, a potential threat is that this might reduce the number of students in tracks that prepare for higher education and thereby decrease the number of college graduates.

4 Application of the model

This section introduces a dynamic model of educational choices in the Flemish context. It applies the methodological contribution discussed in section 2 to be able to let students change their study effort in response to counterfactual policy changes. The notation is similar to section 2 but needs to be adapted to a more realistic context. Throughout the model, i refers to a student, t the time period in years, $j = 1, ..., J$ are mutually exclusive study programs and $j = 0$ is an outside option, i.e. not attending school. The program choice is given by $d_{it} = j$ if j is chosen. After each year t in high school, students obtain a vector of performance outcomes $g_{it+1} \in G$ that, combined with the track and grade they are currently in, defines their choice set Φ_{it+1}. At the end of high school, it reflects the high school degree. Students have to make two decisions in each year: their study program (d_{it}) and the distribution of their performance outcomes through a single index y_{it}. As in section 2, we close the model at college entrance but from there we also predict college graduation as a function of individual characteristics and (endogenous) high school outcomes to evaluate the impact of counterfactuals that take place in high school. I assume a finite time model by assuming students can no longer attend high school in $t = 10$.

4.1 The choice set

Each study program in high school belongs to one of four tracks: academic (*acad*), middle-theoretical (*midt*), middle-practical (*midp*), and vocational (*voc*). Within the academic track, students can also choose for math-intensive programs (*math*), and/or classical languages (*clas*) in the curriculum. In the middle-theoretical track, they can also choose for a math-intensive program. The tracks are available throughout secondary education, i.e. grade 7 to 12 (and 13 in the vocational track). The *clas* option starts at the same time, while the *math* options start in grade 9. Next to the full-time education system, there is also a part-time vocational program (*part*). This program is available from the moment a student is 15 years old and does not have a grade structure.

The choice set of students, Φ_{it}, is restricted. First, students can never upgrade tracks according to the following hierarchy: $acad > midt > midp > voc > part$, except for the first two grades in which mobility between *acad*, *midt* and *midp* is allowed. Second, within those tracks, they can choose *math* and *clas*, but only if they also did this from the first grade these options were available.[21] Finally, from grade 11 on, students who want to stay in full-time education must stay in the same program.

Students progress in secondary education by obtaining a certificate at the end of the year.[22] As explained in the institutional context, the flexibility of a B-certificate can have different implications for the choice set. Therefore, I use the certificate data to create variables that capture the permission for a student to enter in each program. Let the performance vector be $g_{it+1} = (g_{it+1}^{track}, g_{it+1}^{math}, g_{it+1}^{clas})$. The main performance outcome in the model is $g_{it+1}^{track} \in \{0, 1, 2, 3, 4\}$. The lowest outcome (0) does not allow any track in the next grade. Each increase corresponds to a track of higher academic level being available (vocational (1), middle-practical (2), middle-theoretical (3), and academic (4)). $g_{it+1}^{math} \in \{0, 1, 2\}$ indicates if a student can go to *math*. Since there are two tracks in which this is possible, we distinguish between the possibility to choose *math*, only in track *midt* ($g_{it+1}^{math} = 1$) or in both *midt* and *acad* ($g_{it+1}^{math} = 2$). $g_{it+1}^{clas} \in \{0, 1\}$ indicates if a student can go to *clas*.

[21] However it is allowed to switch from *acad* without extra math to *midt* with extra math in a later grade.

[22] There is one exception. Students can enroll in grade 8 of the vocational track without having succeeded grade 7. Therefore, the lowest performance outcome is a B-certificate and the effort costs of students in grade 7 of the vocational track is captured by a fixed component only.

After obtaining a high school degree, students can decide to enroll in different higher education options or choose the outside option (not going to school). I distinguish between three levels of higher education, STEM and non-STEM majors, and I distinguish between different locations. From the age of 18, students also have the option to leave school without a high school degree. I assume this is a terminal choice, i.e. they never return to secondary education.

4.2 End-of-year performance

Track restrictions

Let the performance measure in $t+1$, be the result of effective study effort y_{it} and a logistically distributed shock η_{it+1}^{track} such that:

$$g_{it+1}^{track} = \tilde{g} \text{ if } \bar{\eta}_{jr}^{\tilde{g}} < \ln y_{it} + \eta_{it+1}^{track} \leq \bar{\eta}_{jr}^{\tilde{g}+1} \tag{13}$$

where $\bar{\eta}_{it}^{\tilde{g}}$ denotes the threshold to obtain at least outcome $\tilde{g}$. This threshold is allowed to differ through the program i is in at time t (j) and the grade he is in (r). At time t, students know y_{it}, but they do not know the realization of g_{it+1}^{track} because of the shock η_{it+1}^{track}. The information students do have at time t is the probability of obtaining an outcome $\tilde{g}$ in a given program ($d_{it} = j$) and grade ($grade_{it} = r$):

$$\Pr(g_{it+1}^{track} = \tilde{g} | y_{it}, d_{it} = j, grade_{it} = r) = F(\ln y_{it} - \bar{\eta}_{jr}^{\tilde{g}}) - F(\ln y_{it} - \bar{\eta}_{jr}^{\tilde{g}+1}) \tag{14}$$

with $F(a) = \frac{\exp(a)}{1+\exp(a)}$ the cumulative distribution function of the performance outcome. Setting $\bar{\eta}_{jr}^{0} = -\infty$ and $\bar{\eta}_{jr}^{5} = +\infty$ guarantees that all probabilities add up to 1 in each program and each grade. Since (13) remains the same when adding or subtracting the same term on all sides, I normalize one of the thresholds $\bar{\eta}_{jr}^{1} = 0$.

One way to interpret effective study effort is that the main performance outcome is modeled as an ordered logit model with index $\ln y_{it}$. An alternative interpretation can be seen by rewriting (14) for the lowest realization of the performance outcome:

$$y_{it} = \frac{1 - \Pr(g^{track}_{it+1} = 0 | y_{it}, d_{it}, grade_{it})}{\Pr(g^{track}_{it+1} = 0 | y_{it}, d_{it}, grade_{it})}. \tag{15}$$

This shows that the effective study effort of a student can be interpreted as the odds of avoiding the lowest outcome (= no track available in the next grade). This model for a discrete performance outcome is a natural extension of the binary case we discussed in section 2 where we defined effective study effort as the odds of obtaining a high school degree. By choosing the odds of avoiding the lowest outcome here, students can change the probability of each realization. If y_{it} is close to zero, they are very likely to obtain the worse outcome. If y_{it} is large, they will probably reach the best outcome.

Note that in the final year of high school, a student can only obtain the best possible performance outcome in his track or fail. The performance outcome then becomes binary and characterizes the high school degree (as in the example in section 2). Similarly, other institutional constraints make it impossible to achieve certain realizations of the performance outcome. This implies that many of the thresholds in (13) are known from the institutional context and not estimated.

Course restrictions

Additional course restrictions are also modeled as an ordered logit (conditional on the realization of g^{track}_{it+1}) to predict g^{math}_{it+1} and g^{clas}_{it+1}. I specify indexes for each course ($math$ and $clas$):

$$\begin{gathered} \alpha^{math}_y \ln y_{it} + S_i' \alpha^{math}_S + \nu_i' \alpha^{math}_\nu + \eta^{math}_{it+1}, \\ \alpha^{clas}_y \ln y_{it} + S_i' \alpha^{clas}_S + \nu_i' \alpha^{clas}_\nu + \eta^{clas}_{it+1}. \end{gathered} \tag{16}$$

with η^{math}_{it+1} and η^{clas}_{it+1} logistically distributed shocks and α^{math}_y and α^{clas}_y measuring how much of the effective study effort matters for each elective course (compared to its impact on track performance). The logit shocks η_{it+1}^{math} and η_{it+1}^{clas} are assumed to be independent of η_{it+1}^{track} but dependence between outcomes is captured by taking into account the outcome g_{it+1}^{track} to influence individual-specific threshold levels.[23] I also allow for comparative advantage in elective courses by estimating the influence of student characteristics that can be observed (S_i) or unobserved (ν_i), with ν_i a vector of dummy variables for each unobserved type the student can be.

Finally, I define $\phi_{ijt}^{\bar{g}}(y_{it})$ as the joint probability of $\bar{g} = \{\bar{g}^{track}, \bar{g}^{clas}, \bar{g}^{math}\}$ which is the product of the three ordered logit probabilities. Note that dependence of $\phi_{ijt}^{\bar{g}}(y_{it})$ on i and t goes through x_{it} and $\nu_i$as it comes from the dependence on the current grade a student is in and the initial characteristics that influence comparative advantages in elective courses.[24]

4.3 Study program

The expected lifetime utility of each program is represented by the conditional value function:

$$
\begin{aligned}
& v_j(x_{it}, \nu_i, y_{it}) + \varepsilon_{ijt} \\
& = u_j(x_{it}, \nu_i, y_{it}) + \beta \sum_{\bar{g} \in G} \phi_{ijt}^{\bar{g}}(y_{it}) \bar{V}(x_{it+1}(\bar{g}), \nu_i) + \varepsilon_{ijt} \text{ for } j \in se
\end{aligned}
\tag{17}
$$

with $v_j(x_{it}, \nu_i, y_{it})$ the conditional value function for student i with observed state variable x_{it} and unobserved type ν_i of choosing program j and effective study effort y_{it} at time t. ε_{ijt} is an extreme value type 1 taste shock. The observed state variable contains the information set students and the econometrician share. This includes the observed student background (S_i), but also time-varying and endogenous variables

[23] If $g_{it+1}^{track} < 3$, then $g_{it+1}^{math} = g_{it+1}^{clas} = 0$. If $g_{it+1} = 3$, then $g_{it+1}^{math} \in \{0, 1\}$ and $g_{it+1}^{clas} = 0$.

[24] The current grade is a direct result of two elements in x_{it} (the grade a student was in last year and the performance outcome) and the current program choice j (as student might have a choice to repeat a grade).

such as the previous program choice and the performance outcome.[25] It also contains t to capture nonstationarities. Because all shocks in the model are assumed to be iid, the unobserved type ν_i will capture persistent differences between students that are not captured by the observables.

The first term is the flow utility of schooling, $u_j(x_{it}, \nu_i, y_{it})$. As in Keane and Wolpin (1997), I interpret flow utility as the negative of an effort cost of going to school. The second term is the expected value of the future, discounted at $\beta \in (0, 1)$. This depends on the ex-ante value functions $\bar{V}(x_{it+1}, \nu_i)$, i.e. the value functions integrated over the future iid shocks. As in Rust (1987), this implies that students do not know future realizations of taste shocks, but they know the distribution. The performance vector g is the only stochastic element in x. Integrating over future states is therefore equivalent to writing a weighted sum over potential outcomes in the set G, with the joint probability of the performance outcome ($\phi_{ijt}^{\bar{g}}(y_{it})$) as a weight.

As explained in section 2, I do not consider the effort costs to be a primitive of the model. Instead, I assume it is the linear sum of a fixed cost ($C_j^0(x_{it}, \nu_i)$), and a variable cost of effort ($c_j(x_{it}, \nu_i)y_{it}$):

$$u_j(x_{it}, \nu_i, y_{it}) = -C_j^0(x_{it}, \nu_i) - c_j(x_{it}, \nu_i)y_{it}. \tag{18}$$

Students will choose the program j and effective study effort y_{it} that gives them highest expected lifetime utility. $C_j^0(x_{it}, \nu_i)$ and $c_j(x_{it}, \nu_i)$ are the primitives of the model to estimate. As in Section 2, the linearity assumption implies a constant, and policy-invariant marginal cost of increasing the odds to avoid the lowest performance outcome.

[25] I do not include choices or outcomes before $t-1$ in x_{it}. This way the model satisfies a first-order Markov property.

4.4 Closing and solving the model

I assume leaving secondary education is a terminal action, i.e. students never return to high school. They either leave the education system or (if they obtained a high school degree) they choose one of the higher education options. To avoid making assumptions on how students expect their wages and college success to evolve, I close the model at higher education enrollment and parameterize its expected lifetime utility. In particular, I assume the conditional value functions for choice options after high school take the following form:

$$v_{ijt}(x_{it}, \nu_i, y_{it}) = \text{Degree}_{it}'\mu^{\text{degree}} + \Psi_j^{HEE}(x_{it}, \nu_i) \text{ if } t = T_i^{SE} + 1 \tag{19}$$

with T_i^{SE} the last period student i spends in high school, $\text{Degree}_{it} \in x_{it}$ a vector of dummy variables for the different types of high school degrees a student can obtain, μ^{degree} a vector of parameters to estimate and $\Psi_j^{HEE}(.)$ a function of the state variables that predicts the higher education enrollment (HEE) decision.[26] The lack of a future value term in this conditional value function requires the parameters to be interpreted as the total expected lifetime utility from enrolling in option j. If the student obtained a high school degree, j is a specific higher education option or an outside option of which the utility, net of the value of a degree, is normalized: $\Psi_0^{HEE} = 0$. I distinguish between three different levels of higher education (professional college, academic college, and university), two majors (STEM and non-STEM), and (for universities) five locations. If the student did not obtain a high school degree, he can only obtain the value of $j = 0$. By normalizing $\Psi_0^{HEE} = 0$, all cost parameters in high school should be interpreted as the one period difference with respect to the expected lifetime value of leaving high school without a degree. Note that this includes potential wages for high school dropouts. Therefore, a high cost of schooling

[26]I estimate a common value of a high school degree, an interaction effect with the academic level of the program and a separate effects for finishing 12th grade in the vocational track and obtaining a high school degree in the vocational track due to the specific nature of the vocational track that requires students to study an additional year in order to obtain the degree.

can also be interpreted as an opportunity cost.

Since only differences in utility are identified, I could set the flow utility of leaving school to 0 for the current and future periods (as I did in the simplified model of Section 2). However, in contrast to a static model, normalizing the utility of one option in every state is not innocuous (Kalouptsidi et al., 2018). Here it would imply an assumption that students only exert effort in school to have the possibility to go to higher education, and not for other benefits that come from obtaining a high school degree (in particular, the better labor market opportunities for students that do not go to college). Therefore, I also estimate the value of a high school degree and set $v_{i0t} = \text{Degree}'_{it}\mu^{\text{degree}}$. As in Eckstein and Wolpin (1999), the value of a degree can be identified from choices in secondary education. In particular, differences in dropout rates between students with low and high chances of obtaining a high school degree help to identify this effect, but exclusion restrictions are needed to separate this from differences in fixed costs. Therefore, I allow distance to college to (endogenously) affect performance in high school, but it is excluded from fixed costs.[27]

Each year students without a high school degree that did not drop out choose between the study programs j that are in their choice set Φ_{it}. If they are legally permitted to do so, they can also drop out by choosing the outside option. They do this by choosing the option with the highest conditional value function, in which the optimal level of the effective study effort is chosen. If they have a high school degree, they can choose between the outside option and different higher education options. This utility is also maximized with respect to the effective study effort y_{it}.[28] Appendix section C describes the full solution of the model.

[27]Also some functional form assumptions help for identification. As explained in the estimation subsection, I restrict the differences in fixed costs over grades to be linear in each track. This helps identification here because higher grades also imply a decrease in uncertainty about eventual graduation and a higher discounted value of the degree.

[28]The decision will often be a joint decision by parents and their child, after advice from teachers. I do not distinguish between these different actors and simply assume some utility function is optimized, regardless of who makes the decision. See Giustinelli (2016) for a paper that does makes this distinction using additional data.

4.5 Graduation in higher education

I simultaneously estimate the parameters of a reduced form conditional logit model with $\Psi_j^{HED}(x_{it}, \nu_i)$ the estimated index that predicts graduation in each campus-level-major combination, conditional on student characteristics, high school program, study delay, and the higher education enrollment decision. Note that I did not make any assumptions on students' expectations of higher education success. I simply estimate parameters for $\Psi_j^{HEE}(.)$ to capture the expected lifetime utility of enrolling in each college option. The reason I can do this is that the counterfactual simulations of this paper will only change the high school system, not the higher education system. Therefore, it will result in different high school outcomes (program, dropout, years of study delay), leading to differences in higher education outcomes, but not to differences in the mapping between high school and higher education outcomes. Therefore, the parameters of $\Psi_j^{HEE}(.)$ and $\Psi_j^{HED}(.)$ are policy-invariant and can be used in counterfactuals. This part of the model is similar to dynamic treatment effect models where some behavioral assumptions can be avoided, while still looking at the causal impact of a counterfactual (Heckman et al., 2016).

4.6 Estimation

As shown in Appendix A, the primitives of the model can be identified after fixing β, normalizing the utility of one option, and specifying the taste shock distributions.[29] To estimate the model, we need to impose further structure because of the limited number of observations in each realization of the state. I summarize the estimation algorithm here but discuss it in detail in Appendix D.

To allow for unobserved heterogeneity, I estimate the model using the two-stage CCP estimator of Arcidiacono and Miller (2011):

[29] I follow Arcidiacono et al. (2016) and set the discount factor $\beta = 0.9$. I obtain almost identical results with $\beta = 0.95$ (see Appendix F).

Stage 1: estimate type distribution and reduced forms

Step 1: initial types

Assume there are two unobserved types in the population and assign each student a random probability to belong to each type. Use these types as weights in what follows.

Step 2: higher education estimates

Impose functional forms for $\Psi_j^{HEE}(.)$ and $\Psi_j^{HED}(.)$ and estimate them as parameters of a conditional logit, using maximum likelihood. Importantly, I allow them to depend on initial characteristics of students (both observed and unobserved) as well as the high school outcomes: final study program and accumulated study delay.

Step 3: reduced forms of high school data

Recover the optimal levels of the effective study effort and the performance thresholds by estimating an ordered logit model with index $\ln y_j^*(x_{it}, \nu_i)$, the log of the optimal value of the effective study effort, conditional on choosing program j. The index is specified as a parametric but flexible function of (x_{it}, ν_i). I also estimate an ordered logit to recover the parameters of the performance outcomes on elective courses. As in Arcidiacono et al. (2016), I obtain predicted values of CCPs by estimating a flexible conditional logit.

Step 4: update types

Use the CCPs in high school, together with predicted performance outcomes, higher education enrollment, and higher education graduation by type to update the individual type probabilities using Bayes rule.

Repeat this until convergence of the joint likelihood of the data.

Stage 2: estimate cost parameters

Use the logit probabilities with the CCP-representation of the conditional value functions to estimate the value of a degree μ^{degree} and a specification for fixed costs $C_j^0(.)$ using maximum likelihood, with type probabilities as weights. I assume fixed costs are functions of a program-specific constant, travel time, and switching costs between tracks and specializations through elective courses. They also differ because of individual characteristics S_i and unobserved type ν_i through an effect that is allowed to change in the level of the track and the elective courses. Finally, grade retention enters through a stock variable (accumulated study delay) and a flow variable (a dummy equal to one when repeating a grade). These effects are also allowed to differ by the academic level.

Finally, marginal costs $c_j(.)$ can be recovered from the FOC at the optimal levels of y_{it} in each program and state $(y_j^*(x_{it}, \nu_i))$, without imposing additional structure. Standard errors are obtained using a bootstrap procedure.[30]

4.7 Identification, ability bias and unobserved types

A first requirement to identify the model is to recover CCPs and state transitions as functions of the observed state variable x_{it} and the unobserved type ν_i. If ν_i would be observed, this step is trivial as we can simply use the observed choices and outcomes for each realization of (x_{it}, ν_i). Magnac and Thesmar (2002) then show that we need to normalize the utility of a reference alternative, specify the discount factor β and the distribution of ε_{ijt} to identify the flow utility in the current policy context (i.e. $u_j^*(x_{it}, \nu_i) \equiv u_j(x_{it}, \nu_i, y_j^*(x_{it}, \nu_i))$). The identification of the indexes that predict higher education enrollment and graduation are simpler because we do not need to separately identify flow utility from the entire impact of (x_{it}, ν_i) (French and Taber, 2011). In section 2 and Appendix A I show that once we recover flow utility and

[30] I sample students with replacement from the observed distribution of the data and use 150 replications. Since the EM algorithm takes some time to converge, I do not correct for estimation error in the probabilities to belong to each type.

state transitions, we can identify fixed costs $C_j^0(x_{it}, \nu_i)$ and marginal costs $c_j(x_{it}, \nu_i)$ by imposing a FOC. Intuitively, differences between flow utilities that follow from good and bad performance outcomes, rationalize the state transitions we observe the periods before.

The remaining identification problem is therefore no different than in a pure discrete choice model as it is situated in the first requirement: we need to identify the distribution of unobserved heterogeneity and its impact on CCPs and the transitions of states. We can then apply the results of Magnac and Thesmar (2002) to identify $u_j^*(x_{it}, \nu_i)$ (Hu and Shum, 2012) and Appendix A shows how to use this and the state transitions to identify the type-specific primitives $C_j^0(x_{it}, \nu_i)$ and $c_j(x_{it}, \nu_i)$.

Unobserved heterogeneity is needed to rationalize decisions and outcomes that differ for students that otherwise look identical. It enters the model in two forms: (1) taste and performance shocks (ε_{ijt} and η_{it}) and (2) a common type ν_i. As shocks are not correlated, only the types capture persistent unobserved heterogeneity and enter the CCPs and state transitions. This not only improves the fit of the model, but it is also important to avoid the potential ability bias on the estimated effects of high school outcomes on higher education outcomes, which are crucial for the policy simulations. It is important to note that I include good measures of initial math and language ability in x_{it}.[31] Nevertheless, these measures of ability might still miss some dimensions of persistent heterogeneity. Therefore, I add two unobserved types. The estimation strategy shows that we can identify types in the parametric framework of this model, but it is important to see that we exploit variation in the data that achieves identification of unobserved heterogeneity and causal effects of high school outcomes under a more flexible structure.

Hu and Shum (2012) prove the identification of a non-stationary first-order Markovian model for CCPs and state transitions at time t using data from $t+1$, t, $t-1$,

[31] A sensitivity analysis in Appendix F shows that including these improves the fit of the model and changes the magnitude of the main effects of simulations, especially if unobserved types are not included.

$t-2$, and $t-3$. They allow for a single unobserved trait that is allowed to transition over time. Because high school takes six years to complete and we add two stages after high school, this shows that no further structure is needed to identify CCPs and state transitions at the end of high school and the enrollment stage of higher education.

Identification is also aided by variables that affect higher education only through their impact in high school. In particular, I assume that (grade-specific) travel time to high school programs influences decisions and outcomes during high school but has no direct effect after secondary education. Therefore, it does not enter the CCPs and reduced forms that predict higher education enrollment and graduation. This allows for the identification of causal effects of high school outcomes on higher education without imposing further structure on the unobserved heterogeneity (Heckman and Navarro, 2007).[32] Note that Heckman and Navarro (2007) do not require an exclusion restriction. One can for example also use an identification at infinity strategy (Abbring, 2010; Heckman et al., 2016).

An alternative model restriction that avoids the full structure is that of a general factor model. Every program choice and performance outcome in the model is indeed a function of a common (discrete) factor. Heckman et al. (2016) use the restrictions imposed by a factor model to identify treatment effects in a dynamic setting. To identify a broad set of distributional treatment effects, they follow Carneiro et al. (2003) and add a system of measurements of the unobserved trait. While this provides clear identification and interpretation of the unobservable, they also mention recent work by Freyberger (2018) that avoids the need for data outside the model as choices and outcomes within the model can play the same role in the identification approach.[33]

[32] As we argue in De Groote and Declercq (2020), this context lends itself to the use of this instrument as students have many school options available to them and parents are therefore not expected to take this into account in their location decisions. Importantly, free school choice is protected by the Belgian constitution and prevents schools from cream-skimming or prioritizing students of the same neighborhood.

[33] Lin (2019) also applies this identification strategy of Freyberger (2018) in the context of a dynamic programming model.

In the case of a lagged dependent variable, allowing for one common factor requires six outcomes or choices in the data to identify the model. The gradual tracking system in Flanders provides a lot of variation for this purpose as we observe many program decisions and performance outcomes, largely exceeding this number.

In line with these identification results, I also do two robustness checks and show the results in Appendix F. First, as explained above, we can relax the exclusion restriction by adding measures of high school travel time to the equations that predict higher education outcomes. I find that travel time to the different high school programs has no direct impact on higher education and the counterfactual results do not change. Second, I additionally add (ordered) logit models that predict measurements of initial skills and parental background to the likelihood function of stage 1 of the estimation approach. I use a discretized measure of students' IQ when they enter high school, as well as answers by their last teacher in elementary school to questions that indicate levels of conscientiousness, extraversion, and agreeableness.[34] Furthermore, I add parental reports of their income category and work situation around the same time. I find that adding these measures does not have an important impact on the counterfactual simulations. Furthermore, they give more insights into the nature of unobserved heterogeneity. I find that unobserved types are important in capturing non-cognitive skills. The impact on IQ, parental income, and work situation has the same sign but it is much smaller (and not statistically significant for the work condition). This can be explained by the inclusion of controls for cognitive ability and SES.[35]

[34] I use the same questions as in Shure (2017).

[35] A potential downside of adding measurements is that it requires the finite number of types to explain measures that are not of direct importance to the model. I therefore do not use them in the main specification. I also do not use the measures as control variables as they are likely measured with error, not available for everyone and they would make the state-space larger, making it more difficult to obtain good estimates of the CCPs and state transitions in stage 1 of the estimator.

5 Estimation results

This section discusses the structural schooling cost estimates and the estimates of higher education outcomes. To check the fit of the model and to run counterfactual simulations, I use these estimates to solve the model. Once I have solved the model backward to find all the conditional value functions and effective study effort levels, I forward simulate all taste and performance shocks to simulate choices of the study program and the distribution of performance. The model fit and details about the simulations are explained in the Appendix section E. The model does a decent job in capturing the patterns in the data such that it can be used for the counterfactual simulations in the next section.

5.1 Effort costs

I focus the discussion on the impact of student characteristics on effort costs using Table 1 and the impact of study delay and switches in Table 2. Appendix Tables A15 and A16 contain the estimates unrelated to student characteristics and the interactions of student characteristics with elective courses. Table A20 shows the intrinsic value of a high school degree.

Table 1: Costs of schooling: student characteristics and academic level

	Fixed costs				**Log of marginal costs**			
	Baseline effect		Interaction with academic level		Baseline effect		Interaction with academic level	
Male	-19.161	(9.714)	17.748	(4.355)	0.740	(0.114)	0.098	(0.061)
Language ability	7.468	(5.485)	-36.328	(5.132)	-0.637	(0.075)	-0.095	(0.088)
Math ability	1.117	(5.081)	-23.197	(4.523)	-0.217	(0.063)	-0.370	(0.051)
High SES	-19.232	(15.555)	-18.782	(5.818)	-0.676	(0.219)	0.016	(0.101)
Type 2	-41.781	(16.422)	85.647	(11.607)	3.265	(0.370)	-0.368	(0.156)

Note: Estimates of a sample of 5,158 students or 33,239 student-year observations. Scale = minutes of daily travel time. The marginal costs in the model are a flexible function of state variables, this table summarizes them by regressing their logarithmic transformation on the same variables that enter the fixed costs. Ability measured in standard deviations. Type 2 = dummy equal to one if student belongs to unobserved type 2 instead of 1. High SES= at least one parent has higher education degree. Level = academic level of high school program (0-3). Bootstrap standard errors in parentheses.

The functional form assumption on the fixed costs is the same as shown in the tables. I divide them by a common fixed cost of travel time such that they can be interpreted in daily minutes of travel time. The marginal costs are a nonlinear function of probabilities in the data and other parameters of the model (see Appendix D). For interpretation purposes only, I perform an OLS regression on the logarithmic transformation of the estimated marginal costs with the same structure as the fixed costs.

Male students have a 19-minute lower fixed cost to attend the benchmark vocational program, but the sign of this effect changes for the most academic programs. The marginal cost estimates reveal that they have a harder time obtaining good performance outcomes. For the same increase in effective study effort, a male student pays twice (exp(0.740)) the cost of a female student.

Higher cognitive ability leads to decreases in fixed costs, except for the benchmark vocational track. Marginal costs are strongly affected too. An increase of 1% of a standard deviation in language ability leads to a decrease of 0.6% in marginal costs in the vocational track (level=0) and 0.9 in the academic track (level=3).[36] The same increase in math ability leads to a decrease of 0.2% in the vocational track, but a much larger 1.3% decrease in the academic track.

Despite the controls for cognitive ability, the parental background still matters. High SES students are more favorable towards programs of higher academic level and marginal costs are lower. The magnitudes are similar to a standard deviation increase in language ability. There is also still a lot of persistent heterogeneity in the data that observable characteristics are not capturing. Appendix Table A17 shows that 30% of students belong to type 1 and 70% to type 2. Type 1 captures a group that experiences little trouble in completing high school in tracks of high academic level, compared to most students (type 2). One reason is that they have much lower fixed costs when they opt for more academic programs, equivalent to 86 minutes of daily

[36] I also estimated a model where academic level is proxied by the hours of academic courses (which varies over both tracks and grades) and obtain similar results.

Table 2: Costs of schooling: repeating and switching

	Fixed costs				**Log of marginal costs**			
	Baseline effect		Interaction with academic level		Baseline effect		Interaction with academic level	
Repeat	274.710	(33.156)	79.903	(13.121)	0.238	(0.194)	-0.461	(0.124)
Study delay	-9.932	(6.884)	10.026	(4.884)	0.637	(0.104)	0.109	(0.064)
Downgrade	169.088	(21.878)			0.174	(0.094)		
Upgrade	325.724	(41.557)			0.262	(0.179)		
Stay in clas	-16.062	(14.652)			0.155	(0.431)		
Stay in math	-203.545	(26.127)			0.653	(0.276)		

Note: Estimates of a sample of 5,158 students or 33,239 student-year observations. Scale = minutes of daily travel time. The marginal costs in the model are a flexible function of state variables, this table summarizes them by regressing their logarithmic transformation on the same variables that enter the fixed costs. Clas= classical languages included. Math= intensive math. Downgrade: switch to lower academic level. Upgrade: switch to higher academic level. Bootstrap standard errors in parentheses.

travel time for each step. Marginal costs are also much lower, making it easier to stay in academic programs. A type 1 student pays only 4% to 12% of the marginal cost of a type 2 student.[37]

Table 2 shows the impact of track choices and grade retention during high school. We see that study delay, i.e. past grade retention, increases marginal costs. This could be a result of demotivation. The same increases in effective study effort might be perceived as more costly for students with study delay because they lose interest in studying. On the contrary, we find decreases in marginal costs when students are repeating a grade in programs of high academic level. This can result from the fact that students see the same course material for a second time, making it easier to succeed. At the same time, the fixed cost estimates show that students strongly dislike repeating a grade. This shows a clear trade-off: students dislike repeating a grade, but it does help them to perform well in more academic programs. Finally, students do not like to switch programs. Both down- and upgrading is associated with much higher fixed costs, indicating a preference of students to stay in the same program.

[37] In the benchmark (=vocational) track this is $\exp(-3.265) = 4\%$, in the academic track it is $\exp(-(3.265 + 3 \times (-0.368))) = 12\%$.

5.2 Higher education

All estimates for higher education outcomes can be found in Appendix Tables A21, A22, and A23. We see that student characteristics that were important in explaining the costs of schooling also have a direct effect on college enrollment and graduation. This is also the case for the unobserved type of students, showing that it is important to control for unobserved heterogeneity when assessing the causal impact of study programs. Because the estimates of high school programs on higher education outcomes are difficult to interpret, I calculate the total Average Treatment effects on the Treated (ATT) of each study program and compare this to a comparison of the raw means in the data in Table 3. The estimate is not the results of a counterfactual

Table 3: Higher education and high school outcomes: difference in means and ATTs

		Enrollment				**Degree**		
		Mean diff		ATT		Mean diff		ATT	
Study program									
Academic									
	clas+math	5.1	(1.1)	1.7	(0.3)	20.1	(2.3)	8.2	(1.9)
	clas	5.2	(0.9)	1.2	(0.2)	16.4	(2.6)	5.4	(2.1)
	math	3.7	(1.0)	2.6	(0.3)	14.0	(2.2)	9.6	(1.8)
	other	*benchmark*				*benchmark*			
Middle-Theoretical									
	math	5.1	(1.1)	4.0	(1.5)	-1.3	(4.2)	5.8	(3.9)
	other	-15.2	(1.8)	-5.9	(2.1)	-26.2	(2.6)	-9.5	(2.8)
Middle-Practical		-39.3	(2.2)	-26.4	(2.8)	-46.6	(2.6)	-20.1	(3.2)
Vocational		-80.7	(1.7)	-64.6	(3.3)	-71.5	(1.9)	-37.2	(3.4)
One year of study delay		-26.0	(1.7)	-4.9	(1.1)	-33.9	(1.4)	-12.3	(1.3)
Data		58.2				44.0			

Note: Effects on enrollment and degree completion after graduating from different high school programs, compared to graduating from the academic track without clas or math option, and the effects of one year of study delay, compared to 0. Average treatment effects on the treated (ATT) make use of the causal estimates of enrollment and graduation equations. ATTs are calculated using indexes, specified in Appendix D, for each individual at the realization of other variables. Effects on obtaining higher education degree are total effects, i.e. they also take into account effects through enrollment. Clas= classical languages included. Math= intensive math. Bootstrap standard errors in parentheses.

simulation of the entire model but it is a "ceteris paribus" causal effect, i.e. it is the effect of one variable if all other variables that were realized at the time of leaving secondary education are kept fixed.[38] Similarly, I calculate the effect of one year of study delay by comparing outcomes for retained students in the counterfactual scenario where they would not have accumulated study delay.

Most estimates point in the same direction as a simple comparison of means in the data, but to a much smaller extent. I find that graduating from the academic track (without classical languages or extra math) leads to an increase in college graduation of 20 %points compared to the middle-practical track. The other track that prepares for higher education, the middle-theoretical track, leads to an increase of 9.5 %points. The estimates show that elective courses mainly matter for the type of higher education but we also see that overall graduation rates are higher when students had classical languages or intensive math in their program.

For study delay, I find a negative impact on higher education enrollment of 5 %points and an even stronger negative impact of 12 %points on obtaining a higher education degree.

6 Counterfactual tracking policies

In the current tracking policy in Flanders, teachers decide if a student has acquired the necessary skills to transition to the next grade in each of the programs. In some cases, they have not acquired the skills to transition to the next grade, regardless

[38]The ATTs are calculated as follows:

$$ATT^{j'} = E_{x,\nu}\left[P_j^{HE}(x_{it_{HE}}(j'),\nu_i) - P_j^{HE}(x_{it_{HE}}(j^0),\nu_i)|d_{iT_i^{SE}} = j'\right] \text{ for } HE = \{HEE, HED\}$$

with $E_{x,\nu}$ an expectations operator over the empirical distribution of the observables x and the estimated distribution of the unobserved types ν. P_j^{HE} is the probability of the higher education outcome (enrollment or graduation) as a function of the state variables. $x_{it_{HE}}(j')$ is the observed state vector of student i in the data at the time the outcome is realized $t = t_{HE}$ and $x_{it_{HE}}(j^0)$ is the same vector but with the graduation track replaced by an arbitrary benchmark program j^0. The ATT then calculates the average effect on HE of graduating high school in j' instead of j^0 for the group of students who graduated from j' in the data.

of their program choice. They then obtain a C-certificate which means they have to repeat the grade. However, in many cases, they are allowed to transition to the next grade but have to switch to a program of lower academic level or drop an elective course. In this case, they obtain a B-certificate. This allows underperforming students to avoid grade retention, but they can still opt for the same program if they are willing to repeat the grade. I compare the current policy to two alternatives:

Counterfactual 1: Repeat
Students are forced to repeat a grade when they obtain a B-certificate. This removes the option to avoid grade retention by switching to a different program if they underperformed this year. It makes the system less flexible and allows us to quantify the importance of the current flexibility.

Counterfactual 2: Downgrade
Students are forced to switch to a different program when they obtain a B-certificate, without repeating the grade. It resembles a policy that will be implemented in Flanders to reduce grade retention and allows us to quantify the importance of offering students the possibility to ignore the advice of teachers.

I first discuss the predicted effect of each policy using the proposed model. I then compare the results of the policy simulations to the impact of initial conditions. Finally, I compare my results to those of a pure discrete choice model. Details about the calculation of welfare effects can be found in the appendix section E.4.

6.1 Policy impact

Table 4 compares the outcomes of the two counterfactuals to the status quo scenario. The "Repeat" policy only shows worse outcomes. It does not manage to significantly increase graduation rates from the academic track. Instead, dropout rates increase by

Table 4: Counterfactual tracking policy

	Status quo	**Policy change B-certificate**			
		Repeat		Downgrade	
Panel A: educational outcomes	*%*	*Change in %points*			
High school					
Academic track	40.02	0.17	(0.30)	-1.13	(0.32)
Middle-theoretical track	16.10	-0.96	(0.29)	-1.52	(0.26)
Middle-practical track	8.14	-1.13	(0.24)	-0.77	(0.26)
Vocational track	21.57	-2.01	(0.32)	5.03	(0.29)
Dropout	14.17	3.94	(0.33)	-1.61	(0.25)
At least 1 B-certificate	37.53	-10.23	(0.71)	-3.49	(0.34)
At least 1 C-certificate	30.69	-0.23	(0.33)	-2.15	(0.24)
At least 1 year of study delay	33.22	9.48	(0.57)	-9.82	(0.55)
Higher education					
Enrollment	58.15	-1.76	(0.24)	-1.40	(0.21)
Graduation	44.25	-1.70	(0.22)	-0.30	(0.18)
Panel B: student welfare		*Change in $1000*			
Total student welfare		-2.14	(0.26)	-1.02	(0.14)
Fixed costs (-)		0.85	(0.12)	-0.48	(0.10)
Variable costs (-)		0.49	(0.08)	0.21	(0.03)
Expected payoff after high school (+)		-0.65	(0.10)	0.32	(0.07)
Taste shocks (+)		-0.15	(0.10)	-1.61	(0.19)

Note: Predictions from the dynamic model. C-certificate: repeat grade. B-certificate = students acquired skills to proceed to next grade but only if they downgrade, i.e. switch to track of lower academic level or drop elective course. Status quo = students can choose to downgrade or repeat grade after obtaining B-certificate, Repeat = students must repeat grade after obtaining B-certificate, Downgrade = students must downgrade and not repeat grade after obtaining B-certificate. Bootstrap standard errors in parentheses.

4 %points, which is a 28% increase in the total number. Not surprisingly, the share of students with grade retention increases by a large amount (9 %points). As a result, enrollment and graduation rates in higher education decrease by 2 %points. I also look at the welfare effects from the student's perspective. I assume an opportunity cost of \$10/hour and interpret it in dollars. Student welfare decreases on average by $\$2,140$, which is mainly driven by the increase in fixed costs and can be explained by the cost of repeating grades. The increase in dropout also decreases the expected payoff after leaving high school because these students can no longer enroll in college. The increase in grade retention and the decrease in college graduates are also expected

to have large negative externalities that are not considered in this exercise. I conclude that the current flexibility in the tracking policy is better than a strict pass or fail policy.

In the "Downgrade" policy, students who obtained a B-certificate are no longer allowed to repeat the grade. This would lead to a decrease in grade retention rates by 10 %points and dropout rates by 1.6 %points. This does come at a cost in the short run. Students switch to programs of lower academic level, which decreases enrollment rates in higher education by 1.4 %points. However, graduation rates only decrease by an insignificant 0.3 %points (or less than 1% of the total number), which can be explained by the strong effect of study delay on graduation. Since the policy restricts the choice set of students, their welfare unambiguously goes down. On average they lose $1,020, despite a reduction in the fixed costs of $480 and an increase in their expected payoff after high school of $320.[39] This is partly because they increase their study effort in response to the policy, leading to a loss of $210 due to variable effort costs, but the main cost of the policy comes from the reduction in their choice set that makes them miss out on $1,610, coming from unobserved taste shocks.

Despite the negative impact on student welfare, it can be argued that the "Downgrade" policy is beneficial for society. First, OECD estimates show that the per capita loss of $1,020 is close to the government saving by providing financing fewer years of schooling.[40] On top of that, gains of a year in tax payments would bring in an additional $1,960. The decrease in drop out and the improved efficiency of higher ed-

[39] Since we close the model after high school, the utility of enrolling in college is the students' expected lifetime utility at the time they leave high school. These expectations can be biased. By simultaneously predicting higher education graduation, we can see that the negative impact of study delay is much larger for graduation than for enrollment (see Table 3). Similarly, the counterfactual impacts the college enrollment rate but it does not have a significant impact on the number of college graduates. If students ultimately care about graduation, rather than enrollment, it suggests that they might underestimate the negative consequences of study delay in the long run. It is therefore possible that the increase in the expected payoff is substantially smaller than the increase in the actual payoff.

[40] The OECD estimates the direct cost to the education system of a student who repeats a grade to be $9,713. Decreasing grade retention rates by 9.82 %points then generate a government saving of $950 per student.

ucation (which is 90% government-funded (OECD, 2012)) would generate even larger returns.

I conclude that the current tracking policy is a good way to guide students in their track choices, rather than having them repeat a grade if they fail. Nevertheless, the choice they currently have to repeat a grade instead of downgrading leads to costly increases in grade retention and dropout, without an impact on graduation from higher education.

6.2 The importance of initial conditions

Although the "Downgrade" policy can be beneficial from a social perspective, many students would be hurt by the policy change. Therefore, I compare the effect of the policy to the impact of initial conditions as savings in high school could be invested to improve initial conditions directly (Gigliotti and Sorensen, 2018; Lafortune et al., 2018). I regress four predicted outcomes on student characteristics (in each policy scenario). I look at the impact on obtaining study delay, dropping out of high school, obtaining a higher education degree, and student welfare. The results can be found in the appendix Table A24.

While the effect of policies do not depend a lot on student characteristics, the impact of initial conditions on outcomes is large. A standard deviation decrease in language ability makes a student 7 %points less likely to obtain study delay, 7 %points less likely to drop out, 15 %points more likely to graduate from college and derive $16,270 more from the current high school system.[41] The impact of math ability, SES, or being female is identical in sign and similar in magnitude on most outcomes. Math ability does have a smaller effect on study delay and gender is more important to explain dropout and less important for higher education. Differences between the

[41]Note that for welfare, only differences are identified. Therefore, the constant is a nuisance parameter and the effect of student background should be interpreted as the effect on student welfare, keeping the utility of the outside good (not attending high school) fixed. This way, it captures changes in fixed and variable effort costs, taste shocks and the expected payoff of leaving high school, but not any effect it might have on the dropout utility.

two unobserved types are about twice the size of any single observable characteristic.

The results put into perspective the impact of the policy change in high school on student outcomes. The "Downgrade" policy can achieve large effects in terms of reducing study delay, resulting in important savings on educational spending. The impact on other outcomes is only marginal compared to the impact of initial conditions. This highlights the importance of preparing students before they enter high school, rather than having them exert additional effort (studying harder or repeating grades in rigorous programs) during high school. This is consistent with the literature on gains from early childhood education through dynamic complementarities (Cunha and Heckman, 2009; Cunha et al., 2010; Heckman and Mosso, 2014). High schools should aim at efficiently fostering the skills that students acquired before entering and refrain from letting students repeat grades to graduate from more academic programs. The savings that result from this could be used to improve these initial conditions directly.[42]

6.3 Bias in the pure discrete choice model

To demonstrate the methodological contribution of this paper, Table 5 compares the counterfactual predictions with those from a pure discrete choice model (i.e. a model in which the distribution of performance is exogenous and not the results of a study effort decisions).

A model in which students cannot adjust their study effort leads to less favorable outcomes in both counterfactuals. For example, the increase in study delay in the "Repeat" policy is 11.6 %points instead of 9.5. The decrease in the "Downgrade" policy is 9.2 %points instead of 9.8. Also in higher education, we see more negative effects if study effort is ignored. Most importantly, we would falsely conclude that there is an important negative impact on higher education graduation from the "Downgrade" policy (−1.1 %points), while the proposed model only estimates an

[42]See Appendix section E.4 for back-of-the-envelope calculations of the expected effects.

Table 5: Bias in the pure discrete choice model

	Policy change B-certificate							
	Repeat				Downgrade			
Effective study effort as a choice variable	Yes	No	Bias		Yes	No	Bias	
Panel A: educational outcomes			*Change in %points*					
High school								
Academic track	0.17	-0.61	-0.78	(0.19)	-1.13	-1.88	-0.75	(0.18)
Middle-theoretical track	-0.96	-1.32	-0.35	(0.19)	-1.52	-1.88	-0.36	(0.17)
Middle-practical track	-1.13	-1.73	-0.60	(0.17)	-0.77	-1.00	-0.23	(0.16)
Vocational track	-2.01	-1.04	0.96	(0.24)	5.03	6.21	1.19	(0.20)
Dropout	3.94	4.70	0.77	(0.20)	-1.61	-1.46	0.15	(0.12)
At least 1 B-certificate	-10.23	-6.32	3.91	(0.44)	-3.49	-0.86	2.63	(0.29)
At least 1 C-certificate	-0.23	0.22	0.44	(0.20)	-2.15	-1.69	0.45	(0.15)
At least 1 year of study delay	9.48	11.61	2.13	(0.36)	-9.82	-9.19	0.63	(0.27)
Higher education								
Enrollment	-1.76	-3.02	-1.27	(0.17)	-1.40	-2.27	-0.87	(0.11)
Graduation	-1.70	-2.69	-0.99	(0.15)	-0.30	-1.12	-0.81	(0.09)
Panel B: student welfare			*Change in $1000*					
Total student welfare	-2.14	-2.23	-0.09	(0.09)	-1.02	-0.99	0.03	(0.05)
Fixed costs (-)	0.85	1.26	0.41	(0.12)	-0.48	-0.63	-0.14	(0.05)
Variable costs (-)	0.49	0.00	-0.49	(0.08)	0.21	0.00	-0.21	(0.03)
Expected payoff after high school (+)	-0.65	-0.95	-0.30	(0.06)	0.32	0.14	-0.17	(0.04)
Taste shocks (+)	-0.15	-0.02	0.12	(0.06)	-1.61	-1.76	-0.14	(0.07)

Note: Predictions of two dynamic models. In a pure discrete choice model students cannot adjust study effort. In the proposed model they can because they choose the distribution of performance through their choice of effective study effort. Changes are with respect to the status quo prediction of each model. C-certificate: repeat grade. B-certificate = students acquired skills to proceed to next grade but only if they downgrade, i.e. switch to track of lower academic level or drop elective course. Status quo = students can choose to downgrade or repeat grade after obtaining B-certificate, Repeat = students must repeat grade after obtaining B-certificate, Downgrade = students must downgrade and not repeat grade after obtaining B-certificate. Opportunity cost of time: $10/h. Bootstrap standard errors in parentheses.

insignificant and small effect of -0.3.

The difference in results can be explained by an increase in study effort. Both counterfactuals make it less favorable to obtain a B-certificate. In a dynamic model with program choice, students can avoid this by choosing a program in which their success rate is higher. In the proposed model they could also change the success rate itself. Although it will be costly to do so, for many students this might be a better option than to switch programs in advance or take the risk to fail. This extra incentive to exert study effort has important implications. Although *actual* study effort remains unobserved, we can see that students adjust it by their willingness to pay the extra cost to have better performance outcomes through an increase in *effective* study effort. This is most clear from the decrease in the number of bad performance outcomes, and especially B-certificates, in both counterfactuals. In both models, the number of B-certificates goes down but the decrease in the pure discrete choice model is much smaller. In the "Repeat" policy, the decrease is 62% of the decrease in a model where students can adjust their study effort. In the "Downgrade" policy it is only 25%. This has important implications. First, there is a smaller increase in study delay in the "Repeat" policy and a stronger decrease in the "Downgrade" policy. Second, more students are staying in more academic programs. This increase in study effort comes at a cost. Both counterfactuals increase the variable costs (a component missing in a pure discrete choice model), but they do this at the benefit of other components of welfare such that the loss in total student welfare does not change significantly. The more favorable higher education outcomes compared to a pure discrete choice model are a result of the decrease in dropout, the increase in students graduating from academic programs, and the decrease in study delay.

7 Conclusion

I estimated a dynamic model of effort choice in secondary education in which students choose the academic level of the study program, as well as the distribution of their performance. I find that policies that encourage students who underperform to opt for programs of lower academic level do not have a negative effect on obtaining a higher education degree and they significantly decrease grade retention and high school dropout. This creates large savings for society that can be reinvested in early childhood education to improve educational outcomes.

The institutional context makes it possible to do clear counterfactuals to investigate the trade-off between costs of academic effort (study effort, grade retention, risk of dropout) and the benefits in the long run (higher education degree). These conclusions are also important for other countries that track students from an early age like Germany, Austria, or the Netherlands. Also in more comprehensive educational systems like the US, we find a similar trade-off at the course-level. Students often retake failed classes to graduate from high school or to increase their chances to be admitted to college. Many colleges explicitly ask for a high GPA and a rigorous academic curriculum in their admission criteria. Students, especially those of lower ability, then face a similar trade-off between studying advanced courses at the risk of retakes and a lower GPA or choosing a curriculum with less advanced courses. The results in this paper suggest these increases in effort during high school can be very costly for students and society, and might not improve future outcomes.

From a methodological perspective, I show that it is possible to allow students to exert different amounts of study effort in counterfactual simulations, despite the lack of data on study effort. I also show that this is important in the application of this paper. Further research can apply the modeling strategy to other contexts. Any model where agents are expected to have some, but imperfect, control over state transitions can benefit from this approach and the data requirements are the same as for a model in which state transition probabilities are not determined within the model.

Future research could also combine this approach with recent extensions of the pure discrete choice model along other dimensions, introducing additional uncertainty about the performance distribution due to imperfect information of students about their ability (Arcidiacono et al., 2016) and endogenous quality of schools or programs due to the quality of peers and effort choices of teachers (Fu and Mehta, 2018). This would allow for counterfactuals that change the educational system more substantially, such as changing the age in which students are first tracked, rather than only constraining choices of students within the current system.

A final area of further research is to empirically test the performance of the model.

I argue that the pure discrete choice model should be adapted because it is unreasonable to assume that counterfactual simulations will not have an impact on study effort. However, it is not testable that students can indeed choose their study effort because the pure discrete choice model is observationally equivalent. This is similar to arguing for the use of a dynamic over a static model. Despite being observationally equivalent (Manski, 1993), dynamic models are often preferred because their assumptions are more reasonable in several contexts and this could impact counterfactual simulations. Similar to the literature on identifying dynamic behavior (Abbring and Daljord, 2020; Magnac and Thesmar, 2002), future research could investigate how exclusion restrictions and policy variation in the data could be exploited to test the pure discrete choice model against one where agents also choose the distribution of state transtions.

References

Abbring, J. H. (2010). Identification of Dynamic Discrete Choice Models. *Annual Review of Economics.*

Abbring, J. H. and Daljord, Ø. (2020). Identifying the discount factor in dynamic discrete choice models. *Quantitative Economics*, 11(2):471–501.

Altonji, J. G. (1995). The Effects of High School Curriculum on Education and Labor Market Outcomes. *The Journal of Human Resources*, 30(3):409.

Altonji, J. G., Blom, E., and Meghir, C. (2012). Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers. *Annual Review of Economics*, 4(1):185–223.

Arcidiacono, P. (2004). Ability sorting and the returns to college major. *Journal of Econometrics*, 121(1-2):343–375.

Arcidiacono, P. (2005). Affirmative action in higher education: How do admission and financial aid rules affect future earnings? *Econometrica*, 73(5):1477–1524.

Arcidiacono, P., Aucejo, E., Maurel, A., and Ransom, T. (2016). College Attrition and the Dynamics of Information Revelation. *Working paper.*

Arcidiacono, P. and Ellickson, P. B. (2011). Practical Methods for Estimation of Dynamic Discrete Choice Models. *Annual Review of Economics*, 3:363–394.

Arcidiacono, P. and Miller, R. A. (2011). Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity. *Econometrica*, 79(6):1823–1867.

Aughinbaugh, A. (2012). The effects of high school math curriculum on college attendance: Evidence from the NLSY97. *Economics of Education Review*, 31(6):861–870.

Beffy, M., Fougere, D., and Maurel, A. (2012). Choosing the field of study in post-secondary education: Do expected earnings matter? *Review of Economics and Statistics*, 94(1):334–347.

Carneiro, P., Hansen, K. T., and Heckman, J. J. (2003). Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College. *International Economic Review*, 44(2):361–422.

Cockx, B., Dejemeppe, M., Launov, A., and Van der Linden, B. (2018a). Imperfect monitoring of job search: Structural estimation and policy design. *Journal of Labor Economics*, 36(1).

Cockx, B., Picchio, M., and Baert, S. (2018b). Modeling the effects of grade retention in high school. *Journal of Applied Econometrics*.

Costrell, R. M. (1994). A Simple Model of Educational Standards. *American Economic Review*, 84(4):956–971.

Cummins, J. R. (2017). Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment. *Economics of Education Review*, 56:40–51.

Cunha, F. and Heckman, J. J. (2009). The Economics and Psychology of Inequality and Human Development. *Journal of the European Economic Association*, 7(2/3):320–364.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78(3):883–931.

De Groote, O. and Declercq, K. (2020). Tracking and specialization of high schools: heterogeneous effects of school choice. *Working paper*.

Declercq, K. and Verboven, F. (2015). Socio-economic status and enrollment in higher education: do costs matter? *Education Economics*, 23(5):532–556.

Declercq, K. and Verboven, F. (2018). Enrollment and degree completion in higher education without admission standards. *Economics of Education Review*, 66:223–244.

Dubois, P., de Janvry, A., and Sadoulet, E. (2012). Effects on School Enrollment and Performance of a Conditional Cash Transfer Program in Mexico. *Journal of Labor Economics*, 30(3):555–589.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5):1739–1774.

Dustmann, C., Puhani, P. A., and Schoenberg, U. (2017). The Long-term Effects of Early Track Choice. *The Economic Journal*, 127(603):1348–1380.

Eckstein, Z. and Wolpin, K. I. (1999). Why youths drop out of high school: The impact of preferences, opportunities, and abilities. *Econometrica*, 67(6):1295–1339.

French, E. and Taber, C. (2011). Identification of Models of the Labor Market. In *Handbook of Labor Economics*, volume 4, pages 537–617. Elsevier.

Freyberger, J. (2018). Non-parametric Panel Data Models with Interactive Fixed Effects. *The Review of Economic Studies*, 85(3):1824–1851.

Fruehwirth, J. C., Navarro, S., and Takahashi, Y. (2016). How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects. *Journal of Labor Economics*, 34(4).

Fu, C. and Mehta, N. (2018). Ability Tracking, School and Parental Effort, and Student Achievement: A Structural Model and Estimation. *Journal of Labor Economics*, 36(4).

Garibaldi, P., Giavazzi, F., Ichino, A., and Rettore, E. (2012). College cost and time to complete a degree: Evidence from tuition discontinuities. *Review of Economics and Statistics*, 94(3):699–711.

Gigliotti, P. and Sorensen, L. C. (2018). Educational resources and student achievement: Evidence from the Save Harmless provision in New York State. *Economics of Education Review*, 66:167–182.

Giustinelli, P. (2016). Group Decision Making with Uncertain Outcomes: Unpacking Child-Parent Choice of The High School Track. *International Economic Review*, 57(2):573–601.

Goodman, J. (2019). The Labor of Division: Returns to Compulsory High School Math Coursework. *Journal of Labor Economics*, 37(4).

Guyon, N., Maurin, E., and McNally, S. (2012). The Effect of Tracking Students by Ability into Different Schools: A Natural Experiment. *Journal of Human Resources*, 47(3):684–721.

Hanushek, E. A. and Woessmann, L. (2006). Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries. *The Economic Journal*, 116(510):C63–C76.

Heckman, J. J., Humphries, J. E., and Veramendi, G. (2016). Dynamic treatment effects. *Journal of Econometrics*, 191(2):276–292.

Heckman, J. J. and Mosso, S. (2014). The Economics of Human Development and Social Mobility. *Annual Review of Economics*, 6(1):689–733.

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

Hotz, V. J. and Miller, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529.

Hu, Y. and Shum, M. (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171(1):32–44.

Hu, Y. and Xin, Y. (2019). Identification and estimation of dynamic structural models with unobserved choices. *Cemmap working paper*, CWP 35/19.

Jacob, B. A. and Lefgren, L. (2009). The Effect of Grade Retention on High School Completion. *American Economic Journal: Applied Economics*, 1(3):33–58.

Joensen, J. S. and Mattana, E. (2017). Student Aid, Academic Achievement, and Labor Market Behavior. *Working paper*.

Joensen, J. S. and Nielsen, H. S. (2009). Is there a causal effect of high school math on labor market outcomes? *Journal of Human Resources*, 44(1):171–198.

Kalouptsidi, M., Scott, P. T., and Souza-Rodrigues, E. (2018). Identification of Counterfactuals in Dynamic Discrete Choice Models.

Kapor, A., Neilson, C. A., and Zimmerman, S. (2018). Heterogeneous Beliefs and School Choice Mechanisms. *Working paper*, page 56.

Kasahara, H. and Shimotsu, K. (2009). Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices. *Econometrica*, 77(1):135–175.

Keane, M. P. and Wolpin, K. I. (1997). The Career Decisions of Young Men. *Journal of Political Economy*, 105(3):473–522.

Kreisman, D. and Stange, K. (2017). Vocational and Career Tech Education in American High Schools: The Value of Depth Over Breadth. *Education Finance and Policy*, Forthcoming.

Lafortune, J., Rothstein, J., and Schanzenbach, D. W. (2018). School Finance Reform and the Distribution of Student Achievement. *American Economic Journal: Applied Economics*, 10(2):1–26.

Lin, D. (2019). Multi-Dimensional Abilities, Task Content of Occupations and Career Choices: A Dynamic Analysis. *Working paper.*

Magnac, T. and Thesmar, D. (2002). Identifying dynamic discrete decision processes. *Econometrica*, 70(2):801–816.

Manacorda, M. (2012). The Cost of Grade Retention. *Review of Economics and Statistics*, 94(2):596–606.

Manski, C. F. (1993). Dynamic choice in social settings: Learning from the experiences of others. *Journal of Econometrics*, 58(1):121–136.

Meer, J. (2007). Evidence on the returns to secondary vocational education. *Economics of Education Review*, 26:559–573.

Nord, C., Roey, S., Perkins, R., Lyons, M., Lemanski, N., Brown, J., and Schuknecht, J. (2011). The Nation's Report Card [TM]: America's High School Graduates. Technical Report NCES 2011-462, US Department of Education, National Center for Education Statistics, Washington, DC.

OECD (2012). *Education at a Glance 2012.* OECD Publishing. OCLC: 812608795.

OECD (2013). *PISA 2012 Results: What Makes Schools Successful (Volume IV).* PISA. OECD Publishing.

OECD (2017). *Education at a Glance 2017.* Education at a Glance. OECD Publishing.

Paserman, M. D. (2008). Job search and hyperbolic discounting: Structural estimation and policy evaluation. *The Economic Journal*, 118(531):1418–1452.

Pekkarinen, T., Uusitalo, R., and Kerr, S. (2009). School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics*, 93(7-8):965–973.

Roller, M. and Steinberg, D. (2020). The distributional effects of early school stratification - non-parametric evidence from Germany. *European Economic Review*, 125.

Rose, H. and Betts, J. R. (2004). The effect of high school courses on earnings. *The Review of Economics and Statistics*, 86(2):497–513.

Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica*, 55(5):999–1033.

Shure, N. (2017). Non-cognitive peer effects in secondary education. *Unpublished manuscript*.

Todd, P. E. and Wolpin, K. I. (2018). Accounting for Mathematics Performance of High School Students in Mexico: Estimating a Coordination Game in the Classroom. *Journal of Political Economy*, (forthcoming).

van den Berg, G. J. and van der Klaauw, B. (2019). Structural Empirical Evaluation of Job Search Monitoring. *International Economic Review*, 60(2):879–903.

A Identification

A.1 Identified objects in pure discrete choice models

Consider the standard set-up of a dynamic discrete choice model. Each period t, agent i chooses an option j. The decision is based on observed characteristics x_{it}, an unobserved type ν_i and unobserved iid taste shocks $\varepsilon_{it} = \{\varepsilon_{i1t}, \varepsilon_{i2t}, ...\}$. The time horizon can be infinite or (if x_{it} includes t) finite. Each period agent i derives some flow utility

$$u_j(x_{it}, \nu_i) + \varepsilon_{ijt}$$

and states transition according to a process that satisfies conditional independence (Rust, 1987):

$$f_j(x_{it+1}|x_{it}, \nu_i) = f_j(x_{it+1}|x_{it}, \nu_i, \varepsilon_{ijt}).$$

Agents maximize expected lifetime utility by choosing the option with the highest conditional value function:

$$v_j(x_{it}, \nu_i) + \varepsilon_{ijt} = u_j(x_{it}, \nu_i) + \beta \int \bar{V}(x_{it+1}, \nu_i) f_j(x_{it+1}|x_{it}, \nu_i) dx_{it+1} + \varepsilon_{ijt}$$

with $\bar{V}(x_{it+1}, \nu_i)$ the expected value of behaving optimally after integrating over the taste shocks.

In the case where there is no unobserved type, Magnac and Thesmar (2002) show that data on x_{it} and the chosen option, identify $u_j(x_{it})$ after specifying the utility of a reference alternative, the discount factor β and the distribution of ε_{ijt}. State transitions $f_j(x_{it+1}|x_{it})$ are nonparametrically identified. We can then use $u_j(.)$ and $f_j(.)$ for counterfactual simulations by assuming they are primitives and therefore invariant to policy changes.

As the iid assumption on unobserved heterogeneity is restrictive, many applications would add an unobserved state ν_i to capture persistent unobserved heterogeneity and identify $u_j(x_{it}, \nu_i)$ and $f_j(x_{it+1}|x_{it}, \nu_i)$. I will allow for this in the rest of this section.[43]

A.2 Identification of alternative primitives

To relax the assumption of policy-invariance, assume instead that the functions $u_j(.)$ and $f_j(.)$ depend on choice behavior and what we identify are therefore the endogenously determined objects $u_j^*(x_{it}, \nu_i)$ and $f_j^*(x_{it+1}|x_{it}, \nu_i)$. The goal is to derive other primitives from the data that are more likely to be policy-invariant.

Assume agents can choose the distribution of state transitions through a single index y_{it} such that $\phi_{j,\widetilde{x},\widetilde{x}'}(y_{it})$ is the probability for i in state $(x_{it}, \nu_i) = \tilde{x}$ to transition to state $\widetilde{x}' = (x_{it+1}, \nu_i)$ after choosing j. The optimal choice of y_{it} in a given program j and state (x_{it}, ν_i) is then given by $y_j^*(x_{it}, \nu_i) = \phi^{-1}_{j,\widetilde{x},\widetilde{x}'}(f_j^*(x_{it+1}|x_{it}, \nu_i))$.[44] We now let this index linearly enter the utility function:

$$u_j(x_{it}, \nu_i, y_{it}) = u_j^0(x_{it}, \nu_i) + u_j^y(x_{it}, \nu_i) y_{it} \tag{20}$$

with $u_j^0(x_{it}, \nu_i)$ a component that is independent of the choice of the index and $u_j^y(x_{it}, \nu_i) \equiv \frac{du_j(x_{it}, \nu_i, y_{it})}{dy_{it}}$ the marginal flow utility from changing y_{it}.

To connect what we observe in the data with the current model, we make the following assumption:

Assumption: In the data, agents in option j choose y_{it} to maximize expected lifetime utility and obtain an interior solution $y_j^*(x_{it}, \nu_i) = \phi^{-1}_{j,\widetilde{x},\widetilde{x}'}(f_j^*(x_{it+1}|x_{it}, \nu_i))$ with $\phi_{j,\widetilde{x},\widetilde{x}'}$ a known function that is invertible and differentiable in y_{it}.

[43] There are several approaches to identify type-specific $u_j(x_{it}, \nu_i)$ and $f_j(x_{it+1}|x_{it}, \nu_i)$, see for example Magnac and Thesmar (2002); Kasahara and Shimotsu (2009); Hu and Shum (2012), and the discussion on identification of the application in this paper.

[44] For a simple case, assume $x_{it} = (x_{i0}, g_{it})$ with initial observed characteristic x_{i0} and a dummy for obtaining a degree g_{it}. Assume students choose the exponential of the index of a logit on obtaining a degree when they are in an option j that gives this possibility. We can then write $\phi_{j,(x_{i0},0),\nu_i,(x_{i0},1)}(\tilde{y}) = \frac{\tilde{y}}{1+\tilde{y}}$ and $\phi_{j,(x_{i0},0),\nu_i,(x_{i0},0)}(\tilde{y}) = 1 - \frac{\tilde{y}}{1+\tilde{y}}$.

In contrast, the pure discrete choice model recovers $f_j^*(x_{it+1}|x_{it}, \nu_i)$ while remaining agnostic about how it was determined. However, when proceeding to counterfactual simulations, it is not updated, i.e. it is implicitly assumed that agents cannot affect it.

As in the pure discrete choice model, we still assume that agents choose the option j that generates the highest expected lifetime utility. Conditional value functions now depend on y_{it}:

$$v_j(x_{it}, \nu_i, y_{it}) + \varepsilon_{ijt} = u_j(x_{it}, \nu_i, y_{it}) + \beta \int \bar{V}(x_{it+1}, \nu_i)\phi_{j,\widetilde{x},\widetilde{x}'}(y_{it})dx_{it+1} + \varepsilon_{ijt}.$$

Solving this for the optimal y_{it}, the following FOC has to be satisfied:

$$u_j^y(x_{it}, \nu_i) = -\beta \int \bar{V}(x_{it+1}, \nu_i)\frac{\partial \phi_{j,\widetilde{x},\widetilde{x}'}(y_{it})}{\partial y_{it}}dx_{it+1} \text{ for } y_{it} = y_j^*(x_{it}, \nu_i) \tag{21}$$

with the left-hand side equal to the marginal flow utility i receives today from increasing y_{it}, and the right-hand side the expected decrease in future utility. Since $\phi_{j,\widetilde{x},\widetilde{x}'}$ is invertible and differentiable, we can identify the optimal value of y_{it} in the data ($y_j^*(x_{it}, \nu_i)$) and calculate the derivative at this point. Identification of all flow utilities with optimal choices also implies the identification of $\bar{V}(x_{it+1}, \nu_i)$. β is taken as given. Therefore, we can identify $u_j^y(x_{it}, \nu_i)$ using this FOC. With $u_j^y(x_{it}, \nu_i)$, $y_j^*(x_{it}, \nu_i)$ and $u_j^*(x_{it}, \nu_i)$ identified, we can use (20) to identify $u_j^0(x_{it}, \nu_i)$.

We have now identified two new primitives of the model: the marginal utility of a change in the index of state transitions $u_j^y(x_{it}, \nu_i)$, and a component in the utility function that is independent of the distribution of state transitions $u_j^0(x_{it}, \nu_i)$. We can do this for different choices of $\phi_{j,\widetilde{x},\widetilde{x}'}(y_{it})$, giving some flexibility to researchers to choose the primitives of the model. Two aspects are important for this choice. First, by choosing $\phi_{j,\widetilde{x},\widetilde{x}'}$, the researcher effectively chooses for which transformation of state transitions the linearity assumption holds, i.e. for which transformation the marginal impact on flow utility can be considered a primitive of the model. Second, the choice of $\phi_{j,\tilde{x},\tilde{x}'}$ should be consistent with a high-level assumption that the FOC (21) is satisfied in the data (as an interior solution is required). If not, the FOC cannot provide the identifying power we need. Note that we only need this to identify primitives from the data. In counterfactual simulations, we can allow for corner solutions.

B Data appendix

B.1 The LOSO dataset

The dataset used for this paper is the LOSO dataset.[45] The first part of the data contains rich information of students and their parents, and choices and performance measures during high school in the region of Flanders (Belgium). We can follow a cohort of students starting high school in 1990. I also include results from follow-up research, called “LOSO-annex”, which looked into the education and labor market career in the first three years after leaving high school (academic years starting in 1996 until 1998 for most students, but later for those with study delay). This data was later enriched by sending questionnaires during 2003-2005 to students that were still in the educational system in the questionnaire before.

The students are not randomly selected over Flanders. Instead, two large subregions of Flanders were defined that are considered to be representative of the entire region.[46] In these regions, almost all schools are included, and within each school, every student is included. The first subregion is in the east part of Flanders and includes the municipalities Hasselt, Genk, Beringen, Leopoldsburg, Herk-de-Stad, and Diest. The second subregion is more to the west and contains the schools in Dender-

[45] See also https://ppw.kuleuven.be/o_en_o/COE/losodatabank.

[46] To test the representativeness of the data, I compared higher education enrollment number (58%) to population data. For Belgium as a whole, I find an almost identical number around the same time period: 56% in 1996 and 57% in 1999 (UNESCO Institute for Statistics, indicator SE.TER.ENRR).

monde, Hamme, and Zele. Data was collected from students, parents, teachers, and schools, and they were actively contacted by researchers on multiple occasions. This is why the data is of high quality and there is very little attrition. Even if a student decides to leave his school for a school that was not initially part of the project, it was still possible to collect the necessary information.

B.2 Sample selection

I only keep the 6,439 students in the dataset that are known as 'proefgroepleerlingen'. These are students that are tracked from the start of high school, even if they move to another school. The dataset also contains a large number of observations of inflow in schools over time but these are not used in this study. From these students, I eventually keep 5,158 students to estimate the model.

The model in this paper captures the main aspects of the education system but also makes some simplifications, implying that it cannot explain every observation in the data. Moreover, some data on the choices or outcomes that are needed for the estimation are missing. Table A1 summarizes the attrition. More details on why observations had to be dropped follow next.

B.3 Data interpretation

Some information in the data is not straightforward to use in the model. Therefore, I create or adjust some of the information to capture the spirit of the educational system with the model, without overly complicating it to capture all anomalies in the data. In particular, I perform the following manipulations.

First, students who are successful in the first grade of the vocational track can go to the first grade of another track. I do not allow for this possibility in the model. Instead, I make these students look as if they entered the non-vocational track after an additional year of study delay in elementary school. Second, B-certificates often exclude specific programs like technical education-science, or accountancy-informatics,

Table A1: Data attrition

	Total number of students	**Loss**	**Relative loss compared to start**	**Reason for dropping**
Data description	6439			
Data received (includes birth data and gender)	6411	-28	0.004	missing data
No students that leave and return to secondary education	6381	-30	0.009	not allowed by model
No time period	6365	-16	0.011	missing data
No performance	6327	-38	0.017	missing data
No study program	6302	-25	0.021	missing data
Do not allow switch from middle to vocational after grade 11	6263	-39	0.027	not allowed by model
Do not allow to skip grades in high school	6260	-3	0.028	not allowed by model
Go down grades in high school	6249	-11	0.030	not allowed by model
Students have to start in the first grade of high school	6246	-3	0.030	not allowed by model
Ignore students that skipped grade in elementary school	6212	-34	0.035	not allowed by model
Students make choices that are inconsistent with the certificate they received and/or track they were in	5936	-276	0.078	not allowed by model
Students that move from part time to full time education	5903	-33	0.083	not allowed by model
Students that drop out illegally	5832	-71	0.094	not allowed by model
No info on choice after leaving high school	5705	-127	0.114	missing data
No info on obtaining a higher education degree within 6 years	5558	-147	0.137	missing data
No info on location of student	5442	-116	0.155	missing data
Missing characteristics or test score of students	5179	-263	0.196	missing data
Do not allow 3 or more years of study delay at start high school	5177	-2	0.196	not allowed by model
Do not allow students to live more than 50 km to closest school of each type	5160	-17	0.199	remove outliers
Do not allow to go to higher education without high school degree	5158	-2	0.199	not allowed by model

and not always entire study programs as defined in the model. In many cases, only "unrealistic" alternatives remain within the same study program that I include in the model (e.g. a program that is not available in any school in the neighborhood). To avoid modeling every single study program, as well as school choice, I instead use a model with aggregated study programs and interpret the certificate data in a specific way.

Certificates that exclude an entire track are straightforward to implement. This already contains 67% of the data on B-certificates. In other cases, I proceed as follows. I always assume a hierarchy: if a low track is excluded, the higher ones are excluded too.[47] I also use a slightly different definition of a B-certificate that is more consistent over the different grades. I ignore the officially called "C-certificates" in grade 7 as they do not restrict entry into grade 8 of the vocational track, and change them to B-certificates that allow the vocational track in the next grade (or A-certificate if the student is already in the vocational track). In other cases in the academic and middle-theoretical track, I use the following procedure. This procedure was established to be in line as much as possible with the spirit of the educational system, as well as to minimize the number of choices in the data that would not be possible to be explained by the model. I make groups of aggregated study programs that are less aggregated than the ones used in the model, but more aggregated than how they appear in the data. This aggregates over very small differences within programs between which a B-certificate is not expected to ever make a distinction, except when teachers (and probably students) are not aware of the existence of the program. A B-certificate then excludes all classical language options if all the aggregated programs with classical languages appear in the list of restrictions. It excludes math options and the entire track if there is an exclusion within all the major aggregated options of these study programs. For exclusion of the middle-practical track, one occurrence of a program

[47]The following example shows that this is reasonable to assume: out of 199 B-certificates that exclude all programs in the middle tracks for students currently in an academic track, 197 certificates also exclude the academic track.

in the track in the list of restrictions restricts the entire track, unless choice behavior and the corresponding grade is not consistent with that.

At this point, we went from explaining 67% of the B-certificate data to explaining 95%. The remaining 5% is assumed to be imposing irrelevant restrictions on the students in the model and are replaced by A-certificates. An important part of this 5% also contains exclusions within the vocational track which are unrelated to the academic level of the program and are therefore outside the scope of this paper.

B.4 Details about study programs

The official distinction between tracks differs slightly from the one proposed in the paper. The official track names are "ASO", "TSO", "KSO", "BSO", and "BUSO" and the distinction for most tracks is made from the third year on (i.e. grade 9). ASO corresponds to the academic track, BSO and BUSO to the vocational track and both TSO and KSO are middle tracks (that differ in their focus on respectively technical education and artistic education). I then split up this middle track according to programs that prepare primarily for higher education (middle-theoretical) and the labor market (middle-practical), which is a common distinction made, e.g. in Cockx et al. (2018b), but also by the researchers that collected the data.[48]

Although this official distinction does not exist in the first two grades of high school, there is a distinction between programs preparing for the different tracks. First of all, there is the distinction between a B-stream, preparing for the vocational track only, and an A-stream, preparing for the other tracks. Within the A-stream one can also distinguish between more or less theoretical programs, based on the hours per week each school can decide what to teach (5 in grade 7 and up to 10 in grade 8). This distinction was made by the LOSO researchers, although not directly linked to

[48] The supply of programs differs between schools in Flanders. Some schools specialize and offer programs in only one track while other schools do not specialize and offer programs in all tracks. In the model I do not distinguish between different schools as they are all regulated in the same way and the restrictions implied by certificates also hold for other schools.

the specific track they prepare for. Therefore, I looked at the most common transition patterns to assign them to a track. In a few cases, the distinction within the A-stream was not made, I then assumed students were in the same track as the year after.

As mentioned in Cockx et al. (2018b), upward mobility is theoretically possible but practically infeasible which is why it rarely occurs in the data. Nevertheless, I do allow for this flexibility in non-vocational tracks in the first two grades as I do see some upward mobility when the official track structure is not yet established. Note that any mobility between grade 11 and grade 12 is forbidden, except for a switch between some programs from a middle track to the vocational track. I do not allow for that in the model and drop the students that do this. I also exclude the following uncommon choices in the model: dropping out of (full time) high school and returning, and repeating the grade in a track of higher academic level or with an elective course that was not chosen before. Furthermore, sometimes rules are not strictly followed. Some cases can be illegal, but in other cases, parents could have asked for special permission from teachers, the ministry of education, or as a result of a court order. These special cases are dropped.

For the higher education options, the distinction between different levels (professional college, academic college, university) is also used in official statistics on Flemish education and corresponds to respectively “Hoger onderwijs van het korte type”, “Hoger onderwijs van het lange type” and “Universiteit”. Today, the distinction between “Hoger onderwijs van het lange type” and “Universiteit” is no longer made but the study programs within them are still similar. To define STEM majors, I use a characterization by the Flemish government (https://www.onderwijskiezer.be/). The different types of (higher) education are associated with large differences in labor market outcomes. To demonstrate this, I use data of the "Vacature Salarisenquête", a large survey of workers in Flanders in 2006, and compare the median wages of 30-39-year-olds (sample size of 20,534 workers). High school dropouts earned a gross monthly wage of 2,039 EUR, high school graduates without a higher education degree earned 2,250 EUR, professional college graduates 2,600 EUR, academic college graduates 3,281 EUR and university graduates 3,490 EUR. Students that graduated in a STEM major earned 3,264 EUR, while students that graduated in a non-STEM major earned 2,800 EUR.

B.5 Distance and travel time data

I use address data of students and schools to obtain coordinates using the Stata command "geocode3". For the schools, I updated this manually when geocode returned an error or was not very precise. I did this for schools with at least 10 student-time observations using Google maps. I then use the "osrmtime" command to calculate travel time by bike to the closest school that offers the study program.[49] Note that all schools attended by students in the sample are used, which includes also schools outside of the ones assigned by the researchers (because students can switch to other schools). I dropped students living more than 50km from any school as they are more likely to be influenced by schools that I do not observe or are outliers because of measurement error when geocoding.

At the higher education level, I look at the distance to the closest school for each option (level and major) if it is not a university and I distinguish between the five Flemish campuses for universities (Leuven, Ghent, Brussels, Antwerp, and Diepenbeek). This is similar to Declercq and Verboven (2018). If students attend a university abroad or in Wallonia, I assign them randomly to one of the Flemish campuses, using a probability distribution that corresponds to the distribution of students going to Flemish universities.

[49] A bike is the most popular mode of transportation. According to government agency VSV, 36% of students use a bike, 30% the bus and 15% a car (source: http://www.vsv.be/sites/default/files/20120903_schoolstart_duurzaam.pdf). Since distance to school is small, travel time by bike is also a good proxy for other modes of transportation.

B.6 Policy relevance

Although similar issues arise in other educational systems, they are particularly important in the current context. Belgium spends 2.8% of its GDP on secondary education, the highest number among OECD countries. Therefore, it is crucial to study the effectiveness of the system in helping students to achieve their future goals in a cost-efficient way. Since 96% of the cost is paid by society, it is also important to see if students have the right incentives within the system to optimize total welfare (OECD, 2017). Belgium has a very high rate of grade retention in secondary education which comes at a large cost. The total cost of a year of study delay in Belgium amounts to at least $48,918/student or 11% of total expenditures on compulsory education, the highest rate in the OECD (OECD, 2013).

C Solution of the model

I assume it is no longer possible to go to secondary education in $T^{\max} = 10$ such that the model can be solved backward. Because of the extreme value assumption on the taste shocks ε_{ijt}, I can write the expected value of lifetime utility in the period where secondary education is no longer allowed, using the logsum formula:

$$\bar{V}(x_{it+1}, \nu_i) = \gamma + \ln \sum_{j \in \Phi(x_{it+1})} \exp(\text{Degree}'_{it+1}\mu^{\text{degree}} + \Psi_j^{HEE}(x_{it+1}, \nu_i)) \text{ if } t+1 = T^{\max}$$

with $\gamma \approx 0.577$ the Euler constant and $\Phi_{it+1} = \Phi(x_{it+1})$ the choice set. $\bar{V}$ is used as an input in t (see (17)). First, students look for the optimal value of the effective study effort in every possible option in secondary education: y^*_{ijt}. As explained in section 2, an interior solution in the data is required and the following FOC should then be satisfied:

$$c_j(x_{it}, \nu_i) = \beta \sum_{\bar{g}} \frac{\partial \phi^{\bar{g}}_{ijt}(y_{it})}{\partial y_{it}} \bar{V}(x_{it+1}(\bar{g}), \nu_i) \text{ if } y_{it} = y^*_{ijt}. \tag{22}$$

As in the simple model of section 2, a sufficient condition to obtain an interior solution is to assume that students always believe there is a positive probability to avoid the worse performance outcome in any program. This avoids that $y^*_{ijt} = 0$. Furthermore, a positive marginal cost makes sure that it is never optimal to exert an infinite level of study effort. The FOC condition equalizes marginal costs and (expected) marginal benefits. As this does not depend on taste shocks ε or performance shocks η, it implies that students with the same state vector (x_{it}, ν_i) will choose the same effort levels in a given program: $y^*_{ijt} = y^*_j(x_{it}, \nu_i)$. In contrast to the simple model in section 2, I do not obtain a closed-form solution for $y^*_j(x_{it}, \nu_i)$. However, I can still estimate the optimal levels in the data (see section D for details about estimation). In counterfactual simulations, I run a grid search to find the new optimum (see section E for details about the simulations).

When students know the optimal levels of effort in each program, they can choose the program with the highest value of $v_j(x_{it}, \nu_i, y^*_j(x_{it}, \nu_i)) + \varepsilon_{ijt}$. This results in the following logit choice probabilities:

$$\Pr(d_{it} = j | x_{it}, \nu_i) = \frac{\exp(v_j(x_{it}, \nu_i, y^*_j(x_{it}, \nu_i)))}{\sum_{j' \in \Phi(x_{it})} \exp(v_{j'}(x_{it}, \nu_i, y^*_{j'}(x_{it}, \nu_i)))} \tag{23}$$

with v_{ijt} given by (17) for options in secondary education and (19) for options after secondary education. $\bar{V}(x_{it}, \nu_i)$ can also be calculated using:

$$\bar{V}(x_{it}, \nu_i) = \gamma + \ln \sum_{j \in \Phi(x_{it})} \exp(v_j(x_{it}, \nu_i, y^*_j(x_{it}, \nu_i))).$$

These steps can be repeated until the first period to solve the entire model.

D Estimation details

I first explain the estimation of the model when the econometrician knows the type ν_i of every student and then allow them to be unobserved.

D.1 Higher education

I propose the following functional forms for higher education enrollment parameters ($\Psi_j^{HEE}(.)$) and graduation parameters ($\Psi_j^{HED}(.)$) and estimate them as parameters of a conditional logit, using maximum likelihood:

$$\begin{aligned}
\Psi_j^{HEE}(x_{it}, \nu_i) = \varphi_j^{HEE,0} & \\
&+ S_i'(\varphi^{HEE,S,0} + \varphi^{HEE,S,level}\text{level_HE}_j + \varphi^{HEE,S,STEM}\text{STEM}_j) \\
&+ \nu_i'(\varphi^{HEE,\nu,0} + \varphi^{HEE,\nu,level}\text{level_HE}_j + \varphi^{HEE,\nu,STEM}\text{STEM}_j) \\
&+ \varphi^{HEE,\text{dist}}\text{distance_HE}_{ij} \\
&+ \tilde{d}'_{iT_i^{SE}}\varphi^{HEE,\text{SE}} \\
&+ \text{delay}_{iT_i^{SE}}(\varphi^{HEE,delay,0} + \varphi^{HEE,delay,level}\text{level_HE}_j + \varphi^{HEE,delay,STEM}\text{STEM}_j) \\
&+ \varphi^{HEE,level\text{x}delay}\text{level_SE}_{iT_i^{SE}}\times\text{delay}_{iT_i^{SE}} \\
&+ X_{ij}'\varphi^{HEE,\text{interact}}
\end{aligned}$$

Level_HE$_j$ is the level of the higher education program. I follow Arcidiacono (2005) and define the level for each type of higher education by the average math ability of the enrolling students. I use professional college as a benchmark (0.20) and calculate differences with academic college (0.59) and university (0.79). Distance_HE$_{ij}$ is the distance in kilometers from the student's home to the chosen option. $\tilde{d}_{iT_i^{SE}}$ is a vector of dummy variables for each possible program a student can graduate from in high school and delay$_{iT_i^{SE}}$ the years of accumulated study delay. Since there are few students in the academic track that do not enroll in higher education, I do not distinguish between elective courses and estimate a common effect of each track on enrollment in the benchmark professional college. I also include a vector of interactions X_{ij} that includes all interactions between characteristics of the high school program the student graduated in (academic level, intensive math, classical languages) and the characteristics of the higher education program (level and STEM major).

I impose a similar model for graduation from higher education. I use a similar functional form for $\Psi_j^{HED}(.)$ as I did for $\Psi_j^{HEE}(.)$, but I also add more interaction effects in X_{ij} to take into account the higher education enrollment decision.[50] In particular, I include dummy variables for choosing the same level, upgrading a level, and choosing the same major. I add a shock that is distributed extreme value type 1 such that I obtain logit probabilities. Since these shocks are iid, it is important to take into account the enrollment decision to capture the correlation between enrollment decisions and the final degree a student obtains.

D.2 Reduced forms of high school data

In section 2, I explained how a measure of performance can be used to back out the optimal level of the effective study effort. This is still possible in the current model and follows from the FOC (22). A first implication of this is that students with the same state vector will choose the same effort levels within each program. Let y^*_{ijt} be the optimal choice of y_{it}, conditional on program choice j. We can now substitute this in the definition of y_{it} (15):

$$y^*_{ijt} = \frac{1 - \Pr(g^{track}_{it+1} = 0 | d_{it} = j, x_{it}, \nu_i)}{\Pr(g^{track}_{it+1} = 0 | d_{it} = j, x_{it}, \nu_i)}$$

with the current grade deterministic in d_{it} and x_{it} and $y^*_{ijt} = y^*_j(x_{it}, \nu_i)$. Note that both x_{it} and ν_i are observed here, therefore y^*_{ijt} is easily obtained from the observed probability to obtain the lowest performance outcome in each j when students behave optimally in the data. However, the finite number of observations and the large state space does not allow me to do this. Therefore, I recover the optimal levels and the performance thresholds by estimating an ordered logit model for the track

[50] In contrast to college enrollment rates, there is sufficient variation in graduation rates within programs of the same track. Therefore, I do not need to restrict the common parameters of the effect of study programs to be the same.

performance outcome with index $\ln y_j^*(x_{it}, \nu_i)$ and cut points $\bar{\eta}_{jr}^{track}$. The functional form of the index is similar to what is imposed for the fixed cost parameters (see below, equation (25)), but I allow for more flexibility by letting each initial observed and unobserved characteristic be track-specific and change (linearly) over different grades. I also allow distance to higher education options to affect performance and I add an effect of the lagged study program (academic level and dummy variables for intensive math and classical languages). Note that some of the thresholds are not identified from the data but from the institutional context that imposes restrictions on mobility (i.e. some thresholds can be ∞). I allow the thresholds to differ not only by different programs but also by the grade a student is in. Because there is little variation in the data, I restrict the program-specific part through three parameters that capture differences in the increase in thresholds for obtaining a higher outcome in each track. This is then assumed to be constant over grades and tracks (see Table A18). The ordered logit model also generates the probabilities for each performance outcome. For elective courses, I use the predicted values of $\ln y_j^*(x_{it}, \nu_i)$ and estimate the specification in equation (16). Both can then be used to construct the joint probabilities $\phi_{ijt}^{\bar{g}}(y_j^*(x_{it}, \nu_i))$.

As in Arcidiacono et al. (2016), I also obtain predicted values of $\Pr(d_{it}|x_{it}, \nu_i)$ (the CCPs) by estimating a flexible conditional logit with an index, similar to the index I used to predict effective study effort. I assume a functional form that is linear in observed and unobserved characteristics for each student characteristic, and I allow for more flexibility than in fixed costs by letting them be track-specific and change linearly over different grades. I also allow distance to higher education options to affect choices, while they are excluded from fixed costs. As explained further, the CCPs will be used to avoid solving the model during estimation and to back out the unobserved types in a first stage.

D.3 Cost estimates

The FOC (22) allows us to write the conditional value functions without an unknown marginal cost function. Substituting the utility function (18) in the conditional value function (17), after substituting marginal costs by (22) gives:

$$\begin{aligned} & v_j(x_{it}, \nu_i, y_j^*(x_{it}, \nu_i)) \\ & = -C_j^0(x_{it}, \nu_i) \\ & + \beta \sum_{\bar{g} \in G} \left[\bar{V}(x_{it+1}(\bar{g}), \nu_i) \left(\phi_{ijt}^{\bar{g}}(y_j^*(x_{it}, \nu_i)) - \left. \frac{\partial \phi_{ijt}^{\bar{g}}(y_{it})}{\partial y_{it}} \right|_{y_{it} = y_j^*(x_{it}, \nu_i)} y_j^*(x_{it}, \nu_i) \right) \right]. \end{aligned} \tag{24}$$

We already recovered $y_j^*(x_{it}, \nu_i)$ and $\phi_{ijt}^{\bar{g}}(y_j^*(x_{it}, \nu_i))$ from the data. $\frac{\partial \phi_{ijt}^{\bar{g}}(y_{it})}{\partial y_{it}}$ can be derived from the distributional assumptions on the performance measure. As explained in the text, it is the product of three ordered logit probabilities. We can apply the chain rule, knowing that for each ordered logit model we can find the derivative with respect to y_{it} recursively:

$$\begin{aligned} \frac{\partial \Pr(g_{it}^a = 0 | d_{it}, x_{it}, \nu_i, y_{it})}{\partial y_{it}} &= -\alpha_y^a \frac{1}{y_{it}} \Pr(g_{it}^a = 0 | d_{it}, x_{it}, \nu_i, y_{it})(1 - \Pr(g_{it}^a = 0 | d_{it}, x_{it}, \nu_i, y_{it})) \\ \frac{\partial \Pr(g_{it}^a = \bar{g} | d_{it}, x_{it}, \nu_i, y_{it})}{\partial y_{it}} &= -\alpha_y^a \frac{1}{y_{it}} \left(\Pr(g_{it}^a \leq \bar{g} | d_{it}, x_{it}, \nu_i, y_{it}) \Pr(g_{it}^a > \bar{g} | d_{it}, x_{it}, \nu_i, y_{it}) \right) \\ &\quad - \sum_{\tilde{g} < \bar{g}} \frac{\partial \Pr(g_{it}^a = \tilde{g} | d_{it}, x_{it}, \nu_i, y_{it})}{\partial y_{it}} \text{ for } \mathring{g} > 0 \end{aligned}$$

with $(a = track, clas, math)$ and $\alpha_y^{track} = 1$.

After solving the model for $\bar{V}(x_{it+1}(\bar{g}), \nu_i)$, we can use the logit probabilities (23) with these conditional value functions to estimate the value of a degree μ^{degree} and a specification for fixed costs $C_j^0(.)$ by using maximum likelihood. I assume the following functional form:[51]

$$
\begin{aligned}
C_j^0(x_{it},\nu_i) = {} & \mu_j^0 + \mu_j^{\text{grade}}\text{grade}_{ijt} \qquad (25)\\
& + S_i'(\mu_j^{S,0} + \mu^{S,\text{level}}\text{level\_SE}_{ijt} + \mu^{S,\text{math}}\text{math}_{ijt} + \mu^{S,\text{clas}}\text{clas}_{ijt})\\
& + \nu_i'(\mu_j^{\nu,0} + \mu^{\nu,\text{level}}\text{level\_SE}_{ijt} + \mu^{\nu,\text{math}}\text{math}_{ijt} + \mu^{\nu,\text{clas}}\text{clas}_{ijt})\\
& + \mu_{time}\text{time}_{ijt}\\
& + \text{retention}_{ijt}'(\mu^{\text{ret},0} + \mu^{\text{ret,level}}\text{level\_SE}_{ijt})\\
& + \mu_{up}\text{upgrade}_{ijt} + \mu_{down}\text{downgrade}_{ijt}\\
& + \mu_{staymath}\text{math}_{ijt} \times \text{math}_{it-1} + \mu_{stayclas}\text{clas}_{ijt} \times \text{clas}_{it-1}.
\end{aligned}
$$

μ is a vector of parameters to estimate. S_i is a vector of time-invariant observed student characteristics, ν_i is a vector of dummy variables that indicate to which type the student belongs, time_{ijt} is the daily commuting time to the closest school that offers the study program in the current grade and grade_{ijt} is the grade a student is in (set such that 1 is the first year of high school). Level_SE_{ijt} is the academic level of the track a student is in with 0 the vocational track, 1 the middle-practical track, 2 the middle-theoretical track, and 3 the academic track and math and clas refer to respectively programs with intensive math and with classical languages. Grade retention is captured by the 2x1 vector: retention_{ijt}. This vector contains a flow variable: a dummy equal to one if the student is currently in the same grade as the year before ("Repeat") and a stock variable that captures the years of study delay accumulated in previous years ("Study delay"). Finally, upgrade_{ijt} and downgrade_{ijt} are dummy variables indicating if a student is currently in a track with at a higher or lower academic level than the year before and $\mu_{staymath}$ and $\mu_{stayclas}$ capture preferences to stay in a program with the same elective courses.

Note that in section 4, the scale of the utility function was implicitly normalized

[51] Note that the part-time track does not have a grade structure. Therefore, I only model its fixed cost. Due to a lack of variation, I only estimate a choice-specific constant, which implies that student background should have the same effect on part-time and full-time dropout.

to unity. Therefore, all parameters μ are identified. However, to directly interpret the cost estimates, I rescale the parameters by dividing by μ_{time}. This way, the cost estimates can be measured in daily commuting time.

Finally, marginal costs $c_j(.)$ can be recovered from the FOC (22) without imposing additional structure.

D.4 CCP estimation

Hotz and Miller (1993) introduced the CCP method as an alternative to solving dynamic models, which is particularly useful if there is a terminal action (Arcidiacono and Ellickson, 2011). Hotz and Miller (1993) show that the future value term can be written as the conditional value function of an arbitrary choice and a nonnegative correction term that depends on its probability in the data:

$$\bar{V}(x_{it+1}, \nu_i) = \gamma + v_{d^*}(x_{it+1}, \nu_i, y_j^*(x_{it}, \nu_i)) - \ln \Pr(d_{it+1}^* | x_{it+1}, \nu_i) \tag{26}$$

with $\gamma \approx 0.577$ the Euler constant, d_{it+1}^* an arbitrary option $j = d^*$ and $v_{d^*}(.)$ the conditional value function of this option.

Case 1: $j = 0$ available in $t + 1$

If it is possible to leave secondary education in $t + 1$, we can choose $j = 0$ as the arbitrary choice and substitute its value function (19) in (26), with $\Psi_0^{HEE}(.) = 0$:

$$\overline{V}(x_{it+1}, \nu_i) = \gamma + \text{Degree}'_{it}\mu^{\text{degree}} - \ln \Pr(d_{it+1}^0 = 1 | x_{it+1}, \nu_i). \tag{27}$$

We can now substitute (27) in (24), such that for all $j \in se$:

$$
\begin{aligned}
& v_j(x_{it}, \nu_i, y_j^*(x_{it}, \nu_i)) \qquad (28) \\
& = -C_j^0(x_{it}, \nu_i) + \beta\gamma \\
& + \beta \sum_{\bar{g} \in G} \left[\begin{array}{c} \left(\text{Degree}'_{it}(\bar{g})\mu^{\text{degree}} - \ln \Pr(d_{it+1} = 0 | x_{it+1}(\bar{g}), \nu_i)\right) \\ \left(\phi_{ijt}^{\bar{g}}(y_j^*(x_{it}, \nu_i)) - \left.\frac{\partial \phi_{ijt}^{\bar{g}}(y_{it})}{\partial y_{it}}\right|_{y_{it} = y_j^*(x_{it}, \nu_i)} y_j^*(x_{it}, \nu_i) \right) \end{array} \right].
\end{aligned}
$$

The benefit of using the outside option $j = 0$ as the arbitrary choice is that this removes the future value terms in the current period conditional value functions. This is because the terminal nature of $j = 0$ allows us to write its conditional value function directly as a function of observables and parameters (see section 4.4). As in Hotz and Miller (1993), a nonparametric estimate of $\Pr(d_{it+1} = 0 | x_{it+1}, \nu_i)$ can be recovered from the data before estimating the model.

These conditional value functions can now be used as inputs in logit probabilities to recover the fixed cost parameters without having to solve the model.

Case 2: $j = 0$ available in $t + \rho_{it}$

For most students, we start modeling choices from the age of 12. At $t + 1$, they are age 13 and do not have that option because of compulsory schooling laws. They will get the outside option $j = 0$ at $t + 6$. I write ρ_{it} to be the number of years it takes before the CCP correction term with the outside option can be applied: $\rho_{it} = \max\{1, 18 - Age_{it}\}$. We now need to repeat the CCP method in future values until the outside option is available. This is an application of finite dependence, introduced in Arcidiacono and Miller (2011). In contrast to their application on problems that have a renewal action in the future, I apply it to the terminal action of choosing to leave secondary education in the outside option (no higher education). The exposition in this section is similar to Arcidiacono and Miller (2011) and Arcidiacono and Ellickson (2011).

The choice probabilities (23) at the optimal levels of the effective study effort can be written by using differenced value functions. Let $v_j^*(x_{it}, \nu_i) \equiv v_j(x_{it}, \nu_i, y_j^*(x_{it}, \nu_i))$ be the conditional value function at the optimal level of effective study effort and $u_j^*(x_{it}, \nu_i) \equiv u_j(x_{it}, \nu_i, y_j^*(x_{it}, \nu_i))$ the flow utility at this level:

$$\Pr(d_{it} = j | x_{it}, \nu_i) = \frac{\exp\left(v_j^*(x_{it}, \nu_i) - v_{j'}^*(x_{it}, \nu_i)\right)}{1 + \sum_{j^\circ \in \Phi(x_{it})} \exp\left(v_{j^\circ}^*(x_{it}, \nu_i) - v_{j'}^*(x_{it}, \nu_i)\right)}$$

$$\begin{aligned}
&\text{with } v_j^*(x_{it}, \nu_i) - v_{j'}^*(x_{it}, \nu_i) \\
&= u_j^*(x_{it}, \nu_i) + \beta \sum_{\bar{g} \in G} \phi_{ijt}^{\bar{g}}(y_{ijt}^*) \bar{V}(x_{it+1}(\bar{g})) \\
&- u_{j'}^*(x_{it}, \nu_i) - \beta \sum_{\bar{g} \in G} \phi_{ij't}^{\bar{g}}(y_{ij't}^*) \bar{V}(x_{it+1}(\bar{g})),
\end{aligned} \tag{29}$$

for any $j^{'} \in \Phi(x_{it})$. Substitute the CCP representation of the future value as a function of the CCP of an arbitrary choice and its conditional value function (26) in (29):

$$\begin{aligned}
&v_j^*(x_{it}, \nu_i) - v_{j'}^*(x_{it}, \nu_i) \\
&= u_j^*(x_{it}, \nu_i) + \beta \sum_{\bar{g} \in G} \phi_{ijt}^{\bar{g}}(y_{ijt}^*) \left(\gamma + v_{d^*}^*(x_{it+1}(\bar{g}), \nu_i) - \ln \Pr(d_{it+1}^* | x_{it+1}(\bar{g}), \nu_i)\right) \\
&- u_{j'}^*(x_{it}, \nu_i) - \beta \sum_{\bar{g} \in G} \phi_{ij't}^{\bar{g}}(y_{ij't}^*) \left(\gamma + v_{d^*}^*(x_{it+1}(\bar{g}), \nu_i) - \ln \Pr(d_{it+1}^* | x_{it+1}(\bar{g}), \nu_i)\right).
\end{aligned} \tag{30}$$

Define the cumulative probability of being in a particular state given the current state variable and choice, and a particular decision sequence $d_i^* = (d_{it}, d_{it+1}^*, d_{it+2}^*, ...d_{it+\rho_{it}}^*)$:

$$\begin{aligned}
&\kappa_\tau^*(g_{i\tau+1} = \bar{g} | x_{it}, \nu_i) = \phi_{id^*\tau}^{\bar{g}}(y_{d^*}^*(x_{i\tau}, \nu_i)) \text{ if } \tau = t \\
&\kappa_\tau^*(g_{i\tau+1} = \bar{g} | x_{it}, \nu_i) = \sum_{\bar{g}_\tau \in G} \phi_{id^*\tau}^{\bar{g}}(y_{d^*}^*(x_{i\tau}, \nu_i)) \kappa_{\tau-1}^*(g_{i\tau} = \bar{g}_\tau | x_{it}, \nu_i) \text{ if } \tau > t
\end{aligned}$$

with $\phi^{\bar{g}}_{id^*\tau}(y^*_{d^*}(x_{i\tau}, \nu_i))$ the probability of receiving performance outcome $\bar{g}$ at time $t = \tau + 1$, in the program a student will be at $t = \tau$ according to the decision sequence d^*_i. Similarly, define $\kappa^{'}_\tau$ to be the transitions in a sequence where the choice in t is different: $d^{'}_i = (d'_{it}, d^*_{it+1}, d^*_{it+2}, ...d^*_{it+\rho_{it}})$.[52] We can then repeat the CCP method in each of the future periods and rewrite (30) as the sum of future flow utilities and CCPs until the outside option becomes available at $t + \rho_{it}$:

$$
\begin{aligned}
& v^*_j(x_{it}, \nu_i) - v^*_{j'}(x_{it}, \nu_i) \\
& = u^*_j(x_{it}, \nu_i) - u^*_{j'}(x_{it}, \nu_i) \\
& + \sum_{\tau=t+1}^{t+\rho_{it}-1} \beta^{\tau-t} \sum_{\bar{g}\in G} [u^*_{d^*}(x_{i\tau}(\bar{g}), \nu_i) - \ln \Pr(d^*_{i\tau}|x_{i\tau}(\bar{g}), \nu_i)]\kappa^*_{\tau-1}(\bar{g}|x_{it}, \nu_i) \\
& - \sum_{\tau=t+1}^{t+\rho_{it}-1} \beta^{\tau-t} \sum_{\bar{g}\in G} [u^*_{d^*}(x_{i\tau}(\bar{g}), \nu_i) - \ln \Pr(d^*_{i\tau}|x_{i\tau}(\bar{g}), \nu_i)]\kappa'_{\tau-1}(\bar{g}|x_{it}, \nu_i) \\
& + \beta^{\rho_{it}} \sum_{\bar{g}\in G} \overline{V}(x_{t+\rho_{it}}(\bar{g}), \nu_i)\kappa^*_{t+\rho_{it}-1}(\bar{g}|x_{it}, \nu_i) \\
& - \beta^{\rho_{it}} \sum_{\bar{g}\in G} \overline{V}(x_{t+\rho_{it}}(\bar{g}), \nu_i)\kappa^{'}_{t+\rho_{it}-1}(\bar{g}|x_{it}, \nu_i).
\end{aligned}
$$

$\overline{V}(x_{t+\rho_{it}}, \nu_i)$, the value of behaving optimally when the outside option is available and can be written as in (27). The calculation of the value function is now possible after choosing the arbitrary options in each period, the prediction of their CCPs, and the predictions of optimal effort in the study program. However, further simplifications follow from a good choice of "arbitrary" options.

Since upward mobility from the lowest track is never allowed, I argue that the arbitrary choices should always be the lowest track available in each period: the vocational track if a student is not 15 years old yet, and the part-time track if the student is older. This choice significantly removes the number of CCPs and future utility terms we need. From the moment students choose the vocational track, they

[52] We can also allow a more general alternative sequence in which the choice in each period is different but here it is sufficient to only let the first choice be different.

can no longer make choices until the part-time track becomes available. Similarly, once students opt for the part-time track, they can no longer make other choices until the outside option is available. Therefore, we only need a CCP at the time a student is switching tracks in the sequence. Moreover, since the part-time track does not follow a grade-structure, and students can never return to the standard grade-structure, the state variables will not evolve anymore in a way that depends on choices made. Arcidiacono and Ellickson (2011) explain that in this case, the future utility terms after choosing that option can be ignored in estimation as they will cancel out in the differenced value functions.

The same procedure is applied within $u_j^*(x_{it}, \nu_i) = -C_j^0(x_{it}, \nu_i) - c_j(x_{it}, \nu_i) y_j^*(x_{it}, \nu_i)$. By replacing the marginal cost of effort by the marginal benefit of effort in the data, future value terms also enter directly into $u_j^*(x_{it}, \nu_i)$ (see (24)). Because $\sum_{\bar{g} \in G} \frac{\partial \phi_{ijt}^{\bar{g}}(y_{it})}{\partial y_{it}} = 0$, all terms that do not depend on performance drop out such that the same simplifications arise because of finite dependence.

D.5 Unobserved heterogeneity

To allow for types to remain unobserved to the econometrician, I follow the two-stage estimator of Arcidiacono and Miller (2011). I assume there are $M = 2$ unobserved types m in the population, with an estimated probability to occur π_m. For interpretability, I model the types as independent from observed student background. A dummy for belonging to type 2 then enters each part of the model as if it were an observed student characteristic. To avoid an initial conditions problem, I condition the type distribution on the age the student starts secondary education: age_start_i. This is because students who accumulated study delay before secondary education will be faced with different opportunities in the model because they will be able to drop out more quickly. Since starting age depends on past grade retention, it is likely correlated with unobserved ability, creating a bias in the estimates. By conditioning the unobserved types on age_start_i, we can allow for this correlation.[53] The loglikelihood function is

$$\ln L_i = \ln \sum_{m=1}^{M} \pi_{m|age\_start} L_i^m$$

with

$$L_i^m = \prod_{t=1}^{T_i^{SE}} L_{it}^{program,m} \times L_{it+1}^{performance,m} \times L_{it}^{ccp,m} \times L_i^{HEE,m} \times L_i^{HED,m}$$

with $L_{it}^{program,m}$ and $L_i^{HEE,m}$ given by logit choice probabilities (23), with conditional value functions (28) and (19). $L_i^{HED,m}$ is given by the conditional logit probabilities on the different possibilities for higher education graduation outcomes. The likelihood contribution of the performance outcome in secondary education is given by ordered logit probabilities $L_{it+1}^{\text{performance},m}$ and $L_{it}^{ccp,m}$ are the CCP predictors. Note that the inclusion of unobserved types makes the function no longer additively separable such that sequential estimation is not possible.

Arcidiacono and Miller (2011) show that additive separability can be restored. The estimation procedure is an adaptation of the EM algorithm. It starts from a random probability of each observation to belong to each type. The entire model can then be estimated as explained above but weighs each observation-type combination by the probability that the student belongs to the type. Afterwards, the joint likelihood of the data conditional on each type is used to update the individual type probabilities, conditional on the data, using Bayes rule. This is repeated until convergence of the likelihood function. I use the two-stage estimator of Arcidiacono and Miller (2011) which implies that in the calculation of the joint likelihood, reduced form estimates of the CCPs are used for $L_{it}^{program,m}$, instead of the choice probabilities from the structural model. This means that the fixed cost parameters and the common

[53]This is similar to Keane and Wolpin (1997), who start their model at age 16 and condition the types on the educational attainment at that age.

component of the value of a degree are recovered in the second stage. Finally, the FOC (22) is used to recover the marginal costs.

Standard errors are obtained using a bootstrap procedure. I sample students with replacement from the observed distribution of the data and use 150 replications. Since the EM algorithm takes some time to converge, I do not correct for estimation error in the probabilities to belong to each type.

E Simulation details

All predicted values are calculated as follows. I first categorize students by their demographic characteristics: gender, language ability, math ability, SES, and the age they start high school. I discretize the observed ability distribution by creating four equally sized groups for each measure. Every student then belongs to one group which is a unique combination of these variables. Within each group, I use the average travel times and distances. Each group is then used to calculate the value functions for each unobserved type. To limit the number of calculations, I drop groups with less than 10 students and verify that this has a negligible effect on the distribution of student characteristics.

E.1 High school

After obtaining the value functions, I proceed to simulation during high school. For each type, I draw 10,000 students using the empirical distribution of the observable characteristics. I also take draws of taste shocks for every option in every period, as well as performance shocks in every period for every performance outcome. The average statistics are then calculated on a total of 20,000 draws. Given the simulated outcomes of high school, I use the closed-form expressions for higher education to calculate enrollment and graduation.

This procedure allows for a substantial total number of draws while needing only a limited number of students to use for a grid search to find the optimal effort level within each possible program. The grid search for effort levels starts at the optimal value of the scalar y_{it} in each program j in the data: $y_j^*(x_{it}, \nu_i)$ and looks for better levels using five sequential loops and an additional step to check for a corner solution. The first loop looks at changes in the log of the effective study effort by 1 unit with a minimum of -5 and a maximum of +5. The second loop divides steps and thresholds by five, the third by 25, the fourth by 125, and the fifth by 625, such that the final precision is 0.0016 (which is about 0.16% for the effective study effort y). Finally, I check if a corner solution is optimal by setting $y = 0$ and changing the performance distribution to predict the worse outcome with probability 1.

Standard errors are obtained by using the different estimates of each bootstrap sample and by repeating the entire procedure for each of them.

E.2 Higher education

Note that to evaluate the impact on higher education outcomes, a structural model in high school is needed as it allows for policy counterfactuals that will not change the primitives of the model, like the fixed cost of a study program, marginal costs of effort within a program or the value of a degree, but it will change student behavior. Without a structural model, we would not be able to assess the effects of changes in policy. For outcomes after secondary education, we do not need to know the same primitives of the model but only the way these outcomes are influenced by high school outcomes, after controlling for observed and unobserved student characteristics. Therefore, I model a reduced form function only. This is similar to the approach in the dynamic treatment effect literature (Heckman et al., 2016), but I only apply it to choices after leaving high school to be able to do counterfactual simulations during secondary education in which students are forward-looking.

The estimated functions of both enrollment and graduation can be used to look at the impact of counterfactual policies in secondary education. Let $x_{it_{HE}}(Policy = 0)$ be the realized state vector of i at time t_{HE} in the status quo scenario, and $x_{it_{HE}}(Policy = p')$ the state vector in the counterfactual scenario. The expected impact on the proportion of students with long-run outcome HE of policy p' is then given by:

$$E_{x,\nu}\left[P_j^{HE}(x_{it_{HE}}(Policy = p'), \nu_i) - P_j^{HE}(x_{it_{HE}}(Policy = 0), \nu_i)\right] \text{ for } HE = \{HEE, HED\}$$

with $E_{x,\nu}$ an expectations operator over the empirical distribution of the observables x and the estimated distribution of the unobserved types ν. P_j^{HE} is the probability of the enrollment decision or higher education degree outcome of each college option as a function of the state variables.

E.3 Fit of the model

Table A2 shows the ability of the model to replicate the actual data. The model does a good job of predicting the patterns in the data such that it can be used for counterfactual simulations. We see that graduation rates in different track and higher education outcomes are predicted very precisely. There is a slight overprediction in the number of students with a B-certificate leading to a small overprediction in the number of students with study delay.

Table A2: Predictions of the model

		Data	Predictions	
High school (% of students)				
Academic		38.27	40.02	(2.07)
	clas+math	5.06	5.03	(0.65)
	clas	6.11	3.18	(0.42)
	math	13.24	14.59	(1.27)
	other	13.86	17.22	(1.29)
Middle-Theoretical		15.86	16.10	(1.24)
	math	2.42	3.11	(0.46)
	other	13.44	12.99	(0.97)
Middle-Practical		11.85	8.14	(1.19)
Vocational		19.43	21.57	(0.89)
Dropout		14.60	14.17	(0.67)
Students with at least 1 B-certificate		35.40	37.53	(0.81)
Students with at least 1 C-certificate		30.01	30.69	(0.77)
Students with at least 1 year of study delay		31.62	33.22	(0.91)
Higher education (% of students)				
Enrollment		58.18	58.15	(0.75)
Graduation		44.01	44.25	(0.75)
University degree		12.43	11.22	(0.55)
Academic college degree		6.05	6.26	(0.38)
Professional college degree		25.53	26.77	(0.69)
Degree in STEM major		17.76	18.01	(0.65)

Note: Clas= classical languages included. Math= intensive math. Observed outcomes in the data and predictions from the proposed dynamic model. Bootstrap standard errors of predicted values in parentheses.

E.4 Welfare

Opportunity cost

I assume an opportunity cost of \$10/hour. This is chosen to approximate the opportunity cost of students in high school and is consistent with Kapor et al. (2018). Students are not allowed to work until they are 15 years old and the wage often depends on their age. In 2012 the minimum wage ranged between €6.8 and €9.7/hour.[54] Only a small amount of taxes is paid on this if they work a limited amount of hours. To compare to OECD estimates, I use the PPP adjusted exchange rate of dollars (0.82), which results in wages between \$8 and \$12. Note that the model is in years while the estimates are scaled in minutes/day. Therefore, I multiply them by the wage per minute (\$10/60) and the 177 school days there are in a year.

Gains from reducing grade retention

The direct cost and the total foregone earnings can be found in Table IV.1.6 in OECD (2013). I subtract the net income (49%) to only capture the externality. This number was calculated by dividing column (7) by column (1) in Table A10.2 in OECD (2012).

Reinvestment of gains

Estimates in the literature for the effect of a one-time "helicopter drop" increase of \$1,000 on the ability distribution are around 1% to 2% of a standard deviation (Gigliotti and Sorensen, 2018; Lafortune et al., 2018). This implies that reinvesting the efficiency gains of the "Downgrade" policy could result in substantial gains for students. Using the estimates in Table A24 and the savings from avoiding grade retention (\$2,910), a 1.5% effect per \$1,000 on each of the observed ability measures in the downgrade policy would bring back \$1,200 in student welfare, increase graduation rates in higher education by 1.3 %points, reduce study delay by 0.4 %points and dropout by 0.6 %points. This in turn also creates additional savings that could be

[54] https://www.jobat.be/nl/artikels/wat-is-het-minimumloon-voor-een-jobstudent/ (consulted March 2018).

reinvested.

The estimates should be interpreted with caution. First, gender, socioeconomic status, and the unobserved type might all be capturing initial skills that are not captured by the language and math ability measures. Therefore, a policy that changes skills might have a bigger effect than estimated now. On the other hand, ability measures could also capture other things that might not respond to increased funding, e.g. parental characteristics that are not captured by the SES dummy.

F Sensitivity analysis

This Appendix discusses the impact of alternative model specifications on the counterfactual simulations of this paper.

Table A3 looks at the impact of observed and unobserved ability. We see important differences, especially on predicted higher education graduation. In the "Repeat" policy we see an underestimation of the decrease in graduating from higher education (-0.61 instead of -1.70 %points), while in the "Downgrade" policy we see an overestimation of the decrease (-0.84 instead of -0.30 %points). These results can be explained by a failure to take into account the ability bias on the estimated effect of tracks. When isolated, both observable ability measures and unobserved types move the estimate closer to the baseline results. However, for the "Repeat" policy it is mainly coming from the inclusion of observable measures of ability, while for the downgrade policy the types help more. Note however that the ability bias was also smaller in the downgrade policy. These results can be explained by the nature of the data. The availability of rich, continuous measures of ability helps a lot to capture the main source of ability bias. This leaves room for unobserved types to capture more subtle differences between students. Therefore, the limiting structure of having a finite number of types becomes a smaller concern with rich data.

Table A4 compares the baseline estimation method with two approaches that use

Table A3: Sensitivity analysis: observed and unobserved ability

	Study delay		**High school dropout**		**Higher education graduation**		**Student welfare**	
			Predicted value in %					
Status quo								
Baseline	33.22	(0.91)	14.17	(0.67)	44.25	(0.75)		
Obs ability: YES, types: NO	33.93	(0.98)	15.24	(0.76)	43.78	(1.37)		
Obs ability: NO, types: YES	31.11	(1.11)	14.62	(0.64)	46.70	(1.76)		
Obs ability: NO, types: NO	31.29	(1.41)	11.43	(1.37)	50.87	(3.75)		
			Change in %points				*Change in $1000*	
Repeat policy								
Baseline	9.48	(0.57)	3.94	(0.33)	-1.70	(0.22)	-2.14	(0.26)
Obs ability: YES, types: NO	10.42	(0.59)	4.44	(0.37)	-1.80	(0.33)	-2.14	(0.24)
Obs ability: NO, types: YES	9.89	(0.61)	4.20	(0.38)	-0.72	(0.41)	-2.22	(0.26)
Obs ability: NO, types: NO	10.64	(0.70)	4.67	(0.55)	-0.61	(0.57)	-2.30	(0.27)
Downgrade policy								
Baseline	-9.82	(0.55)	-1.61	(0.25)	-0.30	(0.18)	-1.02	(0.14)
Obs ability: YES, types: NO	-10.35	(0.78)	-1.46	(0.34)	-0.57	(0.24)	-0.96	(0.14)
Obs ability: NO, types: YES	-9.22	(0.65)	-1.83	(0.24)	-0.26	(0.19)	-0.72	(0.11)
Obs ability: NO, types: NO	-9.16	(0.70)	-1.13	(0.38)	-0.84	(0.32)	-0.69	(0.11)

Note: Predictions from the dynamic model under alternative specifications. Obs ability refers to the variables on initial math and language ability. If types = YES, it means that two unobserved types are allowed for in the estimation. Status quo = students can choose to downgrade or repeat grade after obtaining B-certificate, Repeat = students must repeat grade after obtaining B-certificate, Downgrade = students must downgrade and not repeat grade after obtaining B-certificate. Bootstrap standard errors in parentheses.

different identifying assumptions. As discussed in section 4.7, travel time to high school programs is excluded from equations that predict higher education enrollment and graduation but exclusion is not required for identification (Heckman and Navarro, 2007). I therefore also estimate a specification in which I add measures of travel time to several programs in the final grade of high school. I add time to the vocational track, and differences in travel time for moving up a track for every other track. I also add the difference in travel time between options with and without intensive math and with and without classical languages. These travel times are interacted in the same way as other observable student characteristics. The resulting simulations are almost identical to the baseline results. In a final specification, I follow Carneiro et al. (2003), Heckman et al. (2016), and Lin (2019) and use additional measurement data to identify unobserved heterogeneity. To do this, I add (ordered) logit models to the likelihood function of stage 1 of the estimation approach to predict the variables listed in Table A5. The results for counterfactual simulations are similar. Furthermore, we can conclude from Table A6 and Table A7 that types especially help to capture non-cognitive skills.

Table A4: Sensitivity analysis: identification unobserved ability

	Study delay		**High school dropout**		**Higher education graduation**		**Student welfare**	
			Predicted value in %					
Status quo								
Baseline	33.22	(0.91)	14.17	(0.67)	44.25	(0.75)		
High school travel time not excluded	33.30	(0.93)	14.32	(0.64)	44.23	(0.81)		
+ Measurements added	34.72	(0.95)	14.53	(0.70)	43.72	(1.23)		
			Change in %points				*Change in $1000*	
Repeat policy								
Baseline	9.48	(0.57)	3.94	(0.33)	-1.70	(0.22)	-2.14	(0.26)
High school travel time not excluded	9.73	(0.45)	3.81	(0.33)	-1.75	(0.21)	-2.06	(0.25)
+ Measurements added	10.21	(0.49)	3.89	(0.37)	-1.84	(0.27)	-1.99	(0.21)
Downgrade policy								
Baseline	-9.82	(0.55)	-1.61	(0.25)	-0.30	(0.18)	-1.02	(0.14)
High school travel time not excluded	-10.85	(0.63)	-1.80	(0.23)	-0.39	(0.17)	-1.03	(0.15)
+ Measurements added	-11.04	(0.69)	-1.88	(0.28)	-0.54	(0.21)	-0.95	(0.14)

Note: Predictions from the dynamic model under alternative specifications. Travel times to high school programs added in equations that predict higher education enrollment and graduation in both alternative specifications. The final specification also adds measurements, summarized in Table A5, to the first stage of the estimation procedure. Bootstrap standard errors in parentheses.

Finally, Table A8 shows that results are robust for using two commonly used discount factors in the literature.

Table A5: Measurements: summary statistics

Variable	Description	Obs	Mean	SD	Min	Max
	Question asked to teacher elementary school (scale of 1 to 5 with 1 the lowest)					
CONS1	Could pay attention in class; has sufficient intellectual capabilities to follow; is smart	3,938	3.635	1.293	1	5
CONS2	Was motivated for school work; wanted to do it really well; worked without reluctance	3,938	3.725	1.243	1	5
CONS3	Could tell a coherent story; explore a topic; stay on the subject	3,936	3.667	1.187	1	5
AGREE1	Did not distrube class intensionally; did not aim to boycot learning	3,917	4.259	1.108	1	5
AGREE2	Held herself to the class rules; waited for her turn; it was not necessary to constantly call her to order	3,935	4.088	1.122	1	5
AGREE3	Was averse to hostilities; was friendly and kind to others; experienced no pleasure in teasing and bullying of others	3,934	3.992	1.104	1	5
EXTRA1	Was open to the teacher; was spontaneous; not defensive	3,935	3.854	1.146	1	5
EXTRA2	Made an energetic and vital impression; looked happy	3,933	3.898	1.064	1	5
EXTRA3	Made contact with fellow students; was open and approachable	3,924	3.989	1.023	1	5
	Definition					
IQ	IQ score, discretized using cutoffs 80, 90, 100, 110 and 120	5,084	3.647	1.361	1	6
Income	Monthly household income in BEF after taxes (1 EUR ≈40 BEF), discretized using cutoffs 40k, 60k, 80k, 100k	5,158	2.471	1.347	1	5
Work	At least one parent is active on the labor market.	4,749	0.861	0.346	0	1

Note: description and summary statistics of measurements of initial traits used in sensitvity checks.

Table A6: Measurements: part 1 of 2

	CONS1		CONS2		CONS3		AGREE1		AGREE2		AGREE3	
Male	0.070	(0.062)	-0.670	(0.066)	-0.226	(0.058)	-1.048	(0.076)	-1.024	(0.066)	-0.818	(0.064)
Language ability	1.624	(0.089)	1.062	(0.071)	1.304	(0.069)	0.367	(0.050)	0.362	(0.049)	0.293	(0.054)
Math ability	0.792	(0.080)	0.581	(0.057)	0.438	(0.059)	0.248	(0.051)	0.289	(0.054)	0.186	(0.054)
High SES	0.476	(0.072)	0.415	(0.076)	0.447	(0.080)	0.075	(0.089)	-0.030	(0.086)	0.032	(0.078)
Type 2	2.239	(0.071)	2.996	(0.076)	2.496	(0.072)	1.898	(0.073)	2.096	(0.068)	2.069	(0.069)
Cut point outcome 2	-3.234	(0.093)	-3.163	(0.091)	-3.327	(0.090)	-3.480	(0.109)	-3.573	(0.102)	-3.582	(0.117)
Cut point outcome 3	-1.012	(0.056)	-1.307	(0.065)	-1.224	(0.058)	-2.255	(0.077)	-1.998	(0.073)	-1.838	(0.070)
Cut point outcome 4	0.615	(0.061)	0.310	(0.057)	0.471	(0.061)	-1.425	(0.068)	-1.049	(0.067)	-0.589	(0.060)
Cut point outcome 5	2.548	(0.073)	2.290	(0.066)	2.650	(0.072)	-0.099	(0.064)	0.588	(0.062)	0.987	(0.061)

Note: Estimates of ordered logit model on measurements used in sensitivity checks. Bootstrap standard errors in parentheses.

Table A7: Measurements: part 2 of 2

	EXTRA1		EXTRA2		EXTRA3		IQ		Income		Work	
Male	-0.547	(0.064)	-0.144	(0.064)	-0.231	(0.059)	0.559	(0.061)	-0.123	(0.050)	0.165	(0.099)
Language ability	0.258	(0.050)	0.310	(0.052)	0.219	(0.048)	1.619	(0.062)	0.142	(0.042)	0.788	(0.082)
Math ability	0.263	(0.046)	0.342	(0.052)	0.209	(0.044)	1.395	(0.079)	0.145	(0.042)	0.087	(0.061)
High SES	0.142	(0.073)	0.317	(0.072)	0.215	(0.069)	0.196	(0.065)	1.662	(0.081)	1.788	(0.251)
Type 2	2.477	(0.071)	2.272	(0.072)	2.021	(0.065)	0.502	(0.061)	0.142	(0.051)	0.160	(0.101)
Cut point outcome 2	-3.099	(0.096)	-3.470	(0.120)	-3.784	(0.136)	-4.021	(0.099)	-0.471	(0.043)		
Cut point outcome 3	-1.349	(0.062)	-1.457	(0.061)	-1.733	(0.073)	-1.975	(0.061)	0.563	(0.043)		
Cut point outcome 4	-0.077	(0.059)	0.055	(0.056)	-0.300	(0.056)	0.144	(0.054)	1.703	(0.046)		
Cut point outcome 5	1.697	(0.061)	1.919	(0.068)	1.485	(0.058)	2.519	(0.060)	2.865	(0.059)		
Cut point outcome 6							5.003	(0.090)				
Constant											1.694	(0.086)

Note: Estimates of (ordered) logit model used in sensitivity checks. Bootstrap standard errors in parentheses.

Table A8: Sensitivity analysis: discount factor

	Study delay		**High school dropout**		**Higher education graduation**		**Student welfare**	
			Predicted value in %					
Status quo								
Baseline ($\beta = 0.9$)	33.22	(0.91)	14.17	(0.67)	44.25	(0.75)		
Alternative ($\beta = 0.95$)	32.95	(0.11)	13.98	(0.68)	44.56	(0.87)		
			Change in %points				*Change in \$1000*	
Repeat policy								
Baseline ($\beta = 0.9$)	9.48	(0.57)	3.94	(0.33)	-1.70	(0.22)	-2.14	(0.26)
Alternative ($\beta = 0.95$)	9.45	(0.53)	3.50	(0.35)	-1.99	(0.21)	-2.53	(0.28)
Downgrade policy								
Baseline ($\beta = 0.9$)	-9.82	(0.55)	-1.61	(0.25)	-0.30	(0.18)	-1.02	(0.14)
Alternative ($\beta = 0.95$)	-10.19	(0.76)	-1.81	(0.28)	-0.26	(0.18)	-1.13	(0.17)

Note: Bootstrap standard errors in parentheses.

G Tables

Table A9: High school program and student background

Study program	Students		Male	Language ability	Math ability	High SES
All	*5158*	*(100.0%)*	*0.50*	*0.00*	*0.00*	*0.28*
Academic	*1974*	*(38.3%)*	*0.40*	*0.71*	*0.64*	*0.49*
clas+math	261	(5.1%)	0.46	1.15	1.05	0.63
clas	315	(6.1%)	0.37	0.94	0.68	0.58
math	683	(13.2%)	0.49	0.74	0.75	0.51
other	715	(13.9%)	0.32	0.41	0.36	0.38
Middle-Theoretical	*818*	(15.9%)	*0.53*	*0.11*	*0.19*	*0.22*
math	125	(2.4%)	0.70	0.32	0.47	0.30
other	693	(13.4%)	0.50	0.07	0.14	0.21
Middle-Practical	*611*	(11.8%)	*0.51*	*-0.06*	*-0.02*	*0.22*
Vocational	*1002*	*(19.4%)*	*0.51*	*-0.76*	*-0.75*	*0.10*
13th grade	609	(11.8%)	0.49	-0.67	-0.69	0.11
12th grade	393	(7.6%)	0.54	-0.89	-0.85	0.08
Dropout	*753*	*(14.6%)*	*0.67*	*-0.92*	*-0.86*	*0.07*
Part-time	431	(8.4%)	0.71	-0.97	-0.90	0.06
Full time	322	(6.2%)	0.62	-0.86	-0.81	0.08

Note: Ability measured using IRT score on tests at start of secondary education. Score normalized to be mean zero and standard deviation 1. High SES= at least one parent has higher education degree. Clas= classical languages included. Math= intensive math. Students in vocational track only obtain full high school degree after an additional 13th grade. dropout split between students directly opting for full time dropout or first choosing part-time option.

Table A10: High school program and higher education outcomes: summary statistics

	Higher education	
Study program	Enrollment	Degree
All	*58.2*	*44.0*
Academic	*96.9*	*84.2*
clas+math	99.2	94.3
clas	99.4	90.5
math	97.8	88.1
other	94.1	74.1
Middle-Theoretical	*82.0*	*51.7*
math	99.2	72.8
other	78.9	47.9
Middle-Practical	*54.8*	*27.5*
Vocational (13th grade)	13.5	*2.6*
Dropout	*0*	*0*

Note: Percentage of all students (including dropouts), conditional on high school program. Clas= classical languages included. Math= intensive math. Students in vocational track only obtain full high school degree after an additional 13th grade.

Table A11: High school program and level and major college degree: summary statistics

	Academic level higher education			**Major**
Study program	University	Academic college	Professional college	STEM
All	*12.4*	*6.0*	*25.5*	*17.8*
Academic				
clas+math	67.0	14.2	13.0	54.0
clas	48.6	10.5	31.4	22.2
math	33.7	18.3	36.2	47.0
other	9.5	6.9	57.8	16.2
Middle-Theoretical				
math	7.2	20.8	44.8	56.0
other	0.7	3.3	43.9	17.5
Middle-Practical	*0.2*	*2.9*	*24.4*	*12.3*
Vocational (13th grade)	*0*	*0.2*	*2.5*	*0.3*

Note: Percentage of all students (including dropouts), conditional on high school program. Three types of higher education options in decreasing order of academic level: university, academic college, professional college. Graduation rates add up to the total rate of 44.0%. Each level has different programs that could be STEM. Graduation from STEM programs is reported. Clas= classical languages included. Math= intensive math.

Table A12: Exclusions because of certificates (in % of certificates)

	Tracks excluded				**Only elective courses excluded**
Current track	Academic	+Middle-Theoretical	+Middle-Practical	+Vocational	
Academic					
grade 7+8	8.9	4.1	4.0	0.8	2.4
grade 9+10	7.9	4.4	4.4	3.8	2.5
grade 11+12	6.2	6.2	6.2	6.2	0
Middle-Theoretical					
grade 7+8	29.2	21.5	19.7	1.2	0.5
grade 9+10	100	16.5	12.4	6.4	0
grade 11+12	100	11.2	11.2	11.2	0
Middle-Practical					
grade 7+8	41.3	33.8	30.5	3.7	0.5
grade 9+10	100	100	22.3	9.3	0
grade 11+12	100	100	15.1	15.1	0
Vocational					
grade 7+8	100	100	100	7.1	0
grade 9+10	100	100	100	13.8	0
grade 11+12+13	100	100	100	13.6	0

Note: Summary of implications of A-, B- and C-certificates. C-certificate: repeat grade, i.e. all tracks excluded, B-certificate can exclude entire tracks or only elective courses. Only electives excl. = math options or classical languages excluded by certificate. Upward mobility always excluded from grade 7 on in the vocational track and from grade 9 on in the other tracks. Track switching from grade 11 on is not possible.

Table A13: Impact performance during secondary education

	Students	**High school**	**Higher education**	
		Dropout	Enrollment	Degree
All	100	14.6	58.2	44.0
At least 1 B-certificate	35.4	21.0	36.6	19.9
At least 1 C-certificate	30.0	38.6	28.9	14.7
At least 1 year of study delay	31.6	26.7	37.7	18.8

Note: First column: share of students for each performance outcome during high school. Column 2-4: share of students for each long run outcome, conditional on obtaining a bad performance outcome in high school. A-certificate: proceed to next grade, C-certificate: repeat grade, B-certificate: repeat or downgrade.

Table A14: Transitions in educational system (in % of students)

Panel A: transitions in high school

	High school outcome					
	Academic	Middle-theoretical	Middle-practical	Vocational	Dropout	Total
First track						
Academic	36.9	10.5	7.1	4.6	3.9	63.0
Middle-theoretical	1.3	4.4	3.8	6.4	3.6	19.4
Middle-practical	0.1	0.9	1.0	3.0	1.9	6.9
Vocational	0	0	0	5.5	5.3	10.8
Total	38.3	15.9	11.8	19.4	14.6	100

Panel B: transitions after high school

	Final outcome				
	Higher education degree	Higher education no degree	High school degree	High school dropout	Total
High school outcome					
Academic	32.2	5.0	1.0	0	38.3
Middle-theoretical	8.2	4.8	2.8	0	15.9
Middle-practical	3.3	3.3	5.3	0	11.8
Vocational	0.3	1.3	17.8	0	19.4
Dropout	0	0	0	14.6	14.6
Total	44.0	14.4	27.0	14.6	100.0

Note: Study program choices of students when they enter and leave secondary education.

Table A15: Costs of schooling: main estimates

	Fixed costs		Log of marginal costs	
Time	1	(.)	-0.001	(0.001)
Grade	9.010	(7.804)	0.277	(0.060)
Academic				
clas+math	76.866	(53.524)	-4.845	(0.844)
clas	-141.766	(47.669)	-3.316	(0.444)
math	-12.592	(49.785)	-3.273	(0.369)
other	-225.277	(49.520)	-1.956	(0.281)
x grade	-22.407	(7.063)	0.068	(0.086)
Middle-theoretical				
math	102.125	(50.984)	-3.397	(0.465)
other	-181.509	(44.300)	-1.923	(0.245)
x grade	-7.705	(5.000)	-0.117	(0.080)
Middle-practical	-25.446	(43.907)	-1.648	(0.335)
x grade	-21.545	(4.873)	-0.218	(0.083)
Vocational	112.558	(46.284)	-4.248	(0.530)
Part-time	270.638	(31.720)		

Note: Estimates of a sample of 5,158 students or 33,239 student-year observations. Scale = minutes of daily travel time. Grade variable starts counting in high school. The marginal costs in the model are a nonparametric function of state variables, this table summarizes them by regressing their logarithmic transformation on the same variables that enter the fixed costs. Bootstrap standard errors in parentheses.

Table A16: Costs of schooling: student characteristics and elective courses

	Fixed costs				Log of marginal costs			
	Interaction with classical languages		Interaction with intensive math		Interaction with classical languages		Interaction with intenstive math	
Male	-4.097	(8.294)	-50.776	(11.655)	0.527	(0.357)	0.866	(0.269)
Language ability	-57.070	(11.140)	29.557	(13.243)	-1.113	(0.410)	-0.824	(0.343)
Math ability	-21.824	(9.085)	-71.002	(14.949)	-0.479	(0.402)	0.542	(0.316)
High SES	-36.216	(9.500)	-25.388	(11.322)	-0.483	(0.391)	0.263	(0.295)
Type 2	83.543	(12.452)	35.493	(11.792)	-0.186	(0.428)	-0.022	(0.331)

Note: Estimates of a sample of 5,158 students or 33,239 student-year observations. Scale = minutes of daily travel time. The marginal costs in the model are a flexible function of state variables, this table summarizes them by regressing their logarithmic transformation on the same variables that enter the fixed costs. Ability measured in standard deviations. Type 2 = dummy equal to one if student belongs to unobserved type 2 instead of 1. High SES= at least one parent has higher education degree. Clas= classical languages included. Math= intensive math. Bootstrap standard errors in parentheses.

Table A17: Type probabilities in %

	Type probabilities	
	Type 1	Type 2
Overall	*29.55*	*70.45*
Age 12	33.07	66.93
Age 13	9.86	90.14
Age 14	10.67	89.33

Note: Estimates of unobserved types in the student population by age they start high school.

Table A18: Performance thresholds

	Performance threshold for outcome	
Increase to obtain outcome 3	0.871	(0.036)
Increase to obtain outcome 4	1.102	(0.043)
Increase to obtain outcome 5	1.744	(0.054)

Note: Optimal y is specific for each grade-track and thresholds for avoiding lowest outcome in them are normalized to 0. These differences are estimated but constrained to be the same over grades and tracks. Constraints on thresholds are used to avoid impossible outcomes because of institutional context. Grade 7-10 allow more than two realizations of main performance outcome.Bootstrap standard errors in parentheses.

Table A19: Performance elective courses

		Performance	
Log effective study effort ($\ln y$)			
	x clas	0.902	(0.339)
	x math	0.124	(0.102)
Male			
	x clas	0.931	(0.469)
	x math	0.058	(0.205)
Language ability			
	x clas	-0.251	(0.524)
	x math	0.053	(0.171)
Math ability			
	x clas	-0.785	(0.532)
	x math	0.488	(0.196)
SES			
	x clas	-0.424	(0.352)
	x math	0.027	(0.193)
Type 2			
	x clas	0.332	(0.517)
	x math	0.392	(0.223)
Cut points clas			
	x grade	-0.096	(0.138)
	x constant	2.222	(2.062)
Cut points math			
	x grade 2, outcome 2	-4.679	(0.624)
	x grade 2, outcome 3	-3.540	(0.556)
	x grade 3, outcome 2	-5.222	(1.313)
	x grade 3, outcome 3	-3.177	(0.574)
	x grade 4, outcome 2	-5.676	(3.625)
	x grade 4, outcome 3	-1.603	(0.528)

Note: Bootstrap standard errors in parentheses.

Table A20: Value of obtaining degree

	Degree values	
High school degree	512.195	(101.307)
x level	108.083	(60.670)
x vocational	-146.819	(99.334)
12th grade certificate vocational track	529.825	(65.461)

Note: Estimates of μ^{degree}. Scale = minutes of daily travel time. Level = academic level of high school program (0-3). Bootstrap standard errors in parentheses.

Table A21: Estimation results higher education (1)

	Higher education			
	Enrollment		Degree	
Male	-1.013	(0.108)	-0.840	(0.129)
x HE level	0.302	(0.206)	0.529	(0.246)
x STEM	0.827	(0.067)	0.598	(0.142)
Language ability	0.340	(0.126)	0.262	(0.144)
x HE level	2.130	(0.230)	0.641	(0.255)
x STEM	-0.186	(0.088)	-0.172	(0.136)
Math ability	0.111	(0.109)	0.611	(0.137)
x HE level	1.433	(0.229)	0.149	(0.335)
x STEM	0.472	(0.098)	-0.154	(0.147)
SES	0.563	(0.126)	0.633	(0.136)
x HE level	1.875	(0.191)	0.643	(0.245)
x STEM	0.084	(0.084)	-0.078	(0.129)
Type 2	-0.613	(0.157)	-1.830	(0.174)
x HE level	-4.741	(0.248)	0.901	(0.310)
x STEM	-0.639	(0.090)	1.006	(0.165)

Note: Estimates of higher education outcomes as specified in Appendix D. HE Level = level of higher education (average ability). Bootstrap standard errors in parentheses.

Table A22: Estimation results higher education (2)

	Higher education			
	Enrollment		Degree	
Academic degree	4.227	(0.266)		
+ clas + math			-0.330	(0.385)
+ clas			-0.009	(0.278)
+ math			0.501	(0.198)
other			benchmark	
Middle-theoretical degree	3.426	(0.205)		
+ math			-0.195	(0.277)
other			-0.442	(0.160)
Middle-practical degree	2.050	(0.161)	-0.721	(0.233)
Vocational degree	benchmark		-2.030	(0.332)
Study delay	0.182	(0.150)	-0.580	(0.271)
High school level x study delay	-0.326	(0.077)	-0.150	(0.121)
HE level				
x high school level	0.163	(0.186)	-0.227	(0.266)
x clas	3.031	(0.296)	1.618	(0.283)
x math	2.596	(0.222)	1.313	(0.258)
x study delay	-0.312	(0.225)	0.034	(0.368)
STEM				
x high school level	-0.455	(0.062)	-0.108	(0.114)
x clas	-0.271	(0.123)	0.480	(0.193)
x math	1.335	(0.086)	0.389	(0.158)
x study delay	-0.234	(0.083)	0.165	(0.148)

Note: Estimates of higher education outcomes as specified in Appendix D. Clas= classical languages included. Math= intensive math. High school level = academic level of high school program (0-3). HE Level = level of higher education (average ability). Bootstrap standard errors in parentheses.

Table A23: Estimation results higher education (3)

	Higher education			
	Enrollment		Degree	
Distance (km)	-0.018	(0.001)	-0.003	(0.001)
Same HE level as enrollment			1.715	(0.112)
Same major as enrollment			2.411	(0.088)
Upgrade HE level			-1.653	(0.283)
University	-3.957	(0.398)	-4.604	(0.447)
Academic college	-2.842	(0.339)	-2.971	(0.317)
Professional college	-1.287	(0.274)	-1.338	(0.177)
STEM	0.289	(0.171)	-1.059	(0.298)

Note: Estimates of higher education outcomes as specified in Appendix D. HE Level = level of higher education (average ability). Bootstrap standard errors in parentheses.

Table A24: OLS regressions on initial conditions and counterfactuals

	Study delay (%)		**High school dropout** (%)		**Higher education graduation** (%)		**Student welfare** ($1000)	
Male	5.53	(1.41)	8.31	(1.02)	-12.40	(1.01)	-6.13	(1.67)
Language ability	-6.76	(1.24)	-7.49	(0.93)	15.38	(1.52)	16.27	(2.29)
Math ability	-1.83	(1.22)	-6.97	(0.86)	14.98	(1.25)	11.58	(1.65)
High SES	-6.46	(1.56)	-4.56	(0.74)	16.60	(1.36)	17.95	(3.02)
Type 2	11.82	(1.44)	14.36	(0.89)	-32.63	(1.32)	-38.20	(5.21)
Constant	24.01	(1.35)	1.34	(0.59)	68.53	(1.33)	70.42	(8.71)
Repeat policy	8.78	(0.73)	1.88	(0.35)	-2.04	(0.27)	-1.76	(0.25)
x male	0.85	(0.69)	1.63	(0.44)	0.05	(0.20)	-0.37	(0.12)
x language ability	-0.37	(0.63)	-1.75	(0.44)	-0.07	(0.18)	0.42	(0.17)
x math ability	-1.38	(0.69)	-0.28	(0.48)	-0.05	(0.20)	0.35	(0.16)
x high SES	-0.72	(0.74)	-1.53	(0.42)	0.16	(0.27)	0.42	(0.17)
x Type 2	0.69	(0.75)	2.38	(0.46)	0.39	(0.22)	-0.45	(0.16)
Downgrade policy	-11.91	(1.02)	-0.20	(0.22)	-0.67	(0.30)	-1.47	(0.23)
x male	-0.91	(0.70)	-1.41	(0.31)	0.30	(0.17)	-0.08	(0.12)
x language ability	-0.05	(0.68)	0.01	(0.29)	0.22	(0.13)	-0.07	(0.11)
x math ability	-0.26	(0.80)	0.95	(0.32)	-0.23	(0.14)	-0.15	(0.12)
x high SES	1.85	(0.88)	0.29	(0.32)	0.13	(0.23)	0.22	(0.17)
x Type 2	2.91	(0.97)	-1.14	(0.30)	0.26	(0.20)	0.62	(0.18)

Note: OLS regression of predictions from the dynamic model. B-certificate = students acquired skills to proceed to next grade but only if they downgrade, i.e. switch to track of lower academic level or drop elective course. Status quo = students can choose to downgrade or repeat grade after obtaining B-certificate, Repeat = students must repeat grade after obtaining B-certificate, Downgrade = students must downgrade and not repeat grade after obtaining B-certificate. Ability measured in standard deviations. Type 2 = dummy equal to one if student belongs to unobserved type 2 instead of 1. High SES= at least one parent has higher education degree. Opportunity cost of time: $10/h. Bootstrap standard errors in parentheses.