# Online Appendix of "Modeling the Effects of Grade Retention in High School"

Bart Cockx* Matteo Picchio† Stijn Baert‡

August 20, 2018

## A Proof of Proposition 1

**Proof.**

*Identification of the unknown components in the first schooling year*

First, observe that, conditional on the age at the start of high school $in_i$,[1] the first high school outcome, i.e. track choice $Y^*_{i01} \equiv tr_{i1} = \mu(\mathbf{x}_i, z_i, in_i; \theta_{01}) + u_{i01}$ is free of selection and can, hence, be considered as a "measurement". The track choice can, however, be selective in the sense that it can be related to the first end-of-year evaluation $Y^*_{i11} = ev_{i1} = \mu(\mathbf{x}_i, tr_{i1}, in_i; \theta_{11}) + u_{i11}$ through the common unobserved determinant $v_{i1}(7)$, which induces dependence between $u_{i01}$ and $u_{i11}$.[2] By the presence of the continuous variable $z_i$ that is excluded from all the other outcome equations (condition 1 in Proposition 1),[3] one can vary track choice and, hence, $u_{i01}$, independently of the end-of-year evaluation in grade 7 and, therefore, independently of $u_{i11}$. As shown by Theorem 1 of Carneiro et al. (2003), this independent variation identifies the joint distribution of $(u_{i01}, u_{i11})$ non-parametrically (up to scale), and the corresponding unobserved threshold parameters of the ordered choice models $\alpha_{j,0}$ ($j \in \{1, ...4\}$) and $\alpha_{k,1}$ ($k \in \{1, 2\}$).[4] Key is that the error terms $u_{i01}$

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

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

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

[1] See further discussion of this point in Section 4.3

[2] By Assumption 3.1, $u_{i01} = \delta_{07} v_{i1}(7) + \epsilon_{i01}$ and $u_{i11} = \delta_{17} v_{i1}(7) + \epsilon_{i11}$, where $\delta_{17} = 1$ by normalization.

[3] Alternatively, in the absence of an exclusion restriction, Carneiro et al. (2003) show that it is sufficient to have two components of $\mathbf{x}$ that vary continuously over $\mathbb{R}$ (condition 2 in our Proposition 1).

[4] Theorem 1 actually assume that the discrete measurements are binary valued and not ordered choices. However, on p. 376 the authors state that the extension to censored random variables, i.e. ordered choices, is straightforward.

and $u_{i11}$ (and the error terms of all other outcomes)[5] do not depend on track choice. This means that there cannot be any higher order dependence that is not captured by the conditional mean, and that the factor loading $\delta_{11}$ on the unobserved determinant $v_{i1}(7)$ cannot depend on $tr_i$, i.e. this excludes *essential heterogeneity* in the effect of track choice.

Based on this result, identification of the unknown components of the error terms $u_{i01}$ and $u_{i11}$ is shown in three steps. First, since by the aforementioned argument the joint distribution of the error terms of the first two outcomes is identified, we can form the following ratio of cross moments (where $\delta_{17} = 1$ by normalization):[6]

$$\frac{E\left[u_{i01}^3 u_{i11}\right]}{E\left[u_{i01} u_{i11}^3\right]} = \frac{\delta_{07}^3 E\left[v_{i1}(7)^4\right]}{\delta_{07} E\left[v_{i1}(7)^4\right]} = \delta_{07}^2 \tag{A-1}$$

using that $u_{i01} = \delta_{07} v_{i1}(7) + \epsilon_{i01}$, $u_{i11} = \delta_{17} v_{i1}(7) + \epsilon_{i11}$, $\delta_{07} \neq 0$, Assumptions 1.2 and 3.2, and the existence of the fourth order moment of the distribution of $v_{i1}(7)$. This identifies $\delta_{07}$ apart from its sign. The sign of $\delta_{07}$ corresponds to the sign of $E\left[u_{i01} u_{i11}\right] = \delta_{07} E\left[v_{i1}(7)^2\right]$, since $E\left[v_{i1}(7)^2\right] > 0$ and finite.

Second, from the the higher order cross moments of the residuals of the two outcome equations, we can can recover all moments of the the unobserved determinants $v_{i1}(7)$: $\forall k > 0$ : $E\left[(\delta_{07} v_{i1}(7) + \epsilon_{i01})^k \ (\delta_{17} v_{i1}(7) + \epsilon_{i11})\right] = \delta_{07}^k E\left(v_{i1}(7)^{k+1}\right)$. Since $\delta_{07}$ is already identified and since a distribution of which all moments are finite (Assumption (iv) in Proposition 1) can be completely characterized by these moments (Billingsley, 1995), this non-parametrically identifies the distribution of $v_{i1}(7)$.

Finally, we can form all the higher order moments of the error terms of each of the two outcome equations, $\forall k > 1 : E\left[(\delta_{07} v_{i1}(7) + \epsilon_{i01})^k\right] = \delta_{07}^k E(v_{i1}^k) + E(\epsilon_{i01}^k)$ and $E\left[(\delta_{11} v_{i1}(7) + \epsilon_{i11})^k\right] = E(v_{i1}(7)^k) + E(\epsilon_{i11}^k)$. Since the first terms of the sum on the right hand-side are already identified, the second terms can be recovered. This enables non-parametric identification of the distribution of $\epsilon_{i01}$ and $\epsilon_{i11}$.

In first period the next outcome is the decision to repeat the grade or not ($Y_{i31} \equiv re_{i1}$) for stu-

Note also that FNT can prove identification of the joint distribution of these error terms without exclusion restriction and without the presence of continuous explanatory variables, because they have continuous outcome variables, i.e. test scores. In the case of continuous outcomes the joint distribution of the error terms can be identified by constructing all (cross) moments of the residuals in the outcome equations. In the case of discrete outcomes, these residuals are "latent", so that their cross moments cannot be directly formed and, hence, stronger identifying assumptions are required.

[5] Further on in the proof we repeatedly use Theorem 1 of Carneiro et al. (2003) to prove identification of the joint distribution of $u_{i01}$ and the error terms $u_{ict}$ of each of the other outcomes for $c \in \{1, \ldots, 4\}$ and $t \in \{1, \ldots T_i\}$.

[6] Instead of forming the ratio of fourth order moments FNT consider the ratio of third order moments. This identification argument works only if $E\left[v_{i1}(7)^3\right] \neq 0$ (and, hence only for asymmetric distributions), because this third moment appears then in the denominator of the ratio. In their Appendix B FNT relax this asymmetry assumption in the case of having measurements of more test scores per student, also using fourth moments. Our identification strategy is inspired by a combination of the arguments mentioned in their main text and in their Appendix B.

dents for whom $Y_{i11} \equiv ev_{i1} = B$. This is because nobody drops out in the seventh grade ($out_{i1} \equiv Y_{i21} = 0$).[7] Using again Theorem 1 of Carneiro et al. (2003) $\theta_{31}$, the associated threshold parameters, and the joint distribution of $(u_{i01}, u_{i31})$ are identified. The latter allows us to form the following cross moment: $E[u_{i01}u_{i31}] = E[(\delta_{07}v_{i1}(7) + \epsilon_{i01})(\delta_{37}v_{i1}(7) + \epsilon_{i31})] = \delta_{07}\delta_{37}E\left(v_{i1}(7)^2\right)$. $\delta_{37}$ is identified, because $\delta_{07} \neq 0$ and $E(v_{i1}(7)^2) \neq 0$ already are. As before, the higher order moments of the third outcome equation then identifies the distribution of $\epsilon_{i31}$. Following the same argument $\theta_{41}$, the associated threshold parameters and the distribution of $\epsilon_{i41}$ are identified as well.

*Identification of the unknown components beyond the first schooling year*

As from period 2 some pupils may have been retained. This means that $u_{ic2}$ depends on $re_{i1}$ through $\delta_{c8}(re_{i1})$ (Assumption 3.1). A consequence is that to identify $\delta_{c8}(re_{i1})$ we have to condition on two sub-populations: the population that has been retained in the previous year ($re_{i1} = 1$) and the one that has not been retained ($re_{i1} = 0$). This is possible (cf. next paragraph) because the selection into retention occurs through dependence on observables and *past* unobservables that have already been identified, while the distribution of the new unobserved persistent shock in grade 8 $v^*_{i1}(8)$ can be identified from the cross moments between the unobservables of individuals who are retained in grade 8, but not in grade 7.

First, consider the error of outcome $c$ for an individual $i$ who is retained in grade 7 ($re_{i1} = 1$). From Assumptions 3.1 and 3.2 we obtain: $u_{ic2} = \delta_{c7}(1)v_{i1}(7) + \epsilon_{ic2}$. We can then apply, as in period 1, Theorem 1 of Carneiro et al. (2003) to prove that the joint conditional distribution $(u_{i01}, u_{ic2})|re_{i1} = 1, \theta_{c2}$ for $re_{i1} = 1$ and the threshold parameters of the corresponding ordered choice are non-parametrically identified. Hence, we can form the following conditional cross moment: $E(u_{i01}, u_{ic2}|re_{i1} = 1) = \delta_{07}\delta_{c7}(1)E\left(v_{i1}(7)^2\right)$. Since $\delta_{07}$ and the distribution of $v_{i1}(7)$ is already identified, this cross moment identifies $\delta_{c7}(1)$.

Second, consider the error of outcome $c = 1$ for an individual who is not retained in grade 7 ($re_{i1} = 0$): $u_{i12} = \delta_{18}(0)\left(\delta^*_7(8)v_{i1}(7) + v^*_{i1}(8)\right)$. Noticing that $\delta_{18}(0) = 1$ by normalization (Assumption 3.1), we can follow a similar argument as in the previous paragraph to show that $\delta^*_7(8)$, $\theta_{12}$ for $re_{i1} = 0$ and the threshold parameters of the corresponding ordered choice. Once $\delta^*_7(8)$ is identified, we can follow a same strategy for outcomes $c \neq 1$ to identify $\delta_{c8}(0)$, $\theta_{c2}$ for $re_{i1} = 0$ and the corresponding threshold parameters.

In order to identify the distribution of the new persistent shock $v^*_{i1}(8)$, we consider the error of outcome $c$ for an individual who is retained in grade 8, but not in grade 7: $u_{ic3} = \delta_{c8}(0\ 1)\,(\delta^*_7(8)\, v_{i1}(7) + v^*_{i1}(8))$. Following a similar argument as in the previous paragraphs we can first form the following cross moments: $E\left(u_{i01}u_{ic3}|re_{i1} = 0, re_{i2} = 1\right) = \delta_{07}\delta_{c8}(0\ 1)\delta^*_7(8)E\left(v_{i1}(7)^2\right)$. Since

[7] In our data the first grade repetition occurs only in grade 8, so that this issue starts only as of period 3. We ignore this here, to demonstrate that identification does not hinge on this particularity. In this case the next outcome in the first period would rather be $dow_{i1} \equiv Y_{i41}$.

$\delta_{07}$, $\delta_7^*(8)$ and the distribution of $v_{i1}$ are already identified these identify $\delta_{c8}(0\ 1)$. Next we form the cross moments of the error terms of in grade 8 of individuals who have been retained in that grade (but not in grade 7): $E\left(u_{ic2}^k u_{ic3} | re_{i1} = 0, re_{i2} = 1\right) = \delta_{c8}(0\ 1)\delta_7^*(8)^{k+1} E\left(v_{i1}(7)^{k+1}\right) + \delta_{c8}(0\ 1) E\left(v_{i1}^*(8)^{k+1}\right)$ for $k > 0$. Since only $E\left(v_{i1}^*(8)^{k+1}\right)$ is the only unknown in this expression, these cross moments identify the distribution of $v_{i1}^*(8)$.

We can proceed in a similar way sequentially over time periods, outcomes, and retention histories until we arrive at the end of the observation period to identify to full joint distribution of grade-varying unobserved heterogeneity $\mathbf{v}_i$, all $\theta_{ct}$ and associated threshold parameters. ■

# B Partial Observability of Track Choices at the Start of High School

In Subsection 4.5 we explained that we do not observe the chosen track at the beginning of high school, i.e. $tr_{i1}$. Mroz et al. (2016) solve this partial observability by considering the marginal likelihood function instead of the conditional one, where the unobserved information is integrated out of the likelihood. Here we follow a similar approach by summing the likelihood over the possible initial track choices at the start of high school. As in Mroz et al. (2016), we take prior information into account to restrict the potential number of initial track choices $tr_{i1}$ over which we sum the likelihood. The following prior information is considered: (i) the initial track choice of pupils starting in the vocational track, i.e. for whom $tr_{i1} = 1$, is observed; (ii) the track choice is known in all grades beyond grade 7; (iii) no student is retained in grade 7 ($re_{i1} = 0$); (iv) as all students are younger than 18 years old in grade 7, no student drops out high school in this grade ($out_{i1} = 0$); (v) students can only downgrade ($0 \leq dow_{i1}$) and if they do, they do not downgrade more than two tracks in a single year ($dow_{i1} \leq 2$). In this appendix, we show how the marginal likelihood function that accommodates for the partial observability of $tr_{i1}$ can be adjusted to take this prior information into account.

First, to focus on the main issues, we simplify the notation. We ignore in the joint distribution function as expressed by Equation (1) in Section 4.1 the subscripts and the conditioning on $in_i$ and the observed and unobserved covariates:

$$D(tr_1, \mathbf{Y}) = D(tr_1) D(ev_1, dow_1 | tr_1) D(Y_2 \ldots Y_T | ev_1, dow_1, tr_1) \qquad \text{(A-2)}$$

where $D(.)$ and $D(.,.|.)$ respectively denote the marginal and (joint) conditional distributions of their arguments and where we recall that $Y_1 = [ev_1\ 0\ 0\ dow_1] \equiv [ev_1\ dow_1]$, because, by (iv), nobody drops out in grade 7 ($out_1 = 0$), and, by (iii), nobody is retained in grade 7 ($re_1 = 0$). $D(Y_2 \ldots Y_T | ev_1, dow_1, tr_1)$ in (A-2) (and also $D(dow_1 + tr_2, dow_1 | ev_1, tr_2)$ in (A-4) below) are conditional on $out_1 = 0$ and $re_1 = 0$. However, in order to avoid burdensome notation we leave this conditioning implicit.

In order to take partial observability of the initial track choice into account, we should sum the joint distribution in Equation (A-2) over $tr_1$ for all pupils who are not in the vocational track, i.e. for whom $tr_1 > 1$. However, given the available prior information, the sum should not be over all four unknown tracks ($2 \leq tr_1 \leq 5$). Recall that by prior information (ii) the track choice in grade 8 ($tr_2$) is known. Together with the fact (v) that pupils can only downgrade, and if they downgrade, they can downgrade at most two tracks ($0 \leq dow_{i1} \leq 2$), the initial track choice $tr_1$ is restricted, depending on the track choice $tr_2$ observed in grade 8.

To see more clearly how the prior information (ii) and (v) restricts the number of tracks over which the joint distribution (A-2) is summed, note first that we do not observe $dow_t$ in grade $t$ directly, but we infer it from the tracks in which pupils are observed in each year beyond grade 7:[8]

$$dow_t = tr_t - tr_{t+1} \tag{A-3}$$

This equation establishes a one-to-one relationship between $tr_1$ and $dow_1$ for any given value of $tr_2$: $tr_1 = dow_1 + tr_2$. This means that if we condition (A-2) on the known value of $tr_2$,[9] summing this equation over the unknown $tr_1$ is equivalent to summing it over the unknown $dow_1$. The advantage of summing it over $dow_1$ is that we can easily impose the prior information that both $0 \leq dow_{i1} \leq 2$ and ($2 \leq tr_1 = dow_1 + tr_2 \leq 5 \Leftrightarrow 2 - tr_2 \leq dow_1 = tr_1 - tr_2 \leq 5 - tr_2$) by setting $\max\{0, 2 - tr_2\} \leq dow_1 \leq \min\{2, 5 - tr_2\}$. In case $tr_2 = 1$, $dow_1 > 0$, because $tr_1 \neq 1$, as we observe the track choice for individuals in VHS at the start, i.e. for $tr_1 = 1$.

In order to take the partial observability for $tr_1 > 1$ into account, we therefore consider Equation (A-2) given $tr_2$, replace $tr_1$ by $dow_1 + tr_2$ and sum it, instead of over $tr_1$, over all possible downgrading decisions $dow_1$, taking the prior information into account:

$$\sum_{dow_1=\max\{0,2-tr_2\}}^{\min\{2,5-tr_2\}} D(dow_1 + tr_2, \mathbf{Y}) =$$

$$\sum_{dow_1=\max\{0,2-tr_2\}}^{\min\{2,5-tr_2\}} D(dow_1 + tr_2) D(ev_1, dow_1 | dow_1 + tr_2)$$

$$\times D(Y_2 \ldots Y_T | dow_1 + tr_2, ev_1, dow_1). \tag{A-4}$$

The sample log-likelihood function in Equation (4) in the main text is modified along these lines.

[8] As no student is retained in grade 7, we observe all track choices for $t > 1$ and $g > 7$.

[9] Note that if both $tr_1$ and $dow_1$ are known, (A-3) implies that $tr_2$ is irrelevant, because it does not add any new information. By contrast, if neither $tr_1$ nor $dow_1$ are known, as in the case of partial observability, $tr_2$ matters, because it adds in new information. That is why it appears when summing over $dow_1$ (which is equivalent to summing over $tr_1$) in Equation (A-4) below, while it is absent in (A-2).

# C The Empirical Specification of the Educational Choices

As mentioned in Subsection 4.4 we assume that all educational choices can be specified as (ordered) logits. As discussed in Subsection 4.3, this is not strictly required for identification. In the following subsections we first describe in detail for each schooling outcome the model specification choices. In the final subsection we discuss the specification of the joint unobserved heterogeneity distribution $G(\mathbf{v}_{i1};\boldsymbol{\rho})$.

## C.1 The Track Choice at the Start of High School

The track choice takes value on $\{\mathrm{VHS}, \mathrm{THS}^-, \mathrm{THS}^+, \mathrm{GHS}^-, \mathrm{GHS}^+\}$, which we relabel for notational convenience and to underline their hierarchical ordering by $\{1,2,3,4,5\}$. The probability density function of the track choice is an ordered logit determined by a linear index in the strictly exogenous observed explanatory variables $\mathbf{z}_i \equiv [z_i\ \mathbf{x}_i]$, including the "instrument" $z_i$, i.e. the day of birth, which is excluded from all other schooling outcome equations, the unobservable determinants of track choice $\mathbf{v}_{i,tr} \equiv \mathbf{v}_{i0} \equiv v_{i0}(7) = \delta_{tr} v_{i1}(7)$, where $\delta_{tr} \equiv \delta_0$ is the loading factor of the unobserved heterogeneity, and the initial delay $in_i$:

$$
\begin{aligned}
\Pr(tr_{i1}=1|\mathbf{z}_i,\mathbf{v}_{i,tr},in_i) &= \Lambda(\alpha_{1,tr}-\mathbf{v}_{i,tr}-\mathbf{z}_i'\boldsymbol{\beta}_{tr}-in_i\gamma_{tr}),\\
\Pr(tr_{i1}=2|\mathbf{z}_i,\mathbf{v}_{i,tr},in_i) &= \Lambda(\alpha_{2,tr}-\mathbf{v}_{i,tr}-\mathbf{z}_i'\boldsymbol{\beta}_{tr}-in_i\gamma_{tr})-\Lambda(\alpha_{1,tr}-\mathbf{v}_{i,tr}-\mathbf{z}_i'\boldsymbol{\beta}_{tr}-in_i\gamma_{tr}),\\
\Pr(tr_{i1}=3|\mathbf{z}_i,\mathbf{v}_{i,tr},in_i) &= \Lambda(\alpha_{3,tr}-\mathbf{v}_{i,tr}-\mathbf{z}_i'\boldsymbol{\beta}_{tr}-in_i\gamma_{tr})-\Lambda(\alpha_{2,tr}-\mathbf{v}_{i,tr}-\mathbf{z}_i'\boldsymbol{\beta}_{tr}-in_i\gamma_{tr}),\\
\Pr(tr_{i1}=4|\mathbf{z}_i,\mathbf{v}_{i,tr},in_i) &= \Lambda(\alpha_{4,tr}-\mathbf{v}_{i,tr}-\mathbf{z}_i'\boldsymbol{\beta}_{tr}-in_i\gamma_{tr})-\Lambda(\alpha_{3,tr}-\mathbf{v}_{i,tr}-\mathbf{z}_i'\boldsymbol{\beta}_{tr}-in_i\gamma_{tr}),\\
\Pr(tr_{i1}=5|\mathbf{z}_i,\mathbf{v}_{i,tr},in_i) &= 1-\Lambda(\alpha_{4,tr}-\mathbf{v}_{i,tr}-\mathbf{z}_i'\boldsymbol{\beta}_{tr}-in_i\gamma_{tr}),
\end{aligned}
\tag{A-5}
$$

where $\delta_{tr}$, $\boldsymbol{\beta}_{tr}$ and $\gamma_{tr}$ are parameters to be estimated, the coefficients $\alpha_{1,tr} < \alpha_{2,tr} < \alpha_{3,tr} < \alpha_{4,tr}$ are the ordered threshold parameters and $\Lambda(.)$ denotes the logistic distribution.

## C.2 The End-of-Year Evaluation

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

$$\begin{aligned}
\Pr(ev_{it}=C|\mathbf{x}_i,\mathbf{v}_{i,ev},in_i,tr_{i1},\Im_{it-1}) &= \Lambda\big[\alpha_{1,ev}-\phi_{ev}(\mathbf{x}_i,\mathbf{v}_{i,ev},in_i,tr_{i1},\Im_{it-1})\big],\\
\Pr(ev_{it}=B|\mathbf{x}_i,\mathbf{v}_{i,ev},in_i,tr_{i1},\Im_{it-1}) &= \Lambda\big[\alpha_{2,ev}-\phi_{ev}(\mathbf{x}_i,\mathbf{v}_{i,ev},in_i,tr_{i1},\Im_{it-1})\big]\\
&\quad -\Lambda\big[\alpha_{1,ev}-\phi_{ev}(\mathbf{x}_i,\mathbf{v}_{i,ev},in_i,tr_{i1},\Im_{it-1})\big],\\
\Pr(ev_{it}=A|\mathbf{x}_i,\mathbf{v}_{i,ev},in_i,tr_{i1},\Im_{it-1}) &= 1-\Lambda\big[\alpha_{2,ev}-\phi_{ev}(\mathbf{x}_i,\mathbf{v}_{i,ev},in_i,tr_{i1},\Im_{it-1})\big],
\end{aligned} \tag{A-6}$$

where $\phi_{ev}(\mathbf{x}_i,\mathbf{v}_{i,ev},\ in_i,tr_{i1},\Im_{it-1})$ is a linear index in its arguments, capturing the impact of observed and unobserved determinants and past educational choices of pupils and teachers.

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

$$\begin{aligned}
\phi_{ev}(\mathbf{x}_i,\mathbf{v}_{i,ev},in_i,tr_{i1},\Im_{it-1}) &= \mathbf{x}_i'\boldsymbol{\beta}_{ev}+\boldsymbol{\iota}_{it}'\mathbf{v}_{i,ev}+in_i\gamma_{ev}+\mathbf{Idow}_{it-1}'\boldsymbol{\pi}_{ev}+\mathbf{Itr}_{it}'\boldsymbol{\eta}_{ev}\\
&+ \mathbf{Igr}_{it}'\boldsymbol{\delta}_{ev}+re_{it-1}\kappa_{ev}+pre_{it-1}\psi_{ev}\\
&\equiv \mathbf{w}_{it}'\boldsymbol{\xi}_{ev}+re_{it-1}\kappa_{ev}+pre_{it-1}\psi_{ev},
\end{aligned} \tag{A-7}$$

where

- $\mathbf{v}_{i,ev}\equiv\mathbf{v}_{i1}$ is the vector of unobserved heterogeneity affecting the evaluation;
- $\boldsymbol{\iota}_{it}'=[1\ \mathbf{Igr}_{it}']$ is a $1\times 7$ indicator vector selecting the unobserved heterogeneity component associated to the grade in which individual $i$ is in period $t$ (taking grade 7 as the reference);
- $\mathbf{Igr}_{it}=\Big[\mathbf{1}_{\{8\}}\left(6+t-\mathbf{1}_{\{\forall t:t>1\}}(t)\sum_{s=1}^{t-1}re_{is}\right)...\mathbf{1}_{\{13\}}\left(6+t-\mathbf{1}_{\{\forall t:t>1\}}(t)\ \ \sum_{s=1}^{t-1}re_{is}\right)\Big]'$ is a column vector of six indicators of the grade at the beginning of the $t$-th year in high school and where grade 7 is the reference grade;
- $\mathbf{Idow}_{it-1}=[\mathbf{1}_1(dow_{it-1})\ \mathbf{1}_2(dow_{it-1})]'$ is a column vector of two indicators that determine whether the student chooses to downgrade one or two tracks at the end of the previous academic year (the reference student does not change track): $\mathbf{1}_A(x)$ defines the indicator function that is equal to one if $x\in A$ and zero otherwise; $dow_{it}\in\{0,1,2\}$ indicates the number of tracks that individual $i$ chooses to downgrade at the end of year $t$ in high school;
- $\mathbf{Itr}_{it}=\Big[\mathbf{1}_{\{2\}}\left(tr_{i1}-\mathbf{1}_{\{\forall t:t>1\}}(t)\sum_{s=1}^{t-1}dow_{is}\right)...\mathbf{1}_{\{5\}}\left(tr_{i1}-\mathbf{1}_{\{\forall t:t>1\}}(t)\sum_{s=1}^{t-1}dow_{is}\right)\Big]'$ is a column vector of four indicators of the track chosen at the beginning of the $t$-th year in high school and where VHS ($tr_{i1}=1$) is the reference track;

- $re_{it-1}$ is an indicator variable equal to one if individual $i$ was retained at the end of the previous academic year $t-1$ (repeating the grade therefore in the current year $t$) and zero otherwise;
- $pre_{it-1} = \mathbf{1}_{\mathbb{N}_0}\left(\mathbf{1}_{\{\forall t:t>2\}}(t)\sum_{s=1}^{t-2} re_{is}\right)$ is an indicator equal to one if the student has ever repeated a grade in high school in years prior to the $(t-1)^{th}$ year.
- $\mathbf{w}'_{it} \equiv \left[\mathbf{x}'_i \ \boldsymbol{\iota}'_{it} \ in_i \ \mathbf{Idow}'_{it-1} \ \mathbf{Itr}'_{it} \ \mathbf{Igr}'_{it}\right]$ and $\boldsymbol{\xi}'_{ev} = \left[\boldsymbol{\beta}'_{ev} \ \mathbf{v}'_{i,ev} \ \gamma_{ev} \ \boldsymbol{\pi}'_{ev} \ \boldsymbol{\eta}'_{ev} \ \boldsymbol{\delta}'_{ev}\right]$

where $\boldsymbol{\xi}'_{ev}$, $\kappa_{ev}$ and $\psi_{ev}$ are parameters to be estimated.[10] Hence, we allow that past high school choices ($\Im_{it-1}$) affect the evaluations in a flexible way. The coefficient $\kappa_{ev}$ is the transitory effect of grade repetition on the subsequent academic performance, while $\psi_{ev}$ is the permanent effect. $\boldsymbol{\delta}_{ev}$ and $\boldsymbol{\eta}_{ev}$ capture that students' ability to get good evaluations depends on the current grade and track, respectively. Finally, $\boldsymbol{\pi}_{ev}$ is the (transitory) effect of having downgraded a track on the academic achievement in the subsequent year.

In the second specification of the linear index $\phi_{ev}(\cdot)$ we allow the short- and long-run effects of grade repetition to be heterogeneous in observed and unobserved abilities. To maintain a tractable model, we simplify by interacting $re_{it-1}$ and $pre_{it-1}$ with the linear index defined in Equation (A-7). Since this linear index is also a function of unobservables, this allows for *essential* heterogeneity in the treatment effect of grade repetition. The specification of the linear index is then given by the following expression:

$$\begin{aligned}\phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_{i1}, \Im_{it-1}) &= \mathbf{w}'_{it}\boldsymbol{\xi}_{ev}\left(1 + re_{it-1}\kappa^0_{ev} + pre_{it-1}\psi^0_{ev}\right)\\ &+ re_{it-1}\kappa_{ev}(1+pre_{it-1}\psi^0_{ev}) + pre_{it-1}\psi_{ev}(1+re_{it-1}\kappa^0_{ev}),\end{aligned} \tag{A-8}$$

where $\kappa^0_{ev}$ and $\psi^0_{ev}$ are parameters to be estimated. If $\kappa^0_{ev}$ and $\psi^0_{ev}$ are jointly equal to $0$, then we go back to the first specification of the linear index in Equation (A-7).

In the last two grades or if a student is in the VHS track, the evaluation is dichotomous, either A or C. In these cases, the ordered logit model described in Equation (A-6) collapses to a logit model with:

$$\Pr(ev_{it} = A|\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_{i1}, \Im_{it-1}) = 1 - \Lambda\left[\alpha_{2,ev} - \phi_{ev}(\mathbf{x}_i, \mathbf{v}_{i,ev}, in_i, tr_{i1}, \Im_{it-1})\right], \tag{A-9}$$

and the probability of getting a C is its complement.

[10] $\mathbf{v}'_{i,ev}$ is a function of parameters once it is replaced by the values of the corresponding points of support.

## C.3 The School Drop-Out

In Belgium, compulsory education ends on 30 June of the year in which a student turns 18. From that date onwards, students are at risk of high school drop-out without diploma. School drop-out is an interesting long-run outcome of grade repetition that, as mentioned in the Introduction, other authors have considered as well. We model it as a binary choice in the following way for pupils at risk ($s_{it} = 1$):[11]

$$\Pr(out_{it}=1|\mathbf{x}_i, \mathbf{v}_{i,out}, in_i, tr_{i1}, \Im_{it-1}, ev_{it})=\Lambda\left[\alpha_{out}+\phi_{out}(\mathbf{x}_i, \mathbf{v}_{i,out}, in_i, tr_{i1}, \Im_{it-1}, ev_{it})\right], \tag{A-10}$$

where $\mathbf{v}_{i,out} \equiv \mathbf{v}_{i2} = \delta_{out}\mathbf{v}_{i1}$, $\delta_{out} \equiv \delta_2$ is the loading factor of the unobserved heterogeneity distribution, and where similar to Equation (A-8),

$$\begin{aligned}\phi_{out}(\mathbf{x}_i, \mathbf{v}_{i,out}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}) = (\mathbf{w}'_{it}\boldsymbol{\xi}_{out}+\mathbf{Iev}'_{it}\boldsymbol{\omega}_{out})(1+re_{it-1}\kappa^0_{out}+pre_{it-1}\psi^0_{out})\\ +re_{it-1}\kappa_{out}(1+pre_{it-1}\psi^0_{out})+pre_{it-1}\psi_{out}(1+re_{it-1}\kappa^0_{out}),\end{aligned} \tag{A-11}$$

where $\boldsymbol{\xi}_{out} \equiv \left[\boldsymbol{\beta}'_{out}\ \mathbf{v}'_{i,out}\ \gamma_{out}\ \boldsymbol{\pi}'_{out}\ \boldsymbol{\eta}'_{out}\ \boldsymbol{\delta}'_{out}\right]$, $\boldsymbol{\omega}_{out}$, $\kappa_{out}$, $\psi_{out}$, $\psi^0_{out}$ and $\kappa^0_{out}$ are parameters to be estimated and $\mathbf{Iev}_{it} = \left[\mathbf{1}_{\{A\}}(ev_{it})\ \mathbf{1}_{\{B\}}(ev_{it})\right]$.[12] Compared to Equation (A-8), $\phi_{out}$ has the extra argument, $\mathbf{Iev}_{it}$, i.e. the end-of-year evaluation. By the sequential ordering assumed in Assumption 2, $\mathbf{Iev}_{it}$ is predetermined with respect to the drop-out choice, so that it can be conditioned upon.

## C.4 The Choice of Repeating the Grade in Case of a B Evaluation

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

$$\begin{aligned}\Pr(re_{it} = 1|\mathbf{x}_i, \mathbf{v}_{i,re}, in_i, tr_{i1}, \Im_{it-1}, ev_{it} = B, out_{it}=0) =\\ \Lambda\left[\alpha_{re} + \phi_{re}(\mathbf{x}_i, \mathbf{v}_{i,re}, in_i, tr_{i1}, \Im_{it-1})\right].\end{aligned} \tag{A-12}$$

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

[12] Because of the limited number of students at risk of a drop-out decision, estimation was only possible if we grouped students with a $B$ and a $C$ into one category, so that for the drop-out decision the indicator $\mathbf{1}_{\{B\}}(ev_{it})$ was excluded. For similar reasons a coarser grouping was also imposed on $\mathbf{Idow_{it-1}}$, $\mathbf{Itr_{it}}$ and $\mathbf{Igr_{it}}$. See the results in Section C for more details.

where $\mathbf{v}_{i,re} \equiv \mathbf{v}_{i3} = \delta_{re}\mathbf{v}_{i1}$ and $\delta_{re} \equiv \delta_3$ is the loading factor of the unobserved heterogeneity distribution. Because we do not have enough $B$ observations to empirically identify heterogeneous effects of past grade repetition on the current decision to repeat the grade, the function $\phi_{re}(\mathbf{x}_i, \mathbf{v}_{i,re}, in_i, tr_{i1}, \Im_{it-1})$ is parameterized as in Equation (A-7), excluding thereby the possibility of heterogeneous retention effects across individuals:

$$\phi_{re}(\mathbf{x}_i, \mathbf{v}_{i,re}, in_i, tr_{i1}, \Im_{it-1}) \quad = \quad \mathbf{w}'_{it}\boldsymbol{\xi}_{re} + re_{it-1}\kappa_{re} + pre_{it-1}\psi_{re}. \tag{A-13}$$

where $\boldsymbol{\xi}_{re} \equiv \left[\boldsymbol{\beta}'_{re}\ \mathbf{v}'_{i,re}\ \gamma_{re}\ \boldsymbol{\pi}'_{re}\ \boldsymbol{\eta}'_{re}\ \boldsymbol{\delta}'_{re}\right]$, $\kappa_{re}$ and $\psi_{re}$ are parameters to be estimated. Note that the choice of repeating the grade must be considered neither for students in VHS nor for those in the last grade, because these students may never obtain a B (see Section 2).

## C.5 The Track Downgrade

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

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

$$\begin{aligned}
\Pr(dow_{it} &= 0|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) = \\
&\Lambda\big[\alpha_{1,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it})\big], \\
\Pr(dow_{it} &= 1|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) = \\
&\Lambda\big[\alpha_{2,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it})\big] \\
&- \Lambda\big[\alpha_{1,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it})\big], \\
\Pr(dow_{it} &= 2|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) = \\
&1 - \Lambda\big[\alpha_{2,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it})\big],
\end{aligned} \tag{A-14}$$

where $\mathbf{v}_{i,dow} \equiv \mathbf{v}_{i4} = \delta_{dow}\mathbf{v}_{i1}$ and $\delta_{dow} \equiv \delta_4$ is the loading factor of the unobserved heterogeneity distribution.

The function $\phi_{dow}(\cdot)$ is a linear index similar to the one specified in Equation (A-8):

$$\begin{aligned}\phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it}) = {} & (\mathbf{w}'_{it}\boldsymbol{\xi}_{dow} + \mathbf{Iev}'_{it}\boldsymbol{\omega}_{dow} + re_{it}\tau_{dow} + re_{it}\mathbf{Iev}'_{it}\boldsymbol{\zeta}_{dow}) \\ & \times (1 + re_{it-1}\kappa^0_{dow} + pre_{it-1}\psi^0_{dow}) \\ & + re_{it-1}\kappa_{dow}(1 + pre_{it-1}\psi^0_{dow}) \\ & + pre_{it-1}\psi_{dow}(1 + re_{it-1}\kappa^0_{dow}). \end{aligned} \tag{A-15}$$

where $\boldsymbol{\xi}_{dow} \equiv \left[\boldsymbol{\beta}'_{dow}\ \mathbf{v}'_{i,dow}\ \gamma_{dow}\ \boldsymbol{\pi}'_{dow}\ \boldsymbol{\eta}'_{dow}\ \boldsymbol{\delta}'_{dow}\right]$, $\boldsymbol{\omega}_{dow}$, $\tau_{dow}$, $\boldsymbol{\zeta}_{dow}$, $\psi_{dow}$, $\kappa_{dow}$, $\psi^0_{dow}$ and $\kappa^0_{dow}$ are parameters to be estimated, and $re_{it} = 1$ for students with a B evaluation who decided to repeat the grade and $re_{it} = 0$ otherwise. As a consequence of Assumption 1, $re_{it}$ and $\mathbf{ev}_{it}$ are predetermined and, hence, can be conditioned upon. We also allow for interactions between the latter two variables.

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

$$\begin{aligned}\Pr(dow_{it} = 0|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) = {} & \\ \Lambda\left[\alpha_{1,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it})\right], & \end{aligned} \tag{A-16}$$

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

Second, students with a B choosing to promote to the next grade are forced to downgrade, so that $dow_{it} \in \{1, 2\}$.[13] Also in this case the ordered logit simplifies to:

$$\begin{aligned}\Pr(dow_{it} = 2|\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it}, out_{it} = 0) = {} & \\ 1 - \Lambda\left[\alpha_{2,dow} - \phi_{dow}(\mathbf{x}_i, \mathbf{v}_{i,dow}, in_i, tr_{i1}, \Im_{it-1}, ev_{it}, re_{it})\right], & \end{aligned} \tag{A-17}$$

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

## C.6 The Specification of the Unobserved Heterogeneity Distribution

In Section 4 we imposed a one factor specification on the distribution of unobserved heterogeneity as to be able to identify essential and grade-varying heterogeneity. The unobserved grade-varying factor $\mathbf{v}_{i1}$ have therefore seven dimensions, one for each grade ($g = 7, ..., 13$). Since only VHS pupils have to attend grade 13 to get the diploma (in all the other tracks the diploma is obtained

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

at end of grade 12), we decided to constrain $v_{i1}(12)$ to be equal to $v_{i1}(13)$. The corresponding distribution $G(\mathbf{v}_{i1};\boldsymbol{\rho})$ is assumed to be discrete, with a finite and, a priori, unknown number $M$ of points of support, which each are vectors of real numbers of dimension $7 \times 1$. This distribution assigns with a probability $p^m \equiv \Pr(\mathbf{v}_{i1} = \mathbf{v}_1^m)$ (with $\sum_{j=1}^{M} p^j = 1$) the vector of unobserved heterogeneity terms over all grades $\mathbf{v}'_{i1}$ to the vector value of the $m^{th}$ point of support:

$$\mathbf{v}'_{i1} = \mathbf{v}_1^m \equiv [v_1^m(7)\ v_1^m(8)\ v_1^m(9)\ v_1^m(10)\ v_1^m(11)\ v_1^m(12)\ v_1^m(12)]$$

where we have set $v_1^m(13) = v_1^m(12)$ and $m \in \{1,\ldots,M\}$. By allowing all numbers $v_1^m(g)$ for $g \in \{7,\ldots,12\}$ to take on unrestricted values in the set of real numbers, we permit an arbitrary dependence structure of the unobserved heterogeneity between grades $g$.

We follow the recommendation of Gaure et al. (2007) by determining the number $M$ of points of support of this distribution on the basis of the Akaike Information Criterion (AIC). The probabilities associated to the points of support sum to one and are specified as logistic transforms:

$$p^m \equiv \Pr(\mathbf{v}_{i1} = \mathbf{v}_{i1}^m) = \frac{\exp(\rho^m)}{\sum_{h=1}^{M}\exp(\rho^h)} \quad \text{with} \quad m = 1,\ldots,M \quad \text{and} \quad \rho_M = 0. \tag{A-18}$$

The sample log-likelihood function in Equation (4) can be rewritten by replacing the integral by the following summation over all $M$ points of support:

$$\ell(\boldsymbol{\theta},\boldsymbol{\rho}) = \sum_{i=1}^{N} \ln\Big[\sum_{m=1}^{M} p^m \mathcal{L}_{im}(\boldsymbol{\theta},\boldsymbol{\rho})\Big], \tag{A-19}$$

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

# D Estimation Results of the Benchmark Model with and without unobserved heterogeneity

## D.1 Estimated Probability Masses of the Discrete Unobserved Heterogeneity Distribution

In Table A-1 we report the estimated probability masses of each point of support and other statistics of the estimated models. The number of points of support are chosen so that to minimize the Akaike Information Criterion (AIC). The resulting number of support points is $M = 3$ for both the specification controlling for grade-constant unobserved heterogeneity and the one controlling for grade-varying unobserved heterogeneity. The preferred model according to the AIC is the one with grade-varying unobserved heterogeneity. This is the reason why in our article we consider it as the benchmark model. The location of the support points and the loading factors of each equation are reported in the next Tables.

Table A-1: Estimated Probability Masses of the Discrete Unobserved Heterogeneity Distribution and Other Statistics of the Estimated Models

| | Without unobserved heterogeneity | | With grade-constant unobserved heterogeneity | | | With grade-varying unobserved heterogeneity | | |
|---|---|---|---|---|---|---|---|---|
| | Coeff. | Std. Err. | Coeff. | | Std. Err. | Coeff. | | Std. Err. |
| *Unobserved heterogeneity probability masses* | | | | | | | | |
| $\rho_1$ | – | – | -0.8966 | *** | 0.2632 | -0.9840 | *** | 0.1506 |
| $\rho_2$ | – | – | -2.0524 | *** | 0.2319 | -1.8887 | *** | 0.2416 |
| $\rho_3$ | – | – | 0.0000 | | – | 0.0000 | | – |
| *Resulting probability masses* | | | | | | | | |
| $p_1$ | – | | 0.2655 | | | 0.2451 | | |
| $p_2$ | – | | 0.0836 | | | 0.0992 | | |
| $p_3$ | – | | 0.6509 | | | 0.6557 | | |
| Log-likelihood | -17,353.8 | | -17,281.2 | | | -17,197.7 | | |
| AIC/$N$ | 8.8715 | | 8.8387 | | | 8.8003 | | |
| Number of parameters | 92 | | 100 | | | 108 | | |
| Number of pupils ($N$) | 3,933 | | 3,933 | | | 3,933 | | |

*Notes:* *** Significant at 1%.

## D.2 Estimation Results of the Equation for the Track Choice at the Beginning of Secondary School

Table A-2: Estimation results of the Track Choice at the Beginning of Secondary school

| | Without unobserved heterogeneity | | | With grade-constant unobserved heterogeneity | | | With grade-varying unobserved heterogeneity | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | Coeff. | | Std. Err. | Coeff. | | Std. Err. | Coeff. | | Std. Err. |
| Calendar day of birth/100 | -0.1649 | *** | 0.0301 | -0.1854 | *** | 0.0346 | -0.1890 | *** | 0.0359 |
| Years of delay at start of secondary school | -1.1880 | *** | 0.1481 | -1.3618 | *** | 0.1806 | -1.3514 | *** | 0.1818 |
| Female | 0.2435 | *** | 0.0607 | 0.2957 | *** | 0.0722 | 0.3122 | *** | 0.0743 |
| Cohort 1980 | -0.1526 | ** | 0.0605 | -0.1793 | ** | 0.0717 | -0.1960 | *** | 0.0740 |
| Father's education/10 | 1.6128 | *** | 0.1147 | 1.8874 | *** | 0.1506 | 1.9675 | *** | 0.1486 |
| Mother's education/10 | 1.7222 | | 0.1261 | 2.0513 | *** | 0.1611 | 2.0902 | *** | 0.1614 |
| *Number of siblings – Reference: No siblings* | | | | | | | | | |
| 1 sibling | -0.1357 | | 0.0949 | -0.1732 | | 0.1128 | -0.1930 | * | 0.1160 |
| 2 siblings | -0.2776 | *** | 0.1040 | -0.3425 | *** | 0.1236 | -0.3789 | *** | 0.1265 |
| 3 or more | -0.3890 | *** | 0.1166 | -0.4997 | *** | 0.1384 | -0.5775 | *** | 0.1426 |
| *Ordered logit thresholds* | | | | | | | | | |
| $\alpha_{1,tr}$ | -3.6181 | *** | 0.1269 | -5.7932 | *** | 0.3749 | -6.4144 | *** | 0.3414 |
| $\ln(\alpha_{2,tr} - \alpha_{1,tr})$ | 0.5734 | *** | 0.0426 | 0.6883 | *** | 0.0495 | 0.6832 | *** | 0.0447 |
| $\ln(\alpha_{3,tr} - \alpha_{2,tr})$ | -0.4637 | *** | 0.0532 | -0.3141 | *** | 0.0610 | -0.3055 | *** | 0.0563 |
| $\ln(\alpha_{4,tr} - \alpha_{3,tr})$ | 0.7777 | *** | 0.0239 | 0.9536 | *** | 0.0421 | 1.0060 | *** | 0.0356 |
| Unobs. heter. loading factor $\delta_{tr}$ | – | | – | 0.7523 | *** | 0.1157 | 0.5506 | *** | 0.1585 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

## D.3 Estimation Results of the Evaluation Equation

Table A-3: Estimation results of the Evaluation Equation

| Variable | Without unobserved heterogeneity Coeff. | | Std. Err. | With grade-constant unobserved heterogeneity Coeff. | | Std. Err. | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|---|---|---|---|---|---|
| *Time-constant variables* | | | | | | | | | |
| Years of delay at start of secondary school | -0.3600 | *** | 0.1060 | -0.7880 | *** | 0.1690 | -0.6881 | *** | 0.1446 |
| Female | 0.4551 | *** | 0.0507 | 0.6161 | *** | 0.0709 | 0.5618 | *** | 0.0668 |
| Cohort 1980 | -0.0504 | | 0.0489 | -0.0892 | | 0.0666 | -0.0908 | | 0.0640 |
| Father's education/10 | 0.0945 | | 0.0908 | 0.5796 | *** | 0.1489 | 0.4895 | *** | 0.1311 |
| Mother's education/10 | 0.2739 | *** | 0.0978 | 0.8463 | *** | 0.1609 | 0.7113 | *** | 0.1431 |
| *Number of siblings – Reference: No siblings* | | | | | | | | | |
| 1 sibling | 0.0783 | | 0.0736 | 0.0423 | | 0.1021 | 0.0380 | | 0.0982 |
| 2 siblings | -0.0813 | | 0.0796 | -0.1748 | | 0.1107 | -0.1781 | * | 0.1056 |
| 3 or more | -0.0994 | | 0.0921 | -0.2634 | ** | 0.1287 | -0.2750 | ** | 0.1218 |
| *Ordered logit thresholds* | | | | | | | | | |
| $\alpha_{1,ev}$ | -4.9890 | *** | 0.1743 | -9.2160 | *** | 0.5382 | -11.0402 | *** | 1.4219 |
| $\ln(\alpha_{2,ev} - \alpha_{1,ev})$ | 0.3102 | *** | 0.0364 | 0.3768 | *** | 0.0370 | 0.3527 | *** | 0.0377 |
| *Time-varying variables* | | | | | | | | | |
| *Track in year $t$ – Reference: VHS* | | | | | | | | | |
| GHS+ | 0.4539 | *** | 0.1649 | -1.5122 | *** | 0.3648 | -1.1482 | *** | 0.2955 |
| GHS− | -1.0408 | *** | 0.1176 | -2.4244 | *** | 0.2475 | -2.0643 | *** | 0.2092 |
| THS+ | -1.4480 | *** | 0.1261 | -2.4171 | *** | 0.2074 | -2.1034 | *** | 0.1901 |
| THS− | -1.3699 | *** | 0.1202 | -2.0922 | *** | 0.1686 | -1.8558 | *** | 0.1606 |
| *Grade in year $t$– Reference: Grade 7* | | | | | | | | | |
| Grade 8 | -0.7613 | *** | 0.1034 | -0.9216 | *** | 0.1088 | -4.6753 | *** | 1.5362 |
| Grade 9[§] | -0.4126 | *** | 0.1128 | -0.7146 | *** | 0.1241 | $\infty$ | | – |
| Grade 10 | -0.5370 | *** | 0.1052 | -0.9725 | *** | 0.1244 | -4.1727 | *** | 1.5712 |
| Grade 11 | -0.4702 | *** | 0.1060 | -1.0158 | *** | 0.1311 | 3.1741 | | 3.7086 |
| Grade 12 if VHS | -0.7217 | *** | 0.2767 | -1.3160 | *** | 0.2708 | -5.1179 | *** | 1.5406 |
| Last grade | 0.5885 | *** | 0.1343 | -0.0732 | | 0.1544 | -4.1320 | *** | 1.5060 |
| *Downgrade at the end of year $t-1$ – Reference: No downgrade* | | | | | | | | | |
| 1-step downgrade | 0.1928 | * | 0.1134 | 0.1139 | | 0.1267 | 0.0690 | | 0.1262 |
| 2-step downgrade | 0.2482 | | 0.1965 | -0.0691 | | 0.2120 | 0.0348 | | 0.2079 |
| Ever retained before year $t-1$ ($\psi_{ev}$) | -1.0267 | *** | 0.1797 | 0.3986 | | 0.2459 | 0.2149 | | 0.3469 |
| Retention at the end of year $t-1$ ($\kappa_{ev}$) | -0.1544 | | 0.2713 | -3.2002 | *** | 0.8451 | -2.4716 | * | 1.4758 |
| Heterogeneous effect of Ever retained before year $t-1$ ($\psi^0_{ev}$) | -0.2132 | * | 0.1199 | 0.0779 | ** | 0.0350 | 0.0574 | | 0.0428 |
| Heterogeneous effect of Retention at the end of year $t-1$ ($\kappa^0_{ev}$) | -0.2741 | ** | 0.1346 | -0.6598 | *** | 0.1191 | -0.4127 | ** | 0.1697 |
| *Unobserved heterogeneity* | | | | | | | | | |
| *Unobserved heterogeneity support points – $\mathbf{v}^1_1$ normalized to $\mathbf{0}$* | | | | | | | | | |
| $v^2_1(7)$ | – | | – | -4.6092 | *** | 0.4885 | -6.5837 | *** | 1.5676 |
| *Grade varying unobserved heterogeneity of $v^2_1(.) - v^2_1(7)$* | | | | | | | | | |
| $v^2_1(8) - v^2_1(7)$ | – | | – | – | | – | 4.3095 | *** | 1.6636 |
| $v^2_1(9) - v^2_1(7)$ | – | | – | – | | – | -15.7057 | *** | 0.2898 |
| $v^2_1(10) - v^2_1(7)$ | – | | – | – | | – | 3.4792 | ** | 1.6762 |
| $v^2_1(11) - v^2_1(7)$ | – | | – | – | | – | -4.2500 | | 3.7719 |
| $v^2_1(12) - v^2_1(7) = v^2_1(13) - v^2_1(7)$ | – | | – | – | | – | 3.7905 | ** | 1.6374 |
| $v^3_1(7)$ | – | | – | -2.6645 | *** | 0.4084 | -4.7825 | *** | 1.4038 |
| *Grade varying unobserved heterogeneity of $v^3_1(.) - v^3_1(7)$* | | | | | | | | | |
| $v^3_1(8) - v^3_1(7)$ | – | | – | – | | – | 3.7260 | ** | 1.5041 |
| $v^3_1(9) - v^3_1(7)$ | – | | – | – | | – | -15.7728 | *** | 0.1573 |
| $v^3_1(10) - v^3_1(7)$ | – | | – | – | | – | 3.2926 | ** | 1.5546 |
| $v^3_1(11) - v^3_1(7)$ | – | | – | – | | – | -4.0617 | | 3.7004 |
| $v^3_1(12) - v^3_1(7) = v^3_1(13) - v^3_1(7)$ | – | | – | – | | – | 4.6492 | *** | 1.5177 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

[§] When we include grade-varying heterogeneity we need to fix the coefficient of grade 9 since it tends to be a very large number.

## D.4 Estimation Results of the Drop-Out Equation

Table A-4: Estimation results of the School Drop-Out Equation

| Variable | Without unobserved heterogeneity Coeff. | | Std. Err. | With grade-constant unobserved heterogeneity Coeff. | | Std. Err. | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|---|---|---|---|---|---|
| *Time-constant variables* | | | | | | | | | |
| Years of delay at start of secondary school | -0.6436 | *** | 0.2414 | -0.7978 | *** | 0.2580 | -0.1486 | | 0.2457 |
| Female | -0.2316 | | 0.1665 | -0.1992 | | 0.1751 | -0.1592 | | 0.1699 |
| Cohort 1980 | 0.2335 | | 0.1631 | 0.2776 | | 0.1709 | 0.2157 | | 0.1628 |
| Father's education/10 | -0.8076 | ** | 0.3139 | -0.7160 | ** | 0.3311 | -0.7298 | ** | 0.3168 |
| Mother's education/10 | -0.2614 | | 0.3192 | -0.0661 | | 0.3526 | -0.1525 | | 0.3241 |
| Presence of siblings | -0.1530 | | 0.2205 | -0.1882 | | 0.2290 | -0.0403 | | 0.2227 |
| Constant $\alpha_{out}$ | -0.3886 | | 0.3863 | 0.8502 | | 0.6724 | 1.0376 | ** | 0.4711 |
| *Time-varying variables* | | | | | | | | | |
| Track in year $t$: VHS[§] | 2.8119 | *** | 0.3325 | 3.3619 | *** | 0.4105 | 2.3220 | *** | 0.3809 |
| Evaluation in year $t$: A[§] | -3.3873 | *** | 0.3081 | -3.9154 | *** | 0.3505 | -4.0952 | *** | 0.3930 |
| Grade in year $t$: final grade[§] | -2.4515 | *** | 0.3686 | -2.6123 | *** | 0.3851 | -3.3014 | *** | 0.4246 |
| 1-step or 2-step downgrade at the end of year $t-1$[§] | -1.0167 | * | 0.5632 | -1.1814 | * | 0.6047 | -0.5673 | | 0.5819 |
| Ever retained before year $t-1$ ($\psi_{out}$) | -0.7461 | *** | 0.2317 | -1.0981 | *** | 0.3106 | -0.6281 | ** | 0.2955 |
| Retention at the end of year $t-1$ ($\kappa_{out}$) | -0.2450 | | 0.3428 | -0.6272 | | 0.4678 | 0.5842 | | 0.5577 |
| Heterogeneous effect of Ever retained before year $t-1$ ($\psi^0_{ev}$) | -0.2991 | *** | 0.0835 | -0.3436 | *** | 0.0752 | -0.2811 | *** | 0.0802 |
| Heterogeneous effect of Retention at the end of year $t-1$ ($\kappa^0_{out}$) | -0.2174 | * | 0.1160 | -0.2652 | *** | 0.1006 | -0.1460 | | 0.1097 |
| Unobserved heter. loading factor $\delta_{out}$ | – | | – | -0.4101 | ** | 0.1918 | -0.3104 | ** | 0.1308 b] |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

[§] We had to group track, evaluation, grade, and track downgrade dummies into broader categories due to the scarce number of observations in some categories if defined at a finer level.

## D.5 Estimation Results of the Resitting Equation for B Students

Table A-5: Estimation results of the Resitting Equation for B Students

| Variable | Without unobserved heterogeneity Coeff. | | Std. Err. | With grade-constant unobserved heterogeneity Coeff. | | Std. Err. | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|---|---|---|---|---|---|
| *Time-constant variables* | | | | | | | | | |
| Years of delay at start of secondary school | -0.5316 | | 0.4986 | -0.6861 | | 0.5090 | -0.8854 | * | 0.5251 |
| Female | -0.1148 | | 0.1854 | -0.0674 | | 0.1933 | -0.0162 | | 0.1971 |
| Cohort 1980 | 0.2018 | | 0.1870 | 0.2362 | | 0.1895 | 0.2666 | | 0.1988 |
| Father's education/10 | 0.7340 | ** | 0.3286 | 0.8598 | ** | 0.3471 | 1.0480 | *** | 0.3624 |
| Mother's education/10 | 0.0190 | | 0.3595 | 0.1461 | | 0.3941 | 0.3033 | | 0.3916 |
| Presence of siblings | 0.5060 | * | 0.2750 | 0.5825 | ** | 0.2777 | 0.6222 | ** | 0.2918 |
| Constant $\alpha_{re}$ | -1.8411 | *** | 0.3380 | -1.0999 | | 0.8302 | -0.5092 | | 0.4950 |
| *Time-varying variables* | | | | | | | | | |
| *Track in year $t$ – Reference: THS+/THS−*[§] | | | | | | | | | |
| GHS+/GHS− | -0.3410 | * | 0.1995 | -0.6003 | *** | 0.2326 | -0.8633 | *** | 0.2258 |
| *Grade in year $t$ – Reference: Grade 8* | | | | | | | | | |
| Grade 9[†] | 0.5879 | ** | 0.2527 | 0.4328 | * | 0.2584 | $+\infty$ | | – |
| Grade 10 | 1.1516 | *** | 0.2139 | 0.9546 | *** | 0.2275 | 1.1661 | ** | 0.5300 |
| 1-step or 2-step downgrade at the end of year $t-1$[§] | 0.4993 | | 0.3332 | 0.3788 | | 0.3453 | 0.4466 | | 0.3574 |
| Ever retained before year $t-1$ ($\psi_{re}$) | -0.3644 | | 0.4318 | -0.2809 | | 0.4683 | -0.1598 | | 0.4779 |
| Retention at the end of year $t-1$ ($\kappa_{re}$) | -2.5382 | ** | 1.1435 | -2.4785 | ** | 1.1477 | -2.3282 | ** | 1.1602 |
| Unobserved heter. loading factor $\delta_{re}$ | – | | – | -0.1732 | | 0.2305 | -0.7488 | *** | 0.0601 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

[§] We had to group track and track downgrade dummies into broader categories due to the scarce number of observations in some categories if defined at a finer level.

[†] When we include grade-varying heterogeneity we need to fix the coefficient of grade 9 since it tends to be a very large number.

## D.6 Estimation Results of the Track Downgrade Equation for GHS/THS Students

Table A-6: Estimation Results of the Track Downgrade Equation for GHS/THS Students

| Variable | Without unobserved heterogeneity Coeff. | | Std. Err. | With grade-constant unobserved heterogeneity Coeff. | | Std. Err. | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|---|---|---|---|---|---|
| *Time-constant variables* | | | | | | | | | |
| Years of delay at start of secondary school | 0.2481 | * | 0.1463 | 0.3019 | * | 0.1565 | 0.3287 | ** | 0.1542 |
| Female | -0.0441 | | 0.0634 | -0.0576 | | 0.0645 | -0.0642 | | 0.0641 |
| Cohort 1980 | -0.0594 | | 0.0633 | -0.0564 | | 0.0633 | -0.0523 | | 0.0633 |
| Father's education/10 | -0.6019 | *** | 0.1173 | -0.6557 | *** | 0.1297 | -0.6942 | *** | 0.1248 |
| Mother's education/10 | -0.3596 | *** | 0.1280 | -0.4287 | *** | 0.1391 | -0.4771 | *** | 0.1338 |
| *Number of siblings – Reference: No siblings* | | | | | | | | | |
| 1 sibling | -0.0291 | | 0.0933 | -0.0255 | | 0.0932 | -0.0165 | | 0.0931 |
| 2 siblings | -0.1076 | | 0.1035 | -0.0943 | | 0.1041 | -0.0740 | | 0.1039 |
| 3 or more | -0.0038 | | 0.1214 | 0.0157 | | 0.1228 | 0.0458 | | 0.1217 |
| *Ordered logit thresholds* | | | | | | | | | |
| $\alpha_{1,dow}$ | 0.9913 | *** | 0.1817 | 1.3774 | *** | 0.3767 | 1.4731 | *** | 0.2748 |
| $\ln(\alpha_{2,dow} - \alpha_{1,dow})$ | 0.4950 | *** | 0.0445 | 0.4951 | *** | 0.0446 | 0.4980 | *** | 0.0446 |
| *Time-varying variables* | | | | | | | | | |
| *Evaluation and retention in year t– Reference: C* | | | | | | | | | |
| A | -1.6592 | *** | 0.1229 | -1.5529 | *** | 0.1507 | -1.4820 | *** | 0.1354 |
| B and not resitting | 1.4570 | *** | 0.1616 | 1.4644 | *** | 0.1620 | 1.4703 | *** | 0.1622 |
| B and resitting | -0.8837 | *** | 0.2424 | -0.8585 | *** | 0.2454 | -0.7789 | *** | 0.2456 |
| *Track in year t – Reference: VHS−* | | | | | | | | | |
| GHS+ | 1.7198 | *** | 0.1314 | 1.8787 | *** | 0.1972 | 2.0138 | *** | 0.1694 |
| GHS− | 0.8603 | *** | 0.1221 | 0.9307 | *** | 0.1415 | 0.9722 | *** | 0.1308 |
| THS+ | 1.2203 | *** | 0.1334 | 1.2430 | *** | 0.1367 | 1.2386 | *** | 0.1348 |
| *Grade in year t– Reference: Grade 8* | | | | | | | | | |
| Grade 7 | -3.8673 | *** | 0.2111 | -3.8833 | *** | 0.2121 | -4.3862 | *** | 0.4195 |
| Grade 9 | -1.2577 | *** | 0.0917 | -1.2454 | *** | 0.0919 | -4.3991 | ** | 1.7462 |
| Grade 10 | -0.5479 | *** | 0.0701 | -0.5231 | *** | 0.0732 | -0.5625 | *** | 0.1080 |
| *Downgrade at the end of year $t-1$ – Reference: No downgrade* | | | | | | | | | |
| 1-step downgrade | -0.4251 | ** | 0.1878 | -0.4203 | ** | 0.1880 | -0.3795 | ** | 0.1862 |
| 2-step downgrade | -0.6925 | | 0.4591 | -0.6551 | | 0.4628 | -0.6368 | | 0.4571 |
| Ever retained before year $t-1$ ($\psi_{dow}$) | 0.7733 | *** | 0.2362 | 0.5827 | ** | 0.2693 | 0.5275 | ** | 0.2364 |
| Retention at the end of year $t-1$ ($\kappa_{dow}$) | 0.7582 | *** | 0.1817 | 0.6013 | *** | 0.1982 | 0.5883 | *** | 0.1812 |
| Heterogeneous effect of Ever retained before year $t-1$ ($\psi^0_{dow}$) | 0.2884 | * | 0.1585 | 0.3037 | * | 0.1624 | 0.3500 | ** | 0.1721 |
| Heterogeneous effect of Retention at the end of year $t-1$ ($\kappa^0_{dow}$) | 0.1772 | | 0.1235 | 0.1778 | | 0.1261 | 0.2380 | * | 0.1370 |
| Unobserved heter. loading factor $\delta_{dow}$ | – | | – | -0.0998 | | 0.0820 | -0.1787 | ** | 0.0907 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

# E Estimation Results if Downgrading Is First Decided upon by Students Getting a B with Grade-Varying Unobserved Heterogeneity

In the benchmark model we assume that the choice to repeat the grade in case of a B precedes the track downgrading decision. However, one might question this assumption: B students might first decide whether to stay in the current track and, if B students decide to remain in the same track, no choice about resitting is left. In this section, we report the estimation results of the model in which B students first are assumed to decide first whether to stay in the current track. After the estimation of this alternative model, which is non-nested to the benchmark one, we discriminate between models on the basis of the Vuong test for strictly non-nested models (Vuong, 1989, p. 316–319). We find that the alternative order of events could be rejected against the one of the benchmark model. The value of the asymptotically standard Normal statistic is 3.798 in favor of the benchmark model and rejects the alternative hypothesis at a $p$-value of 0.0001.

## E.1 Estimated Probability Masses of the Discrete Unobserved Heterogeneity Distribution if Downgrading Is First Decided upon by Students Getting a B

Table A-7: Estimated Probability Masses of the Discrete Unobserved Heterogeneity Distribution and Other Statistics of the Model if Downgrading Is First Decided upon by Students Getting a B

| | With grade-varying unobserved heterogeneity | | |
|---|---|---|---|
| | Coeff. | | Std. Err. |
| *Unobserved heterogeneity probability masses* | | | |
| $\rho_1$ | -0.9822 | *** | 0.1531 |
| $\rho_2$ | -1.8634 | *** | 0.2429 |
| $\rho_3$ | 0.0000 | | – |
| *Resulting probability masses* | | | |
| $p_1$ | 0.2448 | | |
| $p_2$ | 0.1014 | | |
| $p_3$ | 0.6538 | | |
| Log-likelihood | -17,227.9 | | |
| AIC/$N$ | 8.8166 | | |
| Number of parameters | 110 | | |
| Number of pupils ($N$) | 3,933 | | |

*Notes:* *** Significant at 1%.

## E.2 Estimation Results of the Equation for the Track Choice at the Beginning of Secondary school if Downgrading Is First Decided upon by Students Getting a B

Table A-8: Estimation results of the Track Choice at the Beginning of Secondary School if Downgrading Is First Decided upon by Students Getting a B

| Variable | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|
| Calendar day of birth/100 | -0.1889 | *** | 0.0359 |
| Years of delay at start of secondary school | -1.3520 | *** | 0.1820 |
| Female | 0.3121 | *** | 0.0742 |
| Cohort 1980 | -0.1965 | *** | 0.0739 |
| Father's education/10 | 1.9724 | *** | 0.1485 |
| Mother's education/10 | 2.0891 | *** | 0.1614 |
| *Number of siblings – Reference: No siblings* | | | |
| 1 sibling | -0.1920 | * | 0.1160 |
| 2 siblings | -0.3815 | *** | 0.1264 |
| 3 or more | -0.5817 | *** | 0.1426 |
| *Ordered logit thresholds* | | | |
| $\alpha_{1,tr}$ | -6.4252 | *** | 0.3407 |
| $\ln(\alpha_{2,tr} - \alpha_{1,tr})$ | 0.6847 | *** | 0.0445 |
| $\ln(\alpha_{3,tr} - \alpha_{2,tr})$ | -0.3084 | *** | 0.0563 |
| $\ln(\alpha_{4,tr} - \alpha_{3,tr})$ | 1.0066 | *** | 0.0357 |
| Unobserved heterogeneity loading factor $\delta_{tr}$ | 0.5516 | *** | 0.1611 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

## E.3 Estimation Results of the Evaluation Equation if Downgrading Is First Decided upon by Students Getting a B

Table A-9: Estimation results of the Evaluation Equation if Downgrading Is First Decided upon by Students Getting a B

| Variable | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|
| *Time-constant variables* | | | |
| Years of delay at start of secondary school | -0.6840 | *** | 0.1458 |
| Female | 0.5561 | *** | 0.0668 |
| Cohort 1980 | -0.0932 | | 0.0638 |
| Father's education/10 | 0.4847 | *** | 0.1318 |
| Mother's education/10 | 0.7079 | *** | 0.1440 |
| *Number of siblings – Reference: No siblings* | | | |
| 1 sibling | 0.0351 | | 0.0982 |
| 2 siblings | -0.1816 | * | 0.1057 |
| 3 or more | -0.2830 | ** | 0.1211 |
| *Ordered logit thresholds* | | | |
| $\alpha_{1,ev}$ | -11.0443 | *** | 1.4468 |
| $\ln(\alpha_{2,ev} - \alpha_{1,ev})$ | 0.3534 | *** | 0.0376 |
| *Time-varying variables* | | | |
| *Track in year $t$ – Reference: VHS* | | | |
| GHS+ | -1.1388 | *** | 0.3034 |
| GHS− | -2.0550 | *** | 0.2138 |
| THS+ | -2.0905 | *** | 0.1930 |
| THS− | -1.8585 | *** | 0.1631 |
| *Grade in year $t$– Reference: Grade 7* | | | |
| Grade 8 | -4.6766 | *** | 1.5674 |
| Grade 9[§] | 12.6568 | | 11.5651 |
| Grade 10 | -4.2873 | *** | 1.5783 |
| Grade 11 | 2.0600 | | 4.1883 |
| Grade 12 if VHS | -5.1805 | *** | 1.5653 |
| Last grade | -4.1756 | *** | 1.5322 |
| *Downgrade at the end of year $t-1$ – Reference: No downgrade* | | | |
| 1-step downgrade | 0.0854 | | 0.1266 |
| 2-step downgrade | 0.0342 | | 0.2092 |
| Ever retained before year $t-1$ ($\psi_{ev}$) | 0.1795 | | 0.3558 |
| Retention at the end of year $t-1$ ($\kappa_{ev}$) | -2.5926 | * | 1.4655 |
| Heterogeneous effect of Ever retained before year $t-1$ ($\psi^0_{ev}$) | 0.0561 | | 0.0438 |
| Heterogeneous effect of Retention at the end of year $t-1$ ($\kappa^0_{ev}$) | -0.4265 | ** | 0.1667 |
| *Unobserved heterogeneity* | | | |
| *Unobserved heterogeneity support points – $\mathbf{v}^1_1$ normalized to $\mathbf{0}$* | | | |
| $v^2_1(7)$ | -6.5628 | *** | 1.6000 |
| *Grade varying unobserved heterogeneity of $v^2_1(.) - v^2_1(7)$* | | | |
| $v^2_1(8) - v^2_1(7)$ | 4.3357 | ** | 1.7106 |
| $v^2_1(9) - v^2_1(7)$ | -13.3026 | | 11.5656 |
| $v^2_1(10) - v^2_1(7)$ | 3.5516 | ** | 1.6996 |
| $v^2_1(11) - v^2_1(7)$ | -3.1056 | | 4.2206 |
| $v^2_1(12) - v^2_1(7)=v^2_1(13) - v^2_1(7)$ | 3.8670 | ** | 1.6791 |
| $v^3_1(7)$ | -4.7856 | *** | 1.4293 |
| *Grade varying unobserved heterogeneity of $v^3_1(.) - v^3_1(7)$* | | | |
| $v^3_1(8) - v^3_1(7)$ | 3.7178 | ** | 1.5299 |
| $v^3_1(9) - v^3_1(7)$ | -13.3307 | | 11.5615 |
| $v^3_1(10) - v^3_1(7)$ | 3.4381 | ** | 1.5571 |
| $v^3_1(11) - v^3_1(7)$ | -2.9476 | | 4.1802 |
| $v^3_1(12) - v^3_1(7)=v^3_1(13) - v^3_1(7)$ | 4.7205 | *** | 1.5438 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

## E.4 Estimation Results of the Drop-Out Equation if Downgrading Is First Decided upon by Students Getting a B

Table A-10: Estimation results of the School Drop-Out Equation if Downgrading Is First Decided upon by Students Getting a B

| Variable | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|
| *Time-constant variables* | | | |
| Years of delay at start of secondary school | -0.8741 | ** | 0.3648 |
| Female | -0.1486 | | 0.2471 |
| Cohort 1980 | -0.1586 | | 0.1709 |
| Father's education/10 | 0.2145 | | 0.1635 |
| Mother's education/10 | -0.7144 | ** | 0.3228 |
| Presence of siblings | -0.1345 | | 0.3272 |
| Constant $\alpha_{out}$ | -0.5751 | | 0.5805 |
| *Time-varying variables* | | | |
| Track in year $t$: THS[§] | -0.1465 | | 0.1100 |
| Evaluation in year $t$: A[§] | 2.3716 | *** | 0.4148 |
| Grade in year $t$: final grade[§] | -4.1158 | *** | 0.4299 |
| 1-step or 2-step downgrade at the end of year $t-1$[§] | -3.2722 | *** | 0.4253 |
| Ever retained before year $t-1$ ($\psi_{out}$) | -0.0370 | | 0.2237 |
| Retention at the end of year $t-1$ ($\kappa_{out}$) | -0.6018 | ** | 0.2992 |
| Heterogeneous effect of Ever retained before year $t-1$ ($\psi^0_{ev}$) | 0.6099 | | 0.5710 |
| Heterogeneous effect of Retention at the end of year $t-1$ ($\kappa^0_{out}$) | -0.2801 | *** | 0.0806 |
| Unobserved heter. loading factor $\delta_{out}$ | 1.0168 | ** | 0.4847 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.
§ We had to group track, evaluation, grade, and track downgrade dummies into broader categories due to the scarce number of observations in some categories if defined at a finer level.

## E.5 Estimation Results of the Resitting Equation for B Students if Downgrading Is First Decided upon by Students Getting a B

Table A-11: Estimation results of the Resitting Equation for B Students if Downgrading Is First Decided upon by Students Getting a B

| Variable | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|
| *Time-constant variables* | | | |
| Years of delay at start of secondary school | -0.9396 | * | 0.5253 |
| Female | 0.1797 | | 0.2121 |
| Cohort 1980 | 0.3592 | * | 0.2125 |
| Father's education/10 | 0.8866 | ** | 0.3962 |
| Mother's education/10 | 0.5750 | | 0.4209 |
| Presence of siblings | 0.8268 | *** | 0.3112 |
| Constant $\alpha_{re}$ | -2.3301 | *** | 0.4192 |
| *Time-varying variables* | | | |
| *Track in year t – Reference: THS+/THS−*[§] | | | |
| GHS+/GHS− | -0.4920 | * | 0.2609 |
| *Grade in year t– Reference: Grade 8* | | | |
| Grade 9 | 15.3336 | | 11.9525 |
| Grade 10 | 1.1072 | * | 0.5965 |
| 1-step or 2-step downgrade at the end of year $t-1$[§] | 0.6850 | * | 0.3773 |
| Ever retained before year $t-1$ ($\psi_{re}$) | 0.1626 | | 0.5271 |
| Retention at the end of year $t-1$ ($\kappa_{re}$) | -2.0478 | * | 1.1930 |
| Unobserved heter. loading factor $\delta_{re}$ | -0.5561 | | 0.6561 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

[§] We had to group track and track downgrade dummies into broader categories due to the scarce number of observations in some categories if defined at a finer level.

## E.6 Estimation Results of the Track Downgrade Equation for GHS/THS Students if Downgrading Is First Decided upon by Students Getting a B

Table A-12: Estimation Results of the Track Downgrade Equation for GHS/THS Students if Downgrading Is First Decided upon by Students Getting a B

| Variable | With grade-varying unobserved heterogeneity Coeff. | | Std. Err. |
|---|---|---|---|
| *Time-constant variables* | | | |
| Years of delay at start of secondary school | 0.3778 | ** | 0.1560 |
| Female | -0.0642 | | 0.0639 |
| Cohort 1980 | -0.0564 | | 0.0630 |
| Father's education/10 | -0.7433 | *** | 0.1252 |
| Mother's education/10 | -0.4734 | *** | 0.1338 |
| *Number of siblings – Reference: No siblings* | | | |
| 1 sibling | -0.0178 | | 0.0932 |
| 2 siblings | -0.0673 | | 0.1038 |
| 3 or more | 0.0368 | | 0.1208 |
| *Ordered logit thresholds* | | | |
| $\alpha_{1,dow}$ | 1.6551 | *** | 0.2888 |
| $\ln(\alpha_{2,dow} - \alpha_{1,dow})$ | 0.4117 | *** | 0.0457 |
| *Time-varying variables* | | | |
| *Evaluation and retention in year t– Reference: C* | | | |
| A | -1.4662 | *** | 0.1373 |
| B | 0.6607 | *** | 0.1483 |
| *Track in year t – Reference: VHS−* | | | |
| GHS+ | 2.1744 | *** | 0.1741 |
| GHS− | 1.1213 | *** | 0.1307 |
| THS+ | 1.3436 | *** | 0.1344 |
| *Grade in year t– Reference: Grade 8* | | | |
| Grade 7 | -4.5312 | *** | 0.4713 |
| Grade 9 | -4.5630 | * | 2.6630 |
| Grade 10 | -0.5725 | *** | 0.1149 |
| *Downgrade at the end of year $t-1$ – Reference: No downgrade* | | | |
| 1-step downgrade | -0.4150 | ** | 0.1789 |
| 2-step downgrade | -0.5763 | | 0.4540 |
| Ever retained before year $t-1$ ($\psi_{dow}$) | 0.4727 | ** | 0.2187 |
| Retention at the end of year $t-1$ ($\kappa_{dow}$) | 0.6569 | *** | 0.2055 |
| Heterogeneous effect of Ever retained before year $t-1$ ($\psi^0_{dow}$) | 0.3098 | * | 0.1819 |
| Heterogeneous effect of Retention at the end of year $t-1$ ($\kappa^0_{dow}$) | 0.3723 | ** | 0.1572 |
| Unobserved heter. loading factor $\delta_{dow}$ | -0.2141 | ** | 0.0980 |

*Notes:* *** Significant at 1%; ** significant at 5%; * significant at 10%.

# F Additional ATTs Based on Counterfactual Simulations

Table A-13: Treatment Heterogeneity of Grade Repetition (C) Relative to Forced Downgrading (B) after Grade 8

| Treatment: C versus B evaluation after grade 8 | Model: grade-varying unobserved heterogeneity | | |
|---|---|---|---|
| | ATT | | 95% CI |
| *I. Evaluation in grade 9: A* | | | |
| First quartile | -0.194 | ** | [-0.384, -0.014] |
| Second quartile | -0.038 | | [-0.186, 0.096] |
| Third quartile | 0.019 | | [-0.097, 0.123] |
| Fourth quartile | 0.064 | | [-0.039, 0.155] |
| *II. High school graduation* | | | |
| First quartile | -0.179 | * | [-0.371, 0.000] |
| Second quartile | -0.033 | | [-0.208, 0.128] |
| Third quartile | 0.024 | | [-0.131, 0.159] |
| Fourth quartile | -0.007 | | [-0.181, 0.145] |
| *III. Delay at start last compulsory year* | | | |
| First quartile | 0.874 | | [0.651, 1.106] |
| Second quartile | 0.721 | ** | [0.538, 0.918] |
| Third quartile | 0.659 | *** | [0.485, 0.844] |
| Fourth quartile | 0.621 | *** | [0.458, 0.801] |

*Notes:* All statistics are based on 999 random simulations of the treated sample that allow for the uncertainty of the estimated parameters. The ATTs are calculated by subtracting the average outcome in case of the counterfactual of forced downgrading from the average outcome in case of a retention. The first quartile is the one with the lowest value for the linear index of the evaluation in grade 8. ***, **, * indicate whether the ATT is significantly different from 0 (1) in panels *I* and *II* (panel *III*) at the 1%, 5%, 10% significance levels, respectively.

Table A-14: ATTs of Grade Repetition (C) Relative to Forced Downgrading (B) after Grade 9 and 10

| | Model: grade-varying unobserved heterogeneity | | | | | |
|---|---|---|---|---|---|---|
| | Treatment: C versus B evaluation after grade 9 | | | Treatment: C versus A evaluation after grade 10 | | |
| | ATT | | 95% CI | ATT | | 95% CI |
| *I. Evaluation in next grade: A* | | | | | | |
| All treated | -0.055 | | [-0.149, 0.038] | – | | – |
| *II. High school diploma* | | | | | | |
| All treated | -0.038 | | [-0.128, 0.048] | -0.035 | | [-0.111, 0.042] |
| *III. Delay at start last compulsory year* | | | | | | |
| All treated | 0.643 | *** | [0.504, 0.767] | 0.558 | *** | [0.440, 0.667] |

*Notes:* All statistics are based on 999 random simulations of the treated sample that allow for the uncertainty of the estimated parameters. The ATTs are calculated by subtracting the average outcome in case of the counterfactual of forced downgrading from the average outcome in case of a retention. Panel *I* is empty for treatments in grade 10 since not all individuals reach grade 11 (and therefore outcomes for the latter year cannot be calculated for all individuals). ***, **, * indicate whether the ATT is significantly different from 0 (1) in panels *I* and *II* (panel *III*) at the 1%, 5%, 10% significance levels, respectively.

# References


Billingsley, P. (1995) *Probability and measure,* $3^{rd}$ ed. (New York: Wiley)

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

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

Mroz, T., G. Picone, F. Sloan, and A. Y. Yashkin (2016) 'Screening for a chronic disease: A multiple stage duration model with partial observability.' *International Economic Review* 57(3), 915–934

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